POD-Based Mixed-Integer Optimal Control of Evolution Systems

Christian Jäkle and Stefan Volkwein(✉)

Department of Mathematics and Statistics, University of Konstanz,
Universitätsstraße 10, 78457 Konstanz, Germany
{Christian.Jaekle,Stefan.Volkwein}@uni-konstanz.de

Abstract. In this chapter the authors consider the numerical treatment of a mixed-integer optimal control problem governed by linear convection-diffusion equations and binary control variables. Using relaxation techniques (introduced by [31] for ordinary differential equations) the original mixed-integer optimal control problem is transferred into a relaxed optimal control problem with no integrality constraints. After an optimal solution to the relaxed problem has been computed, binary admissible controls are constructed by a sum-up rounding technique. This allows us to construct – in an iterative process – binary admissible controls such that the corresponding optimal state and the optimal cost value approximate the original ones with arbitrary accuracy. However, using finite element (FE) methods to discretize the state and adjoint equations often yield to extensive systems which make the frequently calculations time-consuming. Therefore, a model-order reduction based on the proper orthogonal decomposition (POD) method is applied. Compared to the FE case, the POD approach yields to a significant acceleration of the CPU times while the error stays sufficiently small.

Keywords: Mixed-integer optimal control · Integer programming · Relaxation methods · Evolution problems · Proper orthogonal decomposition

1 Introduction

A simplified optimal control problem is considered which is motivated by energy efficient building operation. The goal is to reach a certain desired temperature distribution in a room while choosing an optimal (underfloor) heating strategy. The temperature is governed by a heat equation with convection which extends our results in [3,4], where no convection was involved in the modeling of the heat transfer. Since the heating is described by a time-depending discrete control, the optimization problem involves continuous and discrete variables. These kinds of problems are considered in [10,11,22,32]. For partial differential equations (PDEs) we refer to the note [24]. In particular, mixed-integer problems for

O. Junge et al. (Eds.): SON 2020, SSDC 304, pp. 238–264, 2020.
https://doi.org/10.1007/978-3-030-51264-4_10

hyperbolic PDEs are considered, e.g., for problems in gas transportation systems [15], electric transmission lines [12] and traffic flow [16,17].

Frequently, integer problems are solved with the branch-and-bound method (see, e.g., [5]) to guarantee global optimality. Especially for finite-dimensional linear integer programming, the branch-and-bound method is the method of choice. However, this is often not possible or very expensive for optimal control problems, where infinite-dimensional control spaces are involved. Therefore, methods are used to approximate an optimal integer solution by sufficiently accurate solutions, which can be computed by techniques from infinite-dimensional optimization. In this work we apply relaxation methods which can be found in [31] for the case of ordinary differential equations and in [16,17] for the case of PDEs. To solve the relaxed optimal control problems, we rely on techniques from PDE-constrained optimization ([25,36]). Utilizing sum-up-rounding strategies (introduced by Sager in [31,33]) we construct discrete controls from the continuous ones; see also [26,27].

To speed-up the numerical solution of the relaxed optimal control problems we apply reduced-order modeling; cf. [1,34], for instance. In this work the relaxed optimal control problems are solved by POD Galerkin projection methods; cf. [14,19]. The POD method is known to be very efficient for dynamical systems. An POD a-posteriori error analysis – developed in [35] for optimal control problems – is extended in such a way that the error of the computed suboptimal POD solution can be controlled. This leads to an efficient and a certified optimization method which is also analytically based on theoretical results in [17].

Let us mention that there are other reduced-order approaches available, e.g., the reduced basis methods; cf. [13]. Especially for non-linear problems, the proper orthogonal decomposition (POD) method is a popular and widely used method. Here, predefined points in time are considered by a previously released dynamic system to build up the so-called snapshot space. The leading eigenfunctions of a singular value decomposition are then chosen as the basis for the reduced space, see for example [14]. It has been shown, that this method has good properties in the context of optimal control problems, especially thanks to an available a-posteriori estimate, see [23,35].

The chapter is organized as follows: In Sect. 2 the mixed-integer optimal control problem is introduced. Its relaxation is explained in Sect. 3. The numerical solution approach is described in Sect. 4 and Sect. 5 is devoted to present numerical results. Finally, we draw some conclusions in Sect. 6.

2 Problem Formulation

Let $\Omega \subset \mathbb{R}^n$, $n \in \{1,2,3\}$, be a bounded domain with Lipschitz-continuous boundary $\Gamma = \partial\Omega$. For $T > 0$ we set $Q = (0,T) \times \Omega$ and $\Sigma = (0,T) \times \Gamma$. Moreover, let H and V denote the standard real and separable Hilbert spaces $L^2(\Omega)$ and $H^1(\Omega)$, respectively, endowed with the 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 associated induced norms. For more details on Lebesgue and Sobolev spaces we refer to [9]. Recall the Hilbert space $W(0,T) = \{\varphi \in L^2(0,T;V) \,|\, \varphi_t \in L^2(0,T;V')\}$ endowed with the common inner product [8, pp. 472–479]. It is well-known that $W(0,T)$ is continuously embedded into $C([0,T];H)$, the space of continuous functions from $[0,T]$ to H. When t is fixed, the expression $\varphi(t)$ stands for the function $\varphi(t,\cdot)$ considered as a function in Ω only.

