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 least 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-posteriori error analysis ensures that the computed (suboptimal) POD solutions are sufficiently accurate. Numerical test illustrates the efficiency of the proposed strategy.

Keywords Convection-diffusion equation · Optimal control · State constraints · Primal-dual active set strategy · Model order reduction

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 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 [15] and a mesh-independent

L. Mechelli (✉) · S. Volkwein
University of Konstanz, Department of Mathematics and Statistics, Konstanz, Germany
e-mail: Luca.Mechelli@uni-konstanz.de; Stefan.Volkwein@uni-konstanz.de

M. Falcone et al. (eds.), *Numerical Methods for Optimal Control Problems*,
Springer INdAM Series 29, https://doi.org/10.1007/978-3-030-01959-4_4

property [16]. 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, 13]. To have sufficiently accurate POD suboptimal solutions, we adapt the a-posteriori error analysis from [9]. 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 captures 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], OS-POD [20] or adaptive strategies like in [1]. However, in this paper, we will only compare the quality of basis generated with arbitrary controls and with the idealized ones generated from the optimal finite element controls. Our motivation comes from the fact that we will utilize the proposed strategy within an economic model predictive control approach [10, Chapter 8], where the POD basis will be eventually updated during the closed-loop realization; cf. [22]. In contrast to [9] we consider economic costs, boundary controls, two-dimensional spatial domains and time- as well as space-dependent convection fields.

The paper is organized in the following way: in Sect. 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 Sect. 3. In Sect. 4 we briefly explain the POD method and the related a-posteriori error estimator is presented in Sect. 5. Numerical Tests are shown in Sect. 6. Finally, some conclusions are drawn in Sect. 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_{o}, 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

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

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 $\Sigma_{\mathsf{o}} = (0, T) \times \Gamma_{\mathsf{o}}$. 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

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

where V' denotes the dual of V. The space $W(0, T)$ is a Hilbert space supplied with the common inner product; cf. [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

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

where $u_{\mathsf{a}} = (u_{\mathsf{a}i})_{1 \le i \le m}$, $u_{\mathsf{b}} = (u_{\mathsf{b}i})_{1 \le i \le m} \in \mathbb{R}^m$ 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*

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

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

Assumption 2.1 *We assume that* $\gamma_o \ge 0$, $\boldsymbol{v} \in L^\infty(0,T;L^\infty(\Omega;\mathbb{R}^d))$ *with* $d \in \{1,2,3\}$, $y_{out} \in L^2(0,T)$, $y_\circ \in H$ *and* $b_1, \ldots, b_m \in L^\infty(\Gamma_c)$.

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

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

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

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

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

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

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

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

is satisfied.

Lemma 2.1 *Let Assumption 2.1 hold. Then:*

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

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

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

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

for a constant $c_y > 0$ which is independent of $y_\circ$, y_{out} 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 $\mathscr{F}$, and another part depending linearly on the control variable. Let $\hat{y} \in W(0,T)$ be the unique solution to the problem

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

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

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

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 $y = \hat{y} + \mathscr{S}u$. ◊

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

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

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 (virtual) control. Then, (4) is replaced by the mixed control-state constraints

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

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

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

Assumption 2.2 *Let the desired states satisfy* $y_Q \in L^2(0,T;H)$ *and* $y_T \in H$. *Furthermore,* $\varepsilon > 0$, σ_Q, $\sigma_T \geq 0$, *and* σ, $\sigma_w > 0$.

The optimal control problem is given by

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

Remark 2.2 Following [19] one can consider the generalized problem

$$\begin{aligned} \min \frac{\sigma_Q}{2} \int_0^T \|y(t) - y_Q(t)\|_H^2 \,\mathrm{d}t + \frac{\sigma_T}{2} \|y(T) - y_T\|_H^2 \\ + \frac{\sigma}{2} \sum_{i=1}^m \|u_i\|_{L^2(0,T)}^2 + \frac{f(\varepsilon)}{2} \|w\|_{\mathcal{W}}^2 \end{aligned} \tag{5a}$$

subject to the modified state equations

$$\begin{aligned} y_t(t,\boldsymbol{x}) - \Delta y(t,\boldsymbol{x}) + \boldsymbol{v}(t,\boldsymbol{x}) \cdot \nabla y(t,\boldsymbol{x}) &= g(\varepsilon) w && \text{a.e. in } Q, \\ \frac{\partial y}{\partial \boldsymbol{n}}(t,\boldsymbol{s}) + y(t,\boldsymbol{s}) &= \sum_{i=1}^m u_i(t) b_i(\boldsymbol{s}) && \text{a.e. on } \Sigma_{\mathsf{c}}, \\ \frac{\partial y}{\partial \boldsymbol{n}}(t,\boldsymbol{s}) + \gamma_{\mathsf{o}} y(t,\boldsymbol{s}) &= \gamma_{\mathsf{o}} y_{\mathsf{out}}(t) && \text{a.e. on } \Sigma_{\mathsf{o}}, \\ y(0,\boldsymbol{x}) &= y_\circ(\boldsymbol{x}), && \text{a.e. in } \Omega \end{aligned} \tag{5b}$$

and to the inequality constraints

$$\begin{aligned} u_{\mathsf{a}i} \leq u_i(t) \leq u_{\mathsf{b}i} && \text{a.e. in } [0,T] \text{ for } i = 1, \ldots, m, \\ y_{\mathsf{a}}(t,\boldsymbol{x}) \leq \mathscr{E}y(t,\boldsymbol{x}) + h(\varepsilon) w(t,\boldsymbol{x}) \leq y_{\mathsf{b}}(t,\boldsymbol{x}) && \text{a.e. in } Q, \end{aligned} \tag{5c}$$

where f, g and h are chosen nonnegative functions defined for $\varepsilon \geq 0$. In [19] convergence of a solution $\bar{x}^{\varepsilon} = (\bar{y}^{\varepsilon}, \bar{u}^{\varepsilon}, \bar{w}^{\varepsilon}) \in \mathcal{X}$ is proved for $\varepsilon \to 0$ in the case of an elliptic state equation and unilateral state constraints. In our future work we will study the application of the arguments in [19] to our parabolic setting and to bilateral state constraints. ◊

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:

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

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

$$\mathscr{T}_\varepsilon : \mathcal{Z} \to \mathcal{Z}, \quad z = (u, w) \mapsto \mathscr{T}_\varepsilon(z) = \begin{pmatrix} u \\ \mathscr{E}\mathscr{S}u + \varepsilon w \end{pmatrix} = \begin{pmatrix} \mathscr{I}_\mathcal{U} & 0 \\ \mathscr{E}\mathscr{S} & \varepsilon \mathscr{I}_\mathcal{W} \end{pmatrix} \begin{pmatrix} u \\ w \end{pmatrix}, \tag{6}$$

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

$$\mathscr{T}_\varepsilon^{-1}(\mathfrak{u}, \mathfrak{w}) = \begin{pmatrix} \mathscr{I}_\mathcal{U} & 0 \\ -\varepsilon^{-1}\mathscr{E}\mathscr{S} & \varepsilon^{-1}\mathscr{I}_\mathcal{W} \end{pmatrix} \begin{pmatrix} \mathfrak{u} \\ \mathfrak{w} \end{pmatrix} = \left(\mathfrak{u}, \frac{1}{\varepsilon}(\mathfrak{w} - \mathscr{E}\mathscr{S}\mathfrak{u}) \right) \tag{7}$$

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

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

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

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

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 2.1, 2.2 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 [11].