In this work we consider the following mixed-integer optimal control problem:

$$\min_{(y,u)} J(y,u) = \frac{1}{2}\int_0^T \int_\Omega |y(t,\boldsymbol{x}) - y^d(t,\boldsymbol{x})|^2 \,\mathrm{d}\boldsymbol{x}\mathrm{d}t + \frac{\gamma}{2}\sum_{i=1}^m \int_0^T |u_i(t)|^2 \,\mathrm{d}t \tag{1a}$$

subject to a convection-diffusion equation

$$y_t(t,\boldsymbol{x}) - \Delta y(t,\boldsymbol{x}) + v(\boldsymbol{x}) \cdot \nabla y(t,\boldsymbol{x}) = f(t,\boldsymbol{x}) + \sum_{i=1}^m u_i(t) b_i(\boldsymbol{x}), \qquad (t,\boldsymbol{x}) \in Q, \tag{1b}$$

$$\frac{\partial y}{\partial n}(t,\boldsymbol{s}) + q(\boldsymbol{s}) y(t,\boldsymbol{s}) = g(t,\boldsymbol{s}), \qquad (t,\boldsymbol{s}) \in \Sigma, \tag{1c}$$

$$y(0,\boldsymbol{x}) = y_\circ(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega \tag{1d}$$

and binary control constraints

$$u(t) \in \{0,1\}^m = \{u^i\}_{i=1}^N \quad \text{in } [0,T] \text{ a.e. (almost everywhere)}, \tag{1e}$$

where the u^i's are 0–1-vectors in $\mathbb{R}^m$ and $N = 2^m$ holds.

The desired temperature fulfills $y^d \in L^\infty(Q)$. For the regularization parameter we have $\gamma > 0$. The convection field v is supposed to be in $L^\infty(\Omega;\mathbb{R}^n)$. The heat source function satisfied $f \in C(\overline{Q})$. For $m \in \mathbb{N}$ we assume that the control shape functions fulfill $b_1,\ldots,b_m \in C(\overline{\Omega})$ and $b_i \geq 0$ on Ω a.e., but at least for one $i \in \{1,\ldots,m\}$ it holds $b_i > 0$ on Ω a.e. The isolation function satisfies $q \in L^\infty(\Gamma)$ with $q \geq 0$ on Γ a.e. The outer temperature is described by g and belongs to $C(\overline{\Sigma})$. Finally, for the initial temperature distribution we have $y_\circ \in C(\overline{\Omega})$.

Since we are interested in weak solutions to the state equation (1b)–(1d), we recall this solution concept for our case: A solution $y \in W(0,T)$ to (1b)–(1d) is understood as a weak solution, i.e., y belongs to $W(0,T)$ and satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle y(t),\varphi\rangle_H + a(y(t),\varphi) = \langle \mathcal{F}(t,u(t)),\varphi\rangle_{V',V} \quad \text{for all } \varphi \in V \text{ in } (0,T], \tag{2a}$$

$$\langle y(0),\varphi\rangle_H = \langle y_\circ,\varphi\rangle_H \quad \text{for all } \varphi \in V, \tag{2b}$$

where the bilinear form $a : V \times V \to \mathbb{R}$ is defined as

$$a(\varphi,\phi) = \int_\Omega \nabla\varphi \cdot \nabla\phi \,\mathrm{d}\boldsymbol{x} + \int_\Omega (v \cdot \nabla\varphi)\phi \,\mathrm{d}\boldsymbol{x} + \int_\Gamma q\varphi\phi \,\mathrm{d}\boldsymbol{s} \quad \text{for } \varphi,\phi \in V$$

and the inhomogeinity $\mathcal{F} : [0,T] \times \mathbb{R}^m \to V'$ is given by

$$\langle \mathcal{F}(t,u),\varphi\rangle_{V',V} = \int_\Omega \Big(f(t) + \sum_{i=1}^m u_i b_i\Big)\varphi \,\mathrm{d}\boldsymbol{x} + \int_\Gamma g(t)\varphi \,\mathrm{d}\boldsymbol{s}$$

for $(t, u) \in [0, T] \times \mathbb{R}^m$, $u = (u_i)_{1 \leq i \leq m}$ and $\varphi \in V$. Note that the mapping $\mathcal{F}(\cdot\,, u)$ is continuous for every $u \in \mathbb{R}^m$. The next proposition follows from the results in [20, Chapter 5].

Proposition 1. *Under the above assumptions on the data the following properties hold:*

1) *The bilinear form $a(\cdot\,,\cdot)$ is continuous and coercive, i.e., there are constants $\eta \geq 0$, $\eta_1 > 0$ and $\eta_2 \geq 0$ satisfying*

$$|a(\varphi, \phi)| \leq \eta \, \|\varphi\|_V \|\phi\|_V \qquad \text{for all } \varphi, \phi \in V,$$
$$|a(\varphi, \varphi)| \geq \eta_1 \, \|\varphi\|_V^2 - \eta_2 \, \|\varphi\|_H^2 \qquad \text{for every } \varphi \in V.$$

2) *For any $u \in U = L^2(0, T; \mathbb{R}^m)$ there exists a unique solution $y \in W(0, T)$ to (2) that satisfies*

$$\|y\|_{W(0,T)} \leq C\big(\|\mathcal{F}(\cdot\,, u(\cdot))\|_{C([0,T];V')} + \|y_\circ\|_H\big)$$

for a constant $C > 0$.