Theorem 2.2 *Let Assumptions 2.1 and 2.2 hold. Suppose that the feasible set $\mathcal{Z}_{ad}^\varepsilon$ is nonempty and that $\bar{z} = (\bar{u}, \bar{w}) \in \mathcal{Z}_{ad}^\varepsilon$ 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*

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}t}\langle \bar{p}(t),\varphi\rangle_H + a(t;\varphi,\bar{p}(t)) + \langle \bar{\beta}(t),\varphi\rangle_H &= \sigma_Q \langle (y_Q - \bar{y})(t),\varphi\rangle_H \ \forall \varphi \in V, \\ \bar{p}(T) &= \sigma_T\big(y_T - \bar{y}(T)\big) \qquad \text{in } H \end{aligned} \tag{8}$$

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

$$\begin{aligned} \sigma\bar{u}_i - \int_{\Gamma_c} b_i\,\bar{p}\,\mathrm{d}s + \bar{\mu}_i = 0 \quad &\text{in } L^2(0,T) \text{ for } i = 1,\dots,m, \\ \sigma_w\bar{w} + \varepsilon\bar{\beta} = 0 \quad &\text{in } \mathcal{W}. \end{aligned} \tag{9}$$

Moreover,

$$\bar{\beta} = \max\big\{0, \bar{\beta} + \eta(\bar{y} + \varepsilon\bar{w} - y_b)\big\} + \min\big\{0, \bar{\beta} + \eta(\bar{y} + \varepsilon\bar{w} - y_a)\big\}, \tag{10a}$$

$$\bar{\mu}_i = \max\big\{0, \bar{\mu}_i + \eta_i(\bar{u}_i - u_{bi})\big\} + \min\big\{0, \bar{\mu}_i + \eta_i(\bar{u}_i - u_{ai})\big\} \tag{10b}$$

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

Remark 2.3 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 Sect. 2.2. Let $\hat{p} \in W(0,T)$ denote the unique solution to the adjoint equation

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

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 $p = \mathscr{A}_1 u$ is the unique solution to

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

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

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

In particular, the solution $\bar{p}$ to (8) 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

$$\sigma \bar{u}_i - \gamma_{\mathsf{c}} \int_{\Gamma_{\mathsf{c}}} b_i \bar{p} \, \mathrm{d}s + \bar{\mu}_i = 0, \quad i = 1, \ldots, m, \tag{11a}$$

$$\sigma_w \bar{w} + \varepsilon \bar{\beta} = 0, \tag{11b}$$

$$\bar{\mu}_i = \max\left\{0, \bar{\mu}_i + \eta_i(\bar{u}_i - u_{\mathsf{b}i})\right\} + \min\left\{0, \bar{\mu}_i + \eta_i(\bar{u}_i - u_{\mathsf{a}i})\right\}, \tag{11c}$$

$$\bar{\beta} = \max\left\{0, \bar{\beta} + \eta(\bar{y} + \varepsilon \bar{w} - y_{\mathsf{b}})\right\} + \min\left\{0, \bar{\beta} + \eta(\bar{y} + \varepsilon \bar{w} - y_{\mathsf{a}})\right\} \tag{11d}$$

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

Remark 2.4 Optimality system (11) can also be expressed as a variational inequality; cf. [17, 25]. Since the admissible set $\mathcal{Z}_{\mathsf{ad}}^\varepsilon$ 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

$$\langle \nabla \hat{J}(\bar{z}), z - \bar{z} \rangle_{\mathcal{Z}} \geq 0 \quad \forall z \in \mathcal{Z}_{\mathsf{ad}}^\varepsilon, \tag{12}$$

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

$$\nabla \hat{J}(z) = \begin{pmatrix} \left(\sigma u_i - \langle b_i, p(\cdot) \rangle_{L^2(\Gamma_{\mathsf{c}})}\right)_{1 \leq i \leq m} \\ \sigma_w w \end{pmatrix} \tag{13}$$

with $p = \hat{p} + \mathscr{A}_1 u$. ◊

3 The Primal-Dual Active Set Strategy (PDASS)

To solve ($\hat{\mathbf{P}}^\varepsilon$) we utilize a semismooth Newton method which can be interpreted as a primal-dual active set strategy; cf. [15, 18, 27]. For more details we refer to [9, 11]. Suppose that $z^k = (u^k, w^k) \in \mathcal{Z}$ is a current iterate for $k \in \mathbb{N}_0$. Then, we set $y^0 = \hat{y} + \mathscr{S} u^0$, $p^0 = \hat{p} + \mathscr{A}_1 u^0 - \sigma_w \mathscr{A}_2 w^0 / \varepsilon$,

$$y^k = \hat{y} + \mathscr{S} u^k, \qquad \beta^k = -\frac{\sigma_w}{\varepsilon} w^k,$$

$$p^k = \hat{p} + \mathscr{A}_1 u^k + \mathscr{A}_2 \beta^k, \qquad \mu_i^k = \int_{\Gamma_{\mathsf{c}}} b_i p^k \, \mathrm{d}s - \sigma u_i^k \text{ for } i = 1, \ldots, m.$$

Now we define the associated active sets

$$\begin{aligned}
\mathcal{A}^{\mathcal{U}}_{\mathsf{a}i}(z^k) &= \big\{t \in [0,T] \,\big|\, \mu_i^k + \sigma(u_i^k - u_{\mathsf{a}i}) < 0 \text{ a.e.}\big\}, \quad i = 1,\dots,m,\\
\mathcal{A}^{\mathcal{U}}_{\mathsf{b}i}(z^k) &= \big\{t \in [0,T] \,\big|\, \mu_i^k + \sigma(u_i^k - u_{\mathsf{b}i}) > 0 \text{ a.e.}\big\}, \quad i = 1,\dots,m,\\
\mathcal{A}^{\mathcal{W}}_{\mathsf{a}}(z^k) &= \left\{(t,\boldsymbol{x}) \in Q \,\middle|\, \beta^k + \frac{\sigma_w}{\varepsilon^2}\big(y^k + \varepsilon w^k - y_{\mathsf{a}}\big) < 0 \text{ a.e.}\right\},\\
\mathcal{A}^{\mathcal{W}}_{\mathsf{b}}(z^k) &= \left\{(t,\boldsymbol{x}) \in Q \,\middle|\, \beta^k + \frac{\sigma_w}{\varepsilon^2}\big(y^k + \varepsilon w^k - y_{\mathsf{b}}\big) > 0 \text{ a.e.}\right\}.
\end{aligned} \tag{14a}$$

The associated inactive sets are defined as

$$\begin{aligned}
\mathcal{I}^{\mathcal{U}}_i(z^k) &= [0,T] \setminus \big(\mathcal{A}^{\mathcal{U}}_{\mathsf{a}i}(z^k) \cup \mathcal{A}^{\mathcal{U}}_{\mathsf{b}i}(z^k)\big) \quad \text{for } i = 1,\dots,m,\\
\mathcal{I}^{\mathcal{W}}(z^k) &= Q \setminus \big(\mathcal{A}^{\mathcal{W}}_{\mathsf{a}}(z^k) \cup \mathcal{A}^{\mathcal{W}}_{\mathsf{b}}(z^k)\big).
\end{aligned} \tag{14b}$$

Now it turns out that the new state y^{k+1} and the new adjoint p^{k+1} are given by the two coupled problems

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

and

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

respectively, which can be expressed as

$$\begin{pmatrix} \mathscr{A}^k_{11} & \mathscr{A}^k_{12}\\ \mathscr{A}^k_{21} & \mathscr{A}^k_{22}\end{pmatrix}\begin{pmatrix} y^{k+1}\\ p^{k+1}\end{pmatrix} = \begin{pmatrix} \mathscr{Q}_1(z^k; y_\circ, u_{\mathsf{a}}, u_{\mathsf{b}}, b_i, \sigma, \gamma_{\mathsf{c}}, y_{\mathsf{out}})\\ \mathscr{Q}_2(z^k; y_{\mathsf{a}}, y_{\mathsf{b}}, y_Q, y_T, \varepsilon, \sigma_w)\end{pmatrix}. \tag{15}$$

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 $\mathscr{A}\colon W(0,T) \to L^2(0,T,V')$ 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 $\varphi \in L^2(0,T;V)$. The new control variable $z^{k+1} = (u^{k+1}, w^{k+1})$ is given by the linear system

$$\begin{aligned}
\int_{\Gamma_c} b_i\, p^{k+1}\,\mathrm{d}s - \sigma u_i^{k+1} &= 0 \qquad \text{in } \mathcal{I}_i^{\mathcal{U}}(z^k), \quad i = 1,\dots,m,\\
u_i^{k+1} &= u_{\mathsf{a}i} \quad \text{in } \mathcal{A}_{\mathsf{a}i}^{\mathcal{U}}(z^k), \quad i = 1,\dots,m,\\
u_i^{k+1} &= u_{\mathsf{b}i} \quad \text{in } \mathcal{A}_{\mathsf{b}i}^{\mathcal{U}}(z^k), \quad i = 1,\dots,m,\\
w^{k+1} &= 0 \qquad \text{in } \mathcal{I}^{\mathcal{W}}(z^k),\\
y^{k+1} + \varepsilon\, w^{k+1} &= y_{\mathsf{a}} \quad \text{in } \mathcal{A}_{\mathsf{a}}^{\mathcal{W}}(z^k),\\
y^{k+1} + \varepsilon\, w^{k+1} &= y_{\mathsf{b}} \quad \text{in } \mathcal{A}_{\mathsf{b}}^{\mathcal{W}}(z^k).
\end{aligned} \tag{16}$$

We resume the previous strategy in Algorithm 1.

Remark 3.1 Algorithm 1 has to be discretized for their numerical realizations. In our tests carried out in Sect. 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 finite elements on a triangular mesh. ◊

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

1: Choose starting value $z^0 = (u^0, w^0) \in \mathcal{Z}$; set $k = 0$ and `flag = false`;
2: Determine $y^0 = \hat{y} + \mathscr{S}u^0$ and $p^0 = \hat{p} + \mathscr{A}_1 u^0 - \sigma_w \mathscr{A}_2 w^0/\varepsilon$;
3: **repeat**
4: Get $\mathcal{A}_{\mathsf{a}i}^{\mathcal{U}}(z^k)$, $\mathcal{A}_{\mathsf{b}i}^{\mathcal{U}}(z^k)$, $\mathcal{I}_i^{\mathcal{U}}(z^k)$, $i = 1,\dots,m$, and $\mathcal{A}_{\mathsf{a}}^{\mathcal{W}}(z^k)$, $\mathcal{A}_{\mathsf{b}}^{\mathcal{W}}(z^k)$, $\mathcal{I}^{\mathcal{W}}(z^k)$ from (14);
5: Compute the solution (y^{k+1}, p^{k+1}) by solving (15);
6: Compute $z^{k+1} = (u^{k+1}, w^{k+1}) \in \mathcal{Z}$ from (16);
7: Set $k = k + 1$;
8: **if** $\mathcal{A}_{\mathsf{a}1}^{\mathcal{U}}(z^k) = \mathcal{A}_{\mathsf{a}1}^{\mathcal{U}}(z^{k-1})$ **and** … **and** $\mathcal{A}_{\mathsf{a}m}^{\mathcal{U}}(z^k) = \mathcal{A}_{\mathsf{a}m}^{\mathcal{U}}(z^{k-1})$ **then**
9: **if** $\mathcal{A}_{\mathsf{b}1}^{\mathcal{U}}(z^k) = \mathcal{A}_{\mathsf{b}1}^{\mathcal{U}}(z^{k-1})$ **and** … **and** $\mathcal{A}_{\mathsf{b}m}^{\mathcal{U}}(z^k) = \mathcal{A}_{\mathsf{b}m}^{\mathcal{U}}(z^{k-1})$ **then**
10: **if** $\mathcal{A}_{\mathsf{a}}^{\mathcal{W}}(z^k) = \mathcal{A}_{\mathsf{a}}^{\mathcal{W}}(z^k)$ **and** $\mathcal{A}_{\mathsf{b}}^{\mathcal{W}}(z^k) = \mathcal{A}_{\mathsf{b}}^{\mathcal{W}}(z^{k-1})$ **then**
11: `flag = true`;
12: **end if**
13: **end if**
14: **end if**
15: **until** `flag = true`;