Remark 1. The bilinear form $a(\cdot\,,\cdot)$ defines a bounded linear operator $\mathcal{A} : V \to V'$ by

$$\langle \mathcal{A}\varphi, \phi \rangle_{V',V} = a(\varphi, \phi) \quad \text{for } \varphi, \phi \in V.$$

Furthermore, the operator $\mathcal{A}$ can also be considered as an unbounded operator on H with domain $\mathscr{D}(\mathcal{A}) = H^2(\Omega) \cap V \cap C(\overline{\Omega})$ which is dense in $C(\overline{\Omega})$. The operator $-\mathcal{A}$ generates a C_0-semigroup on $C(\overline{\Omega})$ and the solution y to (2) belongs to $W(0, T) \cap C(\overline{Q})$; cf. [29, Chapter 5]. Utilizing the continuity assumptions for f, $b_1, \ldots, b_m$, g and $y_\circ$ we can write (2) as the Cauchy problem

$$\dot{y}(t) = -\mathcal{A}y(t) + \mathcal{F}(t, u(t)) \text{ for } t \in (0, T], \quad y(0) = y_\circ$$

posed in $C(\overline{\Omega})$. It is proved in [28, Theorem 4.3] that $-\mathcal{A}$ also generates a holomorphic semigroup on $C(\overline{\Omega})$. ◊

Throughout this work the binary problem (1a)–(1e) is called **(BN)**. Its cost value at an admissible solution is denoted by J^{BN}. Furthermore, we introduce a relaxed problem, where (1e) is replaced by the relaxation

$$u(t) \in [0, 1]^m \text{ in } [0, T] \text{ a.e.,} \tag{1e'}$$

Problem (1a)–(1d) together with (1e') is denoted by **(RN)**. We write J^{RN} for the objective value obtained by an admissible solution for **(RN)**. Let us mention that (1a)–(1d) together with (1e') does not involve any integrality constraints. Thus, solution methods from continuous optimization can be applied.

3 Relaxation Method

Commonly, mixed-integer problems are solved with the branch-and-bound method (see e.g. [5]) to guarantee global optimality. However, for optimal control problems this is often computationally too expensive. In order to get an optimal control problem without any integer restrictions we apply therefore the approach in [17] which leads to convexified relaxed problems that can be solved by available techniques from PDE-constrained optimization; see [18,35], for instance.

3.1 Convexification

Using (1e) we introduce the following representation of the control variable

$$\beta(t) \in \{0,1\}^N, \ \sum_{i=1}^N \beta_i(t) = 1 \text{ and } u(t) = \sum_{i=1}^N \beta_i(t) u^i \quad \text{for } t \in [0,T].$$

To solve our mixed-integer optimal control problem **(BN)** we consider the following convexification (cf. [17, Section 2])

$$\min_{(y_\beta,\beta)} \frac{1}{2} \int_0^T \int_\Omega |y_\beta(t,\boldsymbol{x}) - y^d(t,\boldsymbol{x})|^2 \,\mathrm{d}\boldsymbol{x}\mathrm{d}t + \frac{\gamma}{2} \sum_{i=1}^N \|u^i\|^2_{\mathbb{R}^m} \int_0^T \beta_i(t)\,\mathrm{d}t \tag{3a}$$

subject to

$$\frac{\mathrm{d}}{\mathrm{d}t} \langle y_\beta(t), \varphi\rangle_H + a(y_\beta(t),\varphi) = \sum_{i=1}^N \beta_i(t) \,\langle \mathcal{F}(t,u^i),\varphi\rangle_{V',V} \quad \forall \varphi \in V \text{ in } (0,T], \tag{3b}$$

$$\langle y_\beta(0),\varphi\rangle_H = \langle y_\circ,\varphi\rangle_H \qquad \forall\varphi \in V, \tag{3c}$$

$$\beta(t) = \big(\beta_i(t)\big)_{1\le i\le N} \in \{0,1\}^N \qquad \text{in } [0,T], \tag{3d}$$

$$\sum_{i=1}^N \beta_i(t) = 1 \qquad \text{in } [0,T]. \tag{3e}$$

The convexification (3) of **(BN)** is called **(BL)** and we write J^{BL} for the objective value obtained by an admissible solution. Of course, **(BL)** still contains the integrality constraint (3d). Therefore, we introduce its relaxation – that we call **(RL)** – by

$$\min_{(y_\alpha,\alpha)} \frac{1}{2} \int_0^T \int_\Omega |y_\alpha(t,\boldsymbol{x}) - y^d(t,\boldsymbol{x})|^2 \,\mathrm{d}\boldsymbol{x}\mathrm{d}t + \frac{\gamma}{2} \sum_{i=1}^N \|u^i\|^2_{\mathbb{R}^m} \int_0^T \alpha_i(t)\,\mathrm{d}t \tag{4a}$$

subject to

$$\frac{\mathrm{d}}{\mathrm{d}t} \langle y_\alpha(t), \varphi\rangle_H + a(y_\alpha(t),\varphi) = \sum_{i=1}^N \alpha_i(t) \langle \mathcal{F}(t,u^i),\varphi\rangle_{V',V} \quad \forall \varphi \in V \text{ in } (0,T], \tag{4b}$$

$$\langle y_\alpha(0),\varphi\rangle_H = \langle y_\circ,\varphi\rangle_H \qquad \forall\varphi \in V, \tag{4c}$$

$$\alpha(t) = \big(\alpha_i(t)\big)_{1\le i\le N} \in [0,1]^N \qquad \text{in } [0,T], \tag{4d}$$

$$\sum_{i=1}^N \alpha_i(t) = 1 \qquad \text{in } [0,T]. \tag{4e}$$

We write J^{RL} for the objective value obtained by an admissible solution.