4 Proper Orthogonal Decomposition

For properly chosen admissible controls $z = (u, w) \in \mathcal{Z}_{\mathsf{ad}}^{\varepsilon}$ we set $y = \hat{y} + \mathscr{S}u$ and $p = \hat{p} + \mathscr{A}_1 u - \frac{\sigma_w}{\varepsilon}\mathscr{A}_2 w$. Then, we introduce the linear subspace

$$\mathcal{V} = \operatorname{span}\left\{ y(t),\, p(t) \,\middle|\, t \in [0,T] \right\} \subset V \tag{17}$$

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

$$y(t) = \sum_{i=1}^{\mathsf{d}} \langle y(t), \psi_i \rangle_V \, \psi_i \quad \text{and} \quad p(t) = \sum_{i=1}^{\mathsf{d}} \langle p(t), \psi_i \rangle_V \, \psi_i \quad \text{a.e. in } [0,T] \tag{18}$$

The method of proper orthogonal decomposition (POD) consist in choosing an orthonormal basis $\{\psi_i\}_{i=1}^{\mathsf{d}}$ in $\mathcal{V}$ such that for every $\ell \in \mathbb{N}$ with $\ell \leq \mathsf{d}$ the mean square error between the snapshots y, p and their corresponding ℓ-th partial sum of (18) is minimized:

$$\begin{aligned} &\min \int_0^T \left\| y(t) - \sum_{i=1}^{\ell} \langle y(t), \psi_i \rangle_V \, \psi_i \right\|_V^2 + \left\| p(t) - \sum_{i=1}^{\ell} \langle p(t), \psi_i \rangle_V \, \psi_i \right\|_V^2 \mathrm{d}t \\ &\text{s.t. } \{\psi_i\}_{i=1}^{\ell} \subset V \text{ and } \langle \psi_i, \psi_j \rangle_V = \delta_{ij} \text{ for } 1 \leq i, j \leq \ell, \end{aligned} \tag{19}$$

where δ_{ij} is the Kronecker delta.

Definition 4.1 A solution $\{\psi_i\}_{i=1}^{\ell}$ to (19) is called a POD basis of rank ℓ. We define the subspace spanned by the first ℓ POD basis functions as $V^\ell = \operatorname{span}\{\psi_1, \ldots, \psi_\ell\}$.

Using a Lagrangian framework, the solution to (19) is characterized by the following optimality conditions (cf. [6, 13]):

$$\mathcal{R}\psi = \lambda\psi, \tag{20}$$

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

$$\mathcal{R}\psi = \int_0^T \langle y(t), \psi \rangle_V \, y(t) + \langle p(t), \psi \rangle_V \, p(t) \, \mathrm{d}t \quad \text{for } \psi \in V$$

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

$$\mathcal{R}\psi_i = \lambda_i \psi_i, \quad \lambda_1 \geq \cdots \geq \lambda_{\mathsf{d}} > 0 \quad \text{and} \quad \lambda_i = 0, \quad \text{for } i > d. \tag{21}$$

Moreover $\mathcal{V} = \operatorname{span}\{\psi_i\}_{i=1}^{\mathsf{d}}$. 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:

$$\int_0^T \Big\| y(t) - \sum_{i=1}^{\ell} \langle y(t), \psi_i \rangle_V \, \psi_i \Big\|_V^2 + \Big\| p(t) - \sum_{i=1}^{\ell} \langle p(t), \psi_i \rangle_V \, \psi_i \Big\|_V^2 \, \mathrm{d}t = \sum_{i=\ell+1}^{\mathsf{d}} \lambda_i .$$

Remark 4.1 For the numerical realization, the Hilbert space V has to be discretized by, e.g., piecewise finite elements and the integral over $[0, T]$ has to be replaced by a trapezoidal approximation; see [13]. ◊

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 $y^\ell = \mathscr{S}^\ell u \in W(0, T)$ is given by

$$\frac{\mathrm{d}}{\mathrm{d}t} \langle y^\ell(t), \psi \rangle_H + a(t; y^\ell(t), \psi) = \langle \mathscr{B}(u(t)), \psi \rangle_{V',V} \quad \forall \psi \in V^\ell \text{ a.e. in } (0, T]. \tag{22}$$

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.,[13]. 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

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

5 A-Posteriori Error Analysis

In this section we present an a-posteriori error estimate which is based on a perturbation argument [8] and has been already utilized in [26]. As done in [9], this estimate can be generalized for the mixed control-state constraints case. As first, suppose that Assumptions 2.1 and 2.2 hold. Recall that the linear, invertible operator $\mathscr{T}_\varepsilon$ has been introduced in (6). 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 parameter ε. Now, we consider instead of ($\hat{\mathbf{P}}^\varepsilon$) the following optimal control problem

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

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 ($\hat{\mathbf{P}}^\varepsilon$). Conversely, if $\bar{\mathfrak{z}}$ solves ($\hat{\mathbf{P}}^\varepsilon$), then $\bar{z} = \mathscr{T}_\varepsilon^{-1}(\bar{\mathfrak{z}})$ is the solution to ($\hat{\mathbf{P}}^\varepsilon$). According to [9] we have the following result:

Theorem 5.1 *Suppose that Assumptions 2.1 and 2.2 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 ($\hat{\mathbf{P}}^\varepsilon$).*
2) *Suppose that a point $\mathfrak{z}^{ap} = (\mathfrak{u}^{ap}, \mathfrak{w}^{ap}) \in \mathfrak{Z}_{ad}$ is computed. We set $z^{ap} = \mathscr{T}_\varepsilon^{-1}(\mathfrak{z}^{ap})$, i.e., $z^{ap} = (u^{ap}, w^{ap})$ fulfills $u^{ap} = \mathfrak{u}^{ap}$ and $w^{ap} = \varepsilon^{-1}(\mathfrak{w}^{ap} - \mathscr{E}\mathscr{S}\mathfrak{u}^{ap})$. Then, there exists a perturbation $\zeta = (\zeta^u, \zeta^w) \in \mathcal{Z}$, which is independent of $\bar{z}$, so that*

$$\|\bar{z} - z^{ap}\|_{\mathcal{Z}} \le \frac{1}{\sigma_z} \|\mathscr{T}_\varepsilon^\star \zeta\|_{\mathcal{Z}} \quad \text{with } \sigma_z = \min\{\sigma, \sigma_w\} > 0. \tag{23}$$

where $\mathscr{T}_\varepsilon^\star$ denotes the adjoint of the operator $\mathscr{T}_\varepsilon$; cf. (7).

Proof Since $\mathscr{T}_\varepsilon$ has a bounded inverse, part 1) follows. The second claim can be shown by adapting the proof of Proposition 1 in [9].

Remark 5.1

1) The perturbation ζ can be computed following [9, Section 1.5].
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, we proceed as in [12, 26] and utilize (23) as an a-posteriori error estimate in the following manner: We set

$$\mathfrak{z}^{\mathsf{ap}} = (\mathfrak{u}^{\mathsf{ap}}, \mathfrak{w}^{\mathsf{ap}}) \in \mathcal{Z} \quad \text{with} \quad \mathfrak{u}^{\mathsf{ap}} = \bar{u}^\ell \text{ and } \mathfrak{w}^{\mathsf{ap}} = \varepsilon \bar{w}^\ell + \mathscr{E}\mathscr{S}^\ell \bar{u}^\ell. \tag{24}$$

From $\bar{z}^\ell \in \mathcal{Z}_{\mathsf{ad}}^{\varepsilon,\ell}$ we infer that $\mathfrak{z}^{\mathsf{ap}} \in \mathfrak{Z}_{\mathsf{ad}}$. It follows from (7) and (24) that

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

fulfills (23). Moreover, we found that

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

Consequently, (23) 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 @ 2.60 GHz and 12 GB RAM. The codes are written in C language and we use the tools of PETSc, [3, 4], and SLEPc, [14, 23], for our numerical computations. In the tests we apply a discrete variant of Algorithm 1. For solving the linear system in step 5 of Algorithm 1, 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 supposed to have four 'heaters', which we call controls for simplicity, placed as shown in Fig. 1, with the following shape functions:

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

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

We choose the physical parameter $\gamma_0 = 0.03$ and as initial condition $y_\circ(x) = |\sin(2\pi x_1)\cos(2\pi x_2)|$ for $x = (x_1, x_2) \in \Omega$, as 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:

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

and

$$\mathcal{V}_{\mathcal{F}_1} = \{x = (x_1, x_2) \,|\, 12x_2 + 4x_1 \ge 3, 12x_2 + 4x_1 \le 13\},$$
$$\mathcal{V}_{\mathcal{F}_2} = \{x = (x_1, x_2) \,|\, x_1 + x_2 \ge 0.5, x_1 + x_2 \le 1.5\}.$$

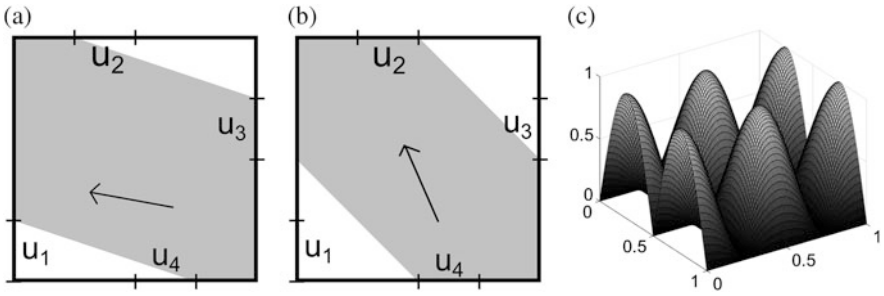

Fig. 1 Spatial domain Ω with the four boundary controls and the velocity fields (grey); initial condition $y_\circ(x)$. (**a**) $t < 0.5$. (**b**) $t \ge 0.5$. (**c**) $y_\circ(x)$

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 $y_T(\boldsymbol{x}) = y_Q(T, \boldsymbol{x})$, as state constraints $y_{\mathsf{a}}(t) = 0.5 + \min(2t, 2.0)$ and $y_{\mathsf{b}} = 3.0$. 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 $u_{\mathsf{a}i} = 0$ and $u_{\mathsf{b}i} = 7$ for $i = 1, \ldots, 4$. 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: Economic Optimal Control

The cost functional weights are $\sigma_T = \sigma_Q = 0$ and $\sigma_w = \sigma = 1$. This choice is motivated by economic optimal 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 [10, Chapter 8], for instance. In this test, as first, we study the behaviour of the PDASS for different values of ε. We will then analyse how this regularization parameter influences the POD approximation and the tightness of the error estimator. Finally, we will compare the POD-M1 and POD-M2 approximation for a fixed value of ε. As can be seen from Fig. 2 and as expected, when ε decreases the minimum temperature in the room gets progressively close to the lower constraints $y_a(t)$ at each time instance, while the average temperature and the maximum one remain for more time inside the constraints' range. The gradual decay of the temperature at the last time steps is due to the terminal condition for the dual variable p: from (8), since in this test $\sigma_T = 0$ holds, we have that $p(T) = 0$. Therefore, the computation of the optimal control, which is affected by this condition, lead to the previously noticed phenomena. As reported in Table 1, the number of PDASS iterations increases when ε decreases: when ε is small, the virtual control w is big in the active points, thus the algorithm employs more iterations to minimize the cost functional, where u and w have the same weights, respecting also the control constraints. It can be shown that $\varepsilon w = (y_{\mathsf{a}} - y)\, \chi_{\mathcal{A}^w_{\mathsf{a}}(z)} + (y_{\mathsf{b}} - y)\, \chi_{\mathcal{A}^w_{\mathsf{b}}(z)}$ holds, hence, the L^2-norm of εw can be used to measure how much the constraints are violated during all the evolution of the solution. As can be seen from Table 1, this value confirms what we already stated commenting Fig. 2. In Table 2, the relative errors between the solution computed with the POD-M2 approximation and the FE one are reported for the same number