In the following theorem we show that the convexification does not change the optimal values of the original problem **(BN)** and the convexified problem **(BL)**. The proof is similar to the one in [31, Theorem 4.6] and therefore adapted from there.

Theorem 1. *If the convexified binary optimal control problem* **(BL)** *has an optimal solution* (y^*_β, β^*) *with objective value* J^{BL}*, then there exists an* m*-dimensional control function* u^* *such that* (y^*, u^*) *is an optimal solution of the binary optimal control problem* **(BN)** *with objective value* J^{BN} *satisfying* $J^{BL} = J^{BN}$*. The converse holds true as well.*

Proof. Assume that (y^*_β, β^*) is a minimizer of **(BL)**. Since it is feasible we have the special order set property (3e) with $\beta^*_i(\cdot) \in \{0, 1\}$ for all $i = 1, \ldots, N$. Thus, there exists one index $1 \leq j(t) \leq N$ for almost all (f.a.a.) $t \in [0, T]$ such that

$$\beta^*_{j(t)} = 1 \text{ and } \beta^*_i = 0, \quad i \neq j(t).$$

The binary control function

$$u^*(t) = u^{j(t)}, \quad t \in [0, T] \text{ a.e.}$$

is therefore well-defined and yields an identical right-hand side function value

$$\mathcal{F}(t, u^*(t)) = \mathcal{F}(t, u^{j(t)}) = \beta^*_{j(t)} \mathcal{F}(t, u^{j(t)}) = \sum_{i=1}^{N} \beta^*_i(t) \mathcal{F}(t, u^i), \quad t \in [0, T] \text{ a.e.}$$

and identical objective function

$$J(y^*, u^*) = J(y^*, u^{j(\cdot)}) = \beta^*_{j(\cdot)} J(y^*_\beta, u^{j(\cdot)}) = \sum_{i=1}^{N} \beta^*_i(\cdot) J(y^*_\beta, u^i)$$

compared to the feasible and optimal solution (y^*_β, β^*) of **(BL)**. Therefore (y^*, u^*) is a feasible solution of **(BN)** with objective value J^{BL}. Next we show – by contradiction – that there exists no admissible solution to **(BN)** with a smaller cost value than J^{BL}. Hence, we assume that a feasible solution $(\hat{y}, \hat{u})$ of **(BN)** exists with objective value $\hat{J}^{BN} < J^{BL}$. Since the set $\{u^1, \ldots, u^N\}$ contains all feasible assignments of $\hat{u}$, there exists again an index function $\hat{j}(\cdot)$ such that $\hat{u}$ can be written as

$$\hat{u}(t) = u^{\hat{j}(t)}, \quad t \in [0, T] \text{ a.e.}$$

With the same arguments above, β is defined as

$$\beta_i(t) = \begin{Bmatrix} 1 & \text{if } i = \hat{j}(t), \\ 0 & \text{otherwise} \end{Bmatrix} \quad \text{for } i = 1, \ldots, 2^N \text{ and } t \in [0, T] \text{ a.e..}$$

Consequently, β is feasible for **(BL)** with objective function value $\hat{J}^{BN} < J^{BL}$ which contradicts the optimality assumption of problem **(BN)**. Thus (y^*, u^*) is an optimal solution of problem **(BN)**.

The converse is proven with the same argumentation starting from the optimal solution **(BN)**.

In the following we want to apply [17, Theorem 1]. For that purpose we additionally suppose that $f(\cdot, \boldsymbol{x})$ and $g(\cdot\,, \boldsymbol{s})$ belong to $W^{1,\infty}(0,T)$ for almost all $\boldsymbol{x} \in \Omega$ and $\boldsymbol{s} \in \Gamma$, respectively. Next we verify assumptions (H0) to (H3) of [17, Theorem 1]:

- (H0): In [20, Theorem 5.13] is proved that **(RL)** has a unique optimal solution.
- (H1): Due to our regularity assumptions for the inhomogeneities f and g, the objective J and $\mathcal{F}$ are (locally) Lipschitz-continuous.
- (H2): Utilizing the $W^{1,\infty}$-regularity for the inhomogeneities f and g again, we notice that $\mathcal{F}(\cdot\,, u^i)$ belongs also $W^{1,\infty}(0,T)$ for any $i = 1, \ldots, N$. Now, (H2) follows from [17, Proposition 1], i.e., there exists a constant $C > 0$ with

$$\left\| \frac{\mathrm{d}}{\mathrm{d}\tau} \left(e^{-\mathcal{A}(t-\tau)} \mathcal{F}(\tau, u^i) \right) \right\|_{C(\overline{\Omega})} \leq C \quad \text{for } 0 < \tau < t < T \text{ a.e. and } 1 \leq i \leq N.$$

- (H3): It follows also that the mapping $t \mapsto \mathcal{F}(t, u^i)$ is essentially bounded in $C(\overline{\Omega})$ for any $i \in \{1, \ldots, N\}$.

Summarizing we have the following result [17, Theorem 1 and Corollary 1]:

Proposition 2. *Let the regularity conditions for the data stated in Sect. 2 hold. Moreover, $f(\cdot\,, \boldsymbol{x})$ and $g(\cdot\,, \boldsymbol{s})$ belong to $W^{1,\infty}(0,T)$ for almost all $\boldsymbol{x} \in \Omega$ and $\boldsymbol{s} \in \Gamma$, respectively. Suppose that (y^*_α, α^*) is the solution to the relaxed problem* **(RL)** *with objective value J^{RL}. Choose an arbitrary $\varepsilon > 0$. Then there exists a feasible solution $(y^*_\varepsilon, u^*_\varepsilon)$ of problem* **(BN)** *satisfying*

$$J^{BN} \leq J^{RL} + \varepsilon.$$

Remark 2. Notice that a feasible solution $(y^*_\varepsilon, u^*_\varepsilon)$ of problem ***(BN)*** can be constructed from (y^*_α, α^*) by sum-up rounding; cf. Sect. 4.2 and [17, Algorithm 1]. $\Diamond$

4 Numerical Solution Method

In the following we describe in detail how to apply the theoretical results in a numerical realization. We utilize Algorithm 1 which is based on the approach described in [17]. To guarantee convergence in a finite number of steps, the sequences of non-negative accuracies $\{\varepsilon_k\}_{k\in\mathbb{N}}$ and the time discretizations should be chosen such that $\varepsilon_k \to 0$ and $\Delta t_k = \max_{i=1,\ldots,\nu^k}\{t^k_i - t^k_{i-1}\} \to 0$, according to Theorem 1 in [17].

4.1 Solution of the Relaxed Problem

Let us introduce two particular solutions for the state and dual equations:

Algorithm 1. Relaxation Method for the FEM Model

```
Choose a time discretization $\mathcal{G}^0 = \{0 = t_0^0 < t_1^0 \ldots < t_{\nu^0}^0 = T\}$, a sequence of non-negative accuracies $\{\varepsilon_k\}_{k\in\mathbb{N}}$ and some fixed tolerance $\varepsilon > 0$.
for $k = 0, 1, \ldots$ do
    Find an optimal control $\alpha^k$ of (RL) which stops with a tolerance of $\varepsilon_k$.
    Set $J_{\mathsf{rel}}^k = \hat{J}(\alpha^k)$.
    if $\alpha^k$ is binary admissible then
        break
    end if
    Using $\mathcal{G}^k$ and $\alpha^k$ to define a piecewise constant function $\beta^k$ as described in 4.3.
    Determine $J^k = \hat{J}(\beta^k)$.
    if $|J_{\mathsf{rel}}^k - J^k| \leq \varepsilon/2$ and $0 < \varepsilon_k \leq \frac{\varepsilon}{2}$ then
        break
    end if
    Choose $\mathcal{G}^{k+1} = \{0 = t_0^{k+1} < t_1^{k+1} \ldots < t_{\nu^{k+1}}^{k+1} = T\}$ such that $\mathcal{G}^k \subset \mathcal{G}^{k+1}$.
end for
Set $y_{\mathsf{bin}}^* = S\beta^k + \hat{y}$, $\beta^* = \beta^k$, $y^* = S\alpha^k + \hat{y}$ and $u^*(t) = \sum_{j=1}^N u^j \beta_j^k(t)$.
```

- $\hat{y} \in W(0,T)$ is the weak solution to

$$\begin{aligned} \hat{y}_t(t,\boldsymbol{x}) - \Delta\hat{y}(t,\boldsymbol{x}) + v(\boldsymbol{x}) \cdot \nabla\hat{y}(t,\boldsymbol{x}) &= f(t,\boldsymbol{x}) && \text{in } Q \text{ a.e.,} \\ \frac{\partial\hat{y}}{\partial n}(t,\boldsymbol{s}) + q(\boldsymbol{s})\hat{y}(t,\boldsymbol{s}) &= g(t,\boldsymbol{s}) && \text{on } \Sigma \text{ a.e.,} \\ \hat{y}(0,\boldsymbol{x}) &= y_\circ(\boldsymbol{x}) && \text{in } \Omega \text{ a.e.} \end{aligned}$$

- Further, $\hat{p} \in W(0,T)$ is the weak solution to

$$\begin{aligned} -\hat{p}_t(t,\boldsymbol{x}) - \Delta\hat{p}(t,\boldsymbol{x}) - \nabla \cdot \big(v(\boldsymbol{x})\hat{p}(t,\boldsymbol{x})\big) &= y^d(t,\boldsymbol{x}) - \hat{y}(t,\boldsymbol{x}) && \text{in } Q \text{ a.e.,} \\ \frac{\partial\hat{p}}{\partial n}(t,\boldsymbol{s}) + \big(q(\boldsymbol{s}) + v(\boldsymbol{s}) \cdot \boldsymbol{n}(\boldsymbol{s})\big)\hat{p}(t,\boldsymbol{s}) &= 0 && \text{on } \Sigma \text{ a.e.,} \\ \hat{p}(T,\boldsymbol{x}) &= 0 && \text{in } \Omega \text{ a.e.,} \end{aligned}$$

  where $\boldsymbol{n}$ denotes the outwart normal vector.