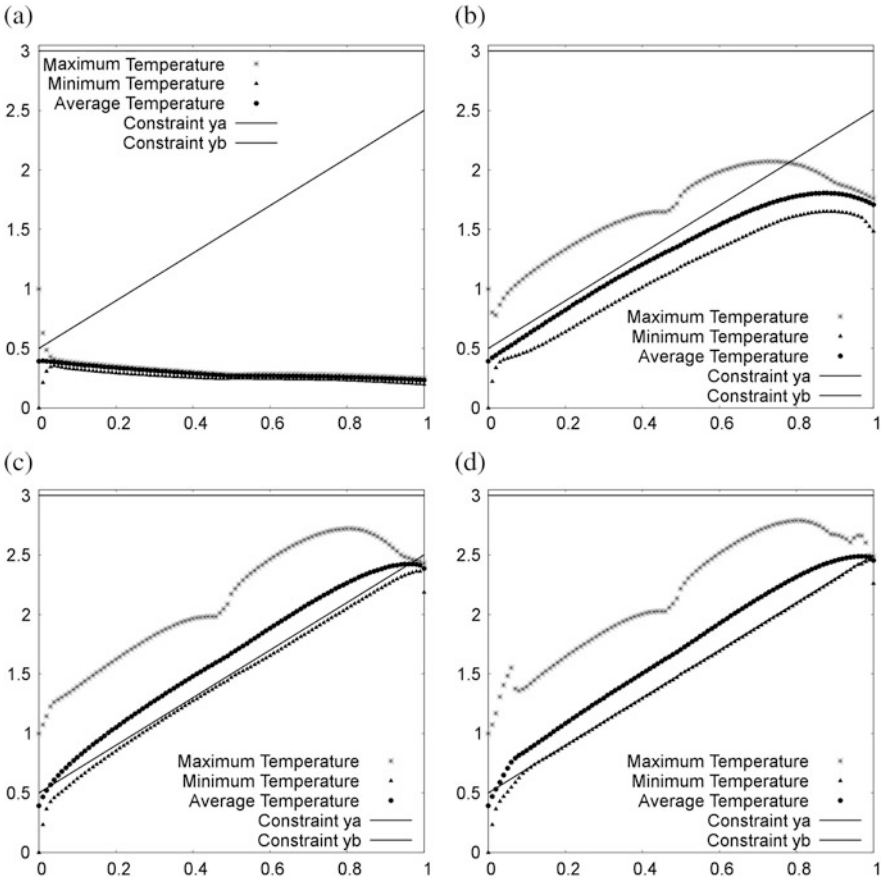

Fig. 2 Test 1: Temperature behaviour at each time-step for different ε. (**a**) $\varepsilon = 1$. (**b**) $\varepsilon = 0.1$. (**c**) $\varepsilon = 0.01$. (**d**) $\varepsilon = 0.001$

Table 1 Test 1: results for the FE discretization for different ε

Spatial discretization	ε	$\hat{J}(z)$	$\|\varepsilon w\|_{\mathcal{W}}$	Iterations
FE	1.0	0.931	1.3563	4
FE	0.1	7.584	0.2874	7
FE	0.01	9.066	0.0216	9
FE	0.001	120.329	0.0150	21

of basis and for different ε. In the last column, we have listed the values of the a-posteriori estimate for the difference $\|u^{\mathrm{FE}} - u^{\mathrm{POD}}\|$, which is defined as

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

Table 2 Test 1: results for the POD-M2 discretization for different ε and same number of basis

Spatial discretization	ε	rel-err(T)	rel-err	rel-err(Act.S.)	$\|u^{\mathrm{FE}} - u^{\mathrm{POD}}\|$	Err.Est.
POD-M2-10 Basis	1.000	0.002	0.003	0	0.0003	0.0004
POD-M2-10 Basis	0.100	0.006	0.004	0.001	0.0076	0.0167
POD-M2-10 Basis	0.010	0.004	0.007	0.024	0.3705	3.3604
POD-M2-10 Basis	0.001	0.700	0.648	0.465	7.359	$\simeq 2 \cdot 10^5$

We also need to clarify how we have computed the relative errors:

$$\begin{aligned}
\text{rel-err}(T) &= \|y^{\mathrm{FE}}(T) - y^{\mathrm{POD}}(T)\|_H / \|y^{\mathrm{FE}}(T)\|_H, \\
\text{rel-err} &= \|y^{\mathrm{FE}} - y^{\mathrm{POD}}\|_{L^2(0,T;H)} / \|y^{\mathrm{FE}}\|_{L^2(0,T;H)}, \\
\text{rel-err(Act.S.)} &= \left| \mathcal{A}^{\mathrm{FE}} \cup \mathcal{A}^{\mathrm{POD}} - \mathcal{A}^{\mathrm{FE}} \cap \mathcal{A}^{\mathrm{POD}} \right| / \left(N_x N_t \right),
\end{aligned}$$

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. The rel-err(Act.S.) in particular points out how much the active sets of state constraints related to the optimal solution computed with the reduced order model are far to the one computed in the FE discretization. As one can see, the POD approximation gets worse as ε decreases. For example, for $\varepsilon = 0.001$ the optimal control computed with the reduced order model is completely far from the one computed with the full order discretization. This is justified from the fact that there are more dynamics to approximate for smaller ε, since the number of iterations of the PDASS algorithm is greater. If we want to obtain, for example, an approximation error less than 0.01 in the case of POD-M2 we have to take at least 4 basis for $\varepsilon = 1$, 9 for $\varepsilon = 0.1$, 28 for $\varepsilon = 0.01$ and 58 for $\varepsilon = 0.001$. In addition, since in Theorem 5.1 w^{ap} depends on ε^{-1} and therefore also the error estimator, we have that its tightness depends on the regularization parameter. The previous statement is confirmed by the data reported in Table 2: the greater is ε the tighter is the error estimator. For example, for $\varepsilon = 1$ we have that it is only 1.3 times greater than the true error, instead it is 5.67 times the true one for $\varepsilon = 0.01$. From now to the end of the subsection, ε is fixed to 0.01. In Table 3 we present some results for Algorithm 1 for the FE and POD approximations using the two different strategies to build the POD bases. The norm of εw and also the cost functional gets closer to their values computed through the FE discretization as soon as the number of basis increases. Moreover, the PDASS algorithm applied to the reduced system converges almost in the same iterations' number of the full one. Even if we are able to solve the reduced linear system of Algorithm 1 around 80–100 times faster than the full one, the total algorithm speed-up is approximatively 4. This is due to the fact that we have to compute the active sets for the state constraints at each algorithm's iteration and this means that the reduced algorithm has to project into the FE discretization the approximated POD solution, compute the active sets, which costs $O(NN_t)$, and project back into the POD subspace those sets. To better compare POD-M1 and POD-M2 approaches, we

Table 3 Test 1: results for the FE and POD discretizations for $\varepsilon = 0.01$

Spatial discretization	POD basis elements	$\hat{J}(z)$	$\|\varepsilon w\|_{\mathcal{W}}$	rel-err(Act.S.)	Iterations	Speed-up
FE	–	9.066	0.0216	–	9	–
POD-M1	10	9.659	0.0339	0.127	10	3.91
POD-M1	15	9.123	0.0223	0.019	10	3.58
POD-M1	20	9.119	0.0221	0.010	9	3.48
POD-M2	10	9.181	0.0252	0.024	9	4.01
POD-M2	15	9.090	0.0229	0.014	9	3.90
POD-M2	20	9.076	0.0218	0.003	9	3.45

Table 4 Test 1: error values for the POD suboptimal solutions

Spatial discretization	POD basis elements	rel-err(T)	rel-err	$\|u^{\mathrm{FE}} - u^{\mathrm{POD}}\|$	Error estimator
POD-M1	10	0.068	0.115	1.344	7.620
POD-M1	15	0.003	0.004	0.174	2.361
POD-M1	20	0.003	0.003	0.136	1.549
POD-M2	10	0.004	0.007	0.371	3.360
POD-M2	15	0.003	0.002	0.128	1.321
POD-M2	20	0.001	0.001	0.065	0.179

also report the relative errors between the solution computed with the full and the reduced systems in Table 4. From this table, as expected, we can notice that the POD basis generated with the optimal solution performs better than the other basis: when the algorithm is getting closer 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. 3, where we plot 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 closer to the FE optimal controls than the ones obtained using the POD-M1 reduced system. This explains why we need an a-posteriori error estimator for the POD basis: we can estimate the quality of our basis and we can decide to consider a greater number of basis or to generate new basis from a different initial control. In Fig. 4, we show the comparison between the true error $\|u^{\mathrm{FE}} - u^{\mathrm{POD}}\|$ and the a-posteriori error estimator. Due to the previous discussion on the quality of the POD approximation, we can notice that as expected it is tighter for POD-M2 than for POD-M1 and it becomes for both approximations tighter and smaller as soon as the number of POD basis increases, although with some oscillation.