The first step in Algorithm 1 is to solve **(RL)**. To do so we use a first-order augmented Lagrange method. Thus, we consider for $c \geq 0$

$$\min \hat{J}(\alpha) + \frac{c}{2}\int_0^T \Big|\sum_{j=1}^N \alpha(t) - 1\Big|^2 \mathrm{d}t \quad \text{s.t.} \quad \alpha \in \mathcal{A}_{\mathsf{ad}} \text{ and } \sum_{j=1}^N \alpha_j(t) = 1, \tag{5}$$

where the penalty term

$$\frac{c}{2}\int_0^T \Big|\sum_{j=1}^N \alpha(t) - 1\Big|^2 \mathrm{d}t$$

is the augmentation term, $\hat{J}(\alpha) = J(y_\alpha, \alpha)$ is the reduced cost functional and y_α solves (4b)–(4c). The set $\mathcal{A}_{\mathsf{ad}}$ is defined by

$$\mathcal{A}_{\mathsf{ad}} = \big\{\alpha \in \mathcal{A} \,\big|\, \alpha_j(t) \in [0,1] \text{ on } [0,T] \text{ a.e. for } j = 1, \ldots, N\big\}$$

and $\mathcal{A} = L^2(0,T;\mathbb{R}^N)$. For $c > 0$ the augmented Lagrangian is given as

$$\mathcal{L}_c(\alpha,\lambda) = \hat{J}(\alpha) + \Big\langle \sum_{j=1}^N \alpha_j(\cdot) - 1, \lambda \Big\rangle_{L^2(0,T)} + \frac{c}{2} \Big\| \sum_{j=1}^N \alpha_j(\cdot) - 1 \Big\|_{L^2(0,T)}^2 .$$

For the inner optimization (i.e., the minimization of $\mathcal{L}_c(\cdot\,\lambda)$ with respect to the primal variable α) we choose a multiplier $\lambda^0 \in \mathcal{A}$ and set $k = 0$. Then, for $k = 0, 1, \ldots$, we solve for $c_k > 0$

$$\min \mathcal{L}_{c_k}(\alpha, \lambda^k) \quad \text{s.t.} \quad \alpha \in \mathcal{A}_{\mathsf{ad}} \qquad (\mathbf{P}_c^k)$$

and set

$$\lambda^{k+1} = \lambda^k + c_k \left(\sum_{j=1}^N \alpha_j(\cdot) - 1 \right).$$

For more details about Lagrangian methods see, e.g., [6, Chapter 3 and 4]. We repeat this process until we have

$$\Big\| \sum_{j=1}^N \alpha_j(\cdot) - 1 \Big\|_{L^2(0,T)}^2 \leq \varepsilon$$

for a given tolerance $\varepsilon > 0$. The optimality conditions are given as

$$\partial_\alpha \mathcal{L}(\bar{\alpha}, \bar{\lambda})(\alpha - \bar{\alpha}) = \langle \nabla_\alpha \mathcal{L}_c(\bar{\alpha}, \bar{\lambda}), \alpha - \bar{\alpha} \rangle_{\mathcal{A}} \geq 0 \quad \text{for all } \alpha \in \mathcal{A}_{\mathsf{ad}},$$

where $\partial_\alpha \mathcal{L}(\bar{\alpha}, \bar{\lambda}) : \mathcal{A} \to \mathbb{R}$ stands for the partial derivative with respect to α, $\mathcal{L}_c(\bar{\alpha}, \bar{\lambda}) \in \mathcal{A}$ is the gradient with respect to α. Moreover, $\bar{\alpha}$ is a local optimal solution to (P_c^k), and we have

$$\begin{aligned}\partial_\alpha \mathcal{L}(\bar{\alpha}, \bar{\lambda})\alpha^\delta &= \langle \nabla_\alpha \mathcal{L}_c(\bar{\alpha}, \bar{\lambda}), \alpha^\delta \rangle_{\mathcal{A}} \\ &= \hat{J}'(\bar{\alpha})\alpha^\delta + \sum_{j=1}^N \langle \bar{\alpha}_j^\delta, \bar{\lambda} \rangle_{L^2(0,T)} + c \sum_{j=1}^N \sum_{l=1}^N \langle \bar{\alpha}_j - 1, \alpha_l^\delta \rangle_{L^2(0,T)}\end{aligned}$$

for all directions $\alpha^\delta \in \mathcal{A}$. For a given point $\alpha \in \mathcal{A}_{\mathsf{ad}}$ and a direction $\alpha^\delta \in \mathcal{A}$ the directional derivative $\hat{J}'(\alpha)\alpha^\delta$ can be computed as follows:

1) Compute for a given $\alpha = (\alpha_i)_{1 \leq i \leq N} \in \mathcal{A}_{\mathsf{ad}}$ the state y_α solving

$$\begin{aligned}(y_\alpha)_t(t, \boldsymbol{x}) - \Delta y_\alpha(t, \boldsymbol{x}) + v(\boldsymbol{x}) \cdot \nabla y_\alpha(t, \boldsymbol{x}) &= \sum_{j=1}^N \left(\sum_{i=1}^m b_i(\boldsymbol{x}) u_i^j \right) \alpha_j && \text{in } Q \text{ a.e.,} \\ \frac{\partial y_\alpha}{\partial n}(t, \boldsymbol{s}) + q(\boldsymbol{s}) y_\alpha(t, \boldsymbol{s}) &= 0 && \text{on } \Sigma \text{ a.e.,} \\ y_\alpha(0, \boldsymbol{x}) &= 0 && \text{in } \Omega \text{ a.e.}\end{aligned}$$

and set $y = \hat{y} + y_\alpha$.