6.2 Test 2: Cost of Tracking Type

For the second test, we fix $\varepsilon = 0.1$ and we use the same data of Test 1, except for the cost functional weights which are chosen in the following way: $\sigma_T = \sigma_Q = 1$

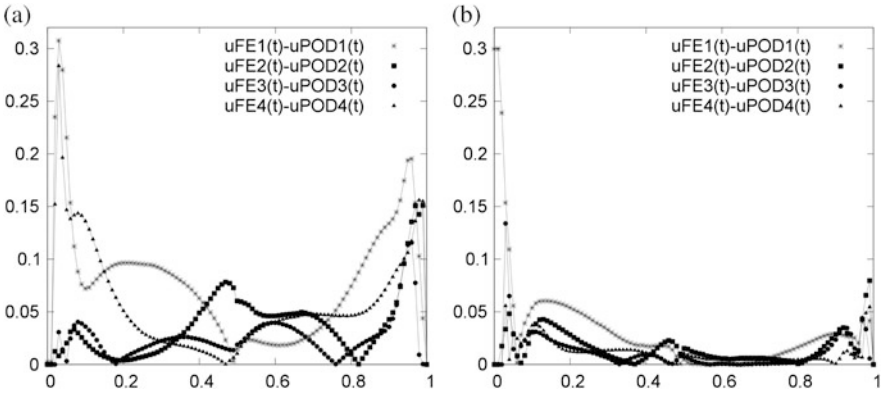

Fig. 3 Test 1: $|u^{FE}(t) - u^{POD}(t)|$ with $\ell = 20$ basis functions. (**a**) POD-M1. (**b**) POD-M2

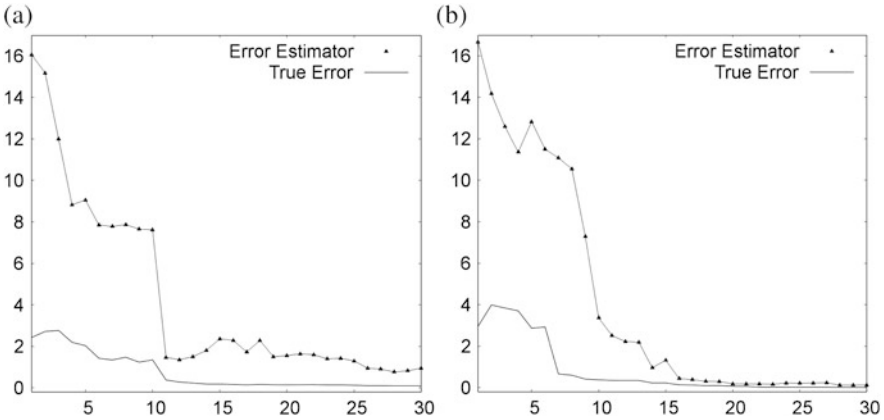

Fig. 4 Test 1: comparison between $\|u^{\mathrm{FE}} - u^{\mathrm{POD}}\|$ and its a-posteriori error estimate. (**a**) POD-M1. (**b**) POD-M2

and $\sigma = 0.01$. Regarding σ_w, we split the section in two parts: as first we study the model's behaviour when its value decreases, then we investigate the case $\sigma_w = 0$. Regarding this last condition, we want to point out that in the continuous model the terms connected to σ_w in the cost functional, adjoint equation and in the error estimator are zero: this means that w is not uniquely defined, since the only condition that w has to satisfy is $y_{\mathsf{a}}(t, \boldsymbol{x}) \leq \mathscr{E}y(t, \boldsymbol{x}) + \varepsilon w(t, \boldsymbol{x}) \leq y_{\mathsf{b}}(t, \boldsymbol{x})$ a.e. in Q, which clearly has no unique solution for fixed values of y_{a}, y_{b}, ε and y. By the way, due to the fact that $\sigma_w = 0$, we can observe that w is not more influencing the computation of the optimal control in the PDASS algorithm, so our optimal control will respect the control constraints and be the minimum of the reduced cost functional $\hat{J}$, but the solution may not be in the state constraints' range. From Table 5 we can noticed that the smaller σ_w is the more the algorithm focuses on reaching the target and the less on respecting the state constraints. In addition, when

Table 5 Test 2: results for the FE discretization for different σ_w

Spatial discretization	σ_w	$\hat{J}(z)$	$\|\varepsilon w\|_{\mathcal{W}}$	$\|y(T) - y_T\|$	$\|y - y_Q\|$	Iterations
FE	1.0000	0.318	0.015	0.159	0.618	9
FE	0.0100	0.311	0.036	0.156	0.624	5
FE	0.0001	0.309	0.161	0.155	0.623	4

Table 6 Test 2: results for the POD-M2 discretization for different σ_w and same number of basis

Spatial discretization	σ_w	rel-err(T)	rel-err	rel-err(Act.S.)	$\|u^{\text{FE}} - u^{\text{POD}}\|$	Err.Est.
POD-M2-10 Basis	1.0000	0.0019	0.0036	0.0014	0.1689	0.3051
POD-M2-10 Basis	0.0100	0.0013	0.0014	0.0007	0.0937	0.1456
POD-M2-10 Basis	0.0001	0.0013	0.0012	0.0005	0.0931	14.1898

σ_w decreases the conditions for the PDASS algorithm are less restrictive, therefore it uses less iteration to compute the solution. As showed in Table 6, also the POD-M2 approximation becomes better when σ_w gets smaller, but there is a worsening in the a-posteriori estimation: this is connected to the term σ_z^{-1} in (23), which makes the estimation increasing. For $\sigma_w = 0$ instead, we have a simplified error estimator, which produces better results compared to the case $\sigma_w > 0$ really small. As in Test 1, in Tables 7 and 8 we report the results of the finite elements solution (FE) and the reduced order ones (POD-M1,POD-M2) for $\sigma_w = 0$, with different choices of basis' number. As can be observed from Table 7, for this choice of parameters we have an improve of the speed-up gained in solving the reduced system, because in this context we do not have to compute the active sets for the state constraints. Therefore, we can have a speed-up for the algorithm similar to the one we get for solving the reduced linear system at each PDASS algorithm's step. In addition, the case $\sigma_w = 0.0001$ (or smaller) is equal to $\sigma_w = 0$, which is not surprising, since this means that already for this value of σ_w, we are almost ignoring the state constraints, due also to the choice on ε, but the advantage of taking $\sigma_w = 0$ is to have a tighter error estimator and a greater speed-up. As last, in this test it is confirmed that the number of POD basis functions needed to approximate the full order model really depends on the choice of the controls used for building the snapshots: we can see that for 4 basis, we can not capture in a good way the FE behaviour with POD-M1 basis, but with 10 basis we get results similar to POD-M2. The optimal trajectories at time $T = 1.0$ are reported in Fig. 5: we can notice that the FE and the POD-M2 ones are similar already for 7 basis, which is not the case for POD-M1.