2) Solve the adjoint equation

$$\begin{aligned} -(p_\alpha)_t(t,\boldsymbol{x}) - \Delta p_\alpha(t,\boldsymbol{x}) - \nabla\cdot\big(v(\boldsymbol{x})p_\alpha(t,\boldsymbol{x})\big) &= -y_\alpha(t,\boldsymbol{x}) && \text{in } Q \text{ a.e.},\\ \frac{\partial p_\alpha}{\partial n}(t,\boldsymbol{s}) + \big(q(\boldsymbol{s}) + v(\boldsymbol{s})\cdot\boldsymbol{n}(\boldsymbol{s})\big)p_\alpha(t,\boldsymbol{s}) &= 0 && \text{on } \Sigma \text{ a.e.},\\ p_\alpha(T,\boldsymbol{x}) &= 0 && \text{in } \Omega \text{ a.e.} \end{aligned}$$

and set $p = \hat{p} + p_\alpha$.

3) Set for $\alpha^\delta \in \mathcal{A}$

$$\begin{aligned} \hat{J}'(\alpha)\alpha^\delta &= \frac{\gamma}{2}\sum_{j=1}^N\Big(\int_0^T\sum_{i=1}^m u_i^j\alpha_j^\delta(t)\,\mathrm{d}t\Big) - \int_0^T\int_\Omega\sum_{j=1}^N\Big(\sum_{i=1}^m b_i(\boldsymbol{x})u_i^j\Big)\alpha_j^\delta(t)p\,\mathrm{d}\boldsymbol{x}\mathrm{d}t\\ &= \sum_{j=1}^N\int_0^T\Big(\frac{\gamma}{2}\sum_{i=1}^m\big(u_i^j\alpha_j^\delta(t)\big) - \alpha_j^\delta(t)\int_\Omega p(t,\boldsymbol{x})\sum_{i=1}^m\big(b_i(\boldsymbol{x})u_i^j\big)\,\mathrm{d}\boldsymbol{x}\Big)\,\mathrm{d}t\\ &= \sum_{j=1}^N\int_0^T\Big(\frac{\gamma}{2}\sum_{i=1}^m u_i^j - \int_\Omega p(t,\boldsymbol{x})\sum_{i=1}^m\big(b_i(\boldsymbol{x})u_i^j\big)\,\mathrm{d}\boldsymbol{x}\Big)\alpha_j^\delta(t)\,\mathrm{d}t\\ &= \Big\langle\Big(\frac{\gamma}{2}\sum_{i=1}^m u_i^j - \int_\Omega p(\cdot,\boldsymbol{x})\sum_{i=1}^m\big(b_i(\boldsymbol{x})u_i^j\big)\,\mathrm{d}\boldsymbol{x}\Big)_{1\le j\le N},\alpha^\delta\Big\rangle_{L^2(0,T)}. \end{aligned}$$

Therefore we can introduce the Riesz representant of the linear, bounded functional $\hat{J}'(\alpha):\mathcal{A}\to\mathbb{R}$ by

$$\Big(\frac{\gamma}{2}\sum_{i=1}^m u_i^j - \int_\Omega p(\cdot,\boldsymbol{x})\sum_{i=1}^m\big(b_i(\boldsymbol{x})u_i^j\big)\,\mathrm{d}\boldsymbol{x}\Big)_{1\le j\le N} =: \nabla\hat{J}(\alpha)\in\mathcal{A}.$$

In particular, $\hat{J}'(\alpha) = \langle\nabla\hat{J}(\alpha),\cdot\rangle_{L^2(0,T)}$ holds true.

Remark 3. The first order optimality conditions for problem (P_c^k) are given by the variational inequality

$$\begin{aligned} &\Big\langle\Big(\frac{\gamma}{2}\sum_{i=1}^m u_i^j - \int_\Omega p(\cdot,\boldsymbol{x})\sum_{i=1}^m\big(b_i(\boldsymbol{x})u_i^j\big)\,\mathrm{d}\boldsymbol{x} + \bar{\lambda}\Big)_{1\le j\le N},\alpha-\bar{\alpha}\Big\rangle_{\mathcal{A}}\\ &\qquad + c\sum_{j=1}^N\sum_{l=1}^N\langle\bar{\alpha}_j - 1,\alpha_l - \bar{\alpha}_l\rangle_{L^2(0,T)} \ge 0 \end{aligned}$$

for all $\alpha = (\alpha_i)_{1\le i\le N}\in\mathcal{A}_{\mathsf{ad}}$, where $\bar{\alpha}$ is the optimal solution to (5). For more details see [20, Chapter 5]. ♢

4.2 Sum-Up Rounding

Assume that we have found an optimal solution α to **(RL)**. The next step in Algorithm 1 is to construct a binary admissible control function β for **(BL)**.

There we need to guarantee that we will not lose the special ordered set property (SOS-1). Therefore we construct β by using the following so-called sum up rounding strategy (compare to [17] and [31, Section 5.1]) as follows.