7 Conclusions

With efficient building operation in mind, we have studied an optimal control problem of a parabolic convection-diffusion equations, with a time-dependent advection field, bilateral constraints for the boundary controls and pointwise state constraints, which have been treated with a Lavrentiev regularization. For solving

Table 7 Test 2: results for the FE and POD discretizations

Spatial discretization	POD basis elements	$\hat{J}(z)$	$\|y(T) - y_T\|$	$\|y - y_Q\|$	Iterations	Speed-up
FE	–	0.309	0.155	0.623	4	–
POD-M1	4	0.404	0.573	0.720	3	86.7
POD-M1	7	0.330	0.350	0.625	3	78.5
POD-M1	10	0.309	0.154	0.623	4	61.4
POD-M2	4	0.342	0.189	0.637	4	76.7
POD-M2	7	0.311	0.151	0.626	4	72.5
POD-M2	10	0.309	0.156	0.623	4	65.1

Table 8 Test 2: error values for the POD suboptimal solutions

Spatial discretization	POD basis elements	rel-err(T)	rel-err	$\|u^{\mathrm{FE}} - u^{\mathrm{POD}}\|$	Error estimator
POD-M1	4	0.091	0.087	1.711	6.170
POD-M1	7	0.056	0.026	0.421	0.781
POD-M1	10	0.002	0.002	0.103	0.166
POD-M2	4	0.044	0.062	1.416	5.770
POD-M2	7	0.004	0.004	0.200	0.379
POD-M2	10	0.001	0.001	0.093	0.142

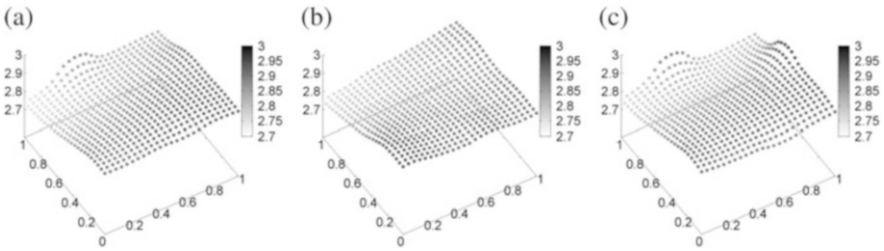

Fig. 5 Test 2: optimal trajectories at time $t = 1.0$. (**a**) FE. (**b**) POD-M1-7Basis. (**c**) POD-M2-7Basis

this optimal control problem we have applied the primal-dual active set strategy presented in [15], which has a super-linear rate of convergence. In order to speed-up the computational time of the algorithm, we have employed the POD method and utilized the a-posteriori error estimator in [9]. In the numerical test section, we have also shown how the variation of the regularization parameter ε and of the cost functional weight σ_w influences the behaviour of the solution and of the POD approximation. In addition, concerning the speed-up due to the POD method, we have noticed that this is reduced because of the computation of the state constraints' active sets, therefore it will be interesting in future work to treat the state constraints with other methods, e.g. the augmented Lagrangian algorithm. As shown in [22], 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.

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

References

1. Afanasiev, K., Hinze, M.: Adaptive control of a wake flow using proper orthogonal decomposition. In: Shape Optimization and Optimal Design. Lecture Notes in Pure and Applied Mathematics, vol. 216, pp. 317–332. Marcel Dekker, New York (2001)
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., 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, Basel (1997)
4. 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, Argonne (2016)
5. Banholzer, S., Beermann, D., Volkwein, S.: POD-based error control for reduced-order bicriterial PDE-constrained optimization. Annu. Rev. Control **44**, 226–237 (2017)
6. Berkooz, G., Holmes, P., Lumley, J.L.: Turbulence, Coherent Structures, Dynamical Systems and Symmetry. Cambridge Monographs on Mechanics. Cambridge University Press, Cambridge (1996)
7. Dautray, R., Lions, J.-L.: Mathematical Analysis and Numerical Methods for Science and Technology. Volume 5: Evolution Problems I. Springer, 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. Grimm, E., Gubisch, M., Volkwein, S.: Numerical analysis of optimality-system POD for constrained optimal control. In: Recent Trends in Computational Engineering - CE2014: Optimization, Uncertainty, Parallel Algorithms, Coupled and Complex Problems. Lecture Notes in Computational Science and Engineering, vol. 105, pp. 297–317. Springer, Cham (2015)
10. Grüne, L., Pannek, J.: Nonlinear Model Predictive Control: Theory and Algorithms, 2nd edn. Springer, London (2017)
11. 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)
12. 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)
13. Gubisch, M., Volkwein, S.: Proper orthogonal decomposition for linear-quadratic optimal control. In: Ohlberger, M., Benner, P., Cohen, A., Willcox, K. (eds.) Model Reduction and Approximation: Theory and Algorithms, pp. 5–66. SIAM, Philadelphia (2017)
14. Hernandez, V., Roman, J.E., Vidal, V.: SLEPc: a scalable and flexible toolkit for the solution of eigenvalue problems. ACM Trans. Math. Softw. **31**(3), 351–362 (2005). http://dx.doi.org/10.1145/1089014.1089019
15. Hintermüller, M., Ito, K., Kunisch, K.: The primal-dual active set strategy as a semismooth Newton method. SIAM J. Optim. **13**, 865–888 (2002)
16. 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)
17. Hinze, M. , Pinnau, R., Ulbrich, M., Ulbrich, S.: Optimization with PDE Constraints. Springer, Berlin (2009)
18. Ito, K., Kunisch, K.: Lagrange Multiplier Approach to Variational Problems and Applications. SIAM, Philadelphia (2008)
19. Krumbiegel, K., Rösch, A.: A virtual control concept for state constrained optimal control problems. Comput. Optim. Appl. **43**, 213–233 (2009)
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: Radu, F.A., Kumar, K., Berre, I., Nordbotten, J.M., Pop, I.S. (eds.) Numerical Mathematics and Advanced Applications ENUMATH 2017. Springer (2018)

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 (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)