Let $\mathcal{G} = \{t_0, t_1, \dots, t_\nu\}$ be a time grid with $0 = t_0 < t_1 < \cdots < t_\nu = T$. Define $\beta = (\beta_1, \dots, \beta_N) : [0, T] \to \{0, 1\}^N$ by

$$\beta_i(t) = p_{i,j}, \quad t \in [t_j, t_{j+1}), \quad i = 1, \dots, N,\ j = 0, \dots, \nu - 1,$$

where for all $i = 1, \dots, N$, $j = 0, \dots, \nu - 1$

$$p_{i,j} = \begin{cases} 1 & \text{if } (\hat{p}_{i,j} \geq \hat{p}_{l,j}\ \forall l \in \{1, \dots, N\} \setminus \{i\}) \text{ and} \\ & (i < l : \forall l \in \{1, \dots, N\} \setminus \{i\} : \hat{p}_{i,j} = \hat{p}_{l,j}) \\ 0 & \text{else} \end{cases}$$

$$\hat{p_{j,i}} = \int_0^{t_{j+1}} \alpha_i(\tau) d\tau - \sum_{l=0}^{j-1} p_{i,l}(t_{l+1} - t_l).$$

Finally, to define a binary admissible control function for **(BN)** $u : [0, T] \to \{0, 1\}^m$ we set

$$u(t) = \sum_{j=1}^{N} u^j \beta_j(t).$$

Thanks to Theorem 1 we get then $J^{BL} = J^{BN}$, where J^{BL} depends on β and J^{BN} depends on u.

4.3 Redefine the Time Discretization

If the values between the cost functions of **(BN)** and **(RL)** are not small enough we need to redefine the time grid to get a better solution. To do so, there are several strategies given in [31, Section 5.3]. However, the simplest way would be just to define the grid in an equidistant way by double it and to use the old solution as a warmstart. In our numerical experiments this is the way we have done.

4.4 The POD Method

The most expensive part in Algorithm 1 is to find an optimal control α of **(RL)**. If one use here, e.g., finite elements to discretize the state and adjoint equations these leads to huge computational time. Therefore, to reduce the cost of the numerical solution method we apply a POD-method. To apply POD to the convexified problem assume that we have already computed a POD basis of rank ℓ and the corresponding POD space $V^\ell = \operatorname{span}\{\psi_1, \dots, \psi_\ell\} \subset V$ is given. Moreover, we assume that we have computed the inhomogeneous part $\hat{y}$ of the solution to the state equation. With the same notations as in Sect. 4.1 we introduce the weak formulation of the homogeneous part of the reduced-order state equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\langle y_\alpha^\ell(t),\psi\rangle_H + a(y_\alpha^\ell(t),\psi) = \Big\langle \sum_{i=1}^m u_i b_i, \psi \Big\rangle_H \qquad \forall \psi \in V^\ell \text{ in } (0,T], \tag{6}$$

$$\langle y_\alpha^\ell(0),\psi\rangle_H = 0 \qquad \forall \psi \in V^\ell. \tag{7}$$

and set $y^\ell = \hat{y} + y_\alpha^\ell$. Moreover we define the POD approximated reduced cost function by

$$\hat{J}^\ell(\alpha) := \frac{1}{2}\,\|y^\ell - y_d\|_{L^2(Q)}^2 + \frac{\gamma}{2}\sum_{j=1}^N \|u^j\|_{\mathbb{R}^m}^2 \int_0^T \alpha_j(t)\,\mathrm{d}t.$$

With the definition of the POD approximated reduced cost we can define the POD approximated reduced problem for the inner optimization of the Lagrange method, more precisely for (P_c^k). For the inner optimization we consider therefore the POD approximated reduced problem

$$\min \mathcal{L}_{c_k}^\ell(\alpha,\lambda^k) \quad \text{s.t.} \quad \alpha \in \mathcal{A}_{\mathsf{ad}} \tag{$\mathbf{P}_c^{\ell,k}$}$$

for a given c_k and λ^k, where we have for $c \geq 0$

$$\mathcal{L}_c^\ell(\alpha,\lambda^k) = \hat{J}^\ell(\alpha) + \Big\langle \sum_{j=1}^N \alpha_j - 1, \lambda^k \Big\rangle_{L^2(0,T)} + \frac{c}{2}\,\Big\| \sum_{j=1}^N \alpha_j - 1 \Big\|_{L^2(0,T)}^2.$$

Then the gradient of $\hat{J}^\ell$ is given by

$$\nabla \hat{J}^\ell(\alpha) = \Big(\frac{\gamma}{2}\sum_{i=1}^m u_i^j - \int_\Omega p^\ell(\cdot,\boldsymbol{x}) \sum_{i=1}^m \big(b_i(\boldsymbol{x}) u_i^j\big)\,\mathrm{d}\boldsymbol{x} \Big)_{1\leq j\leq N} \in \mathcal{A},$$

where $p^\ell = \hat{p} + p_\alpha^\ell$ holds true und p_α^ℓ solves the adjoint problem

$$-\frac{\mathrm{d}}{\mathrm{d}t}\,\langle p_\alpha^\ell(t),\psi\rangle_H + a(\psi, p_\alpha^\ell(t)) = -\langle y_\alpha^\ell(t),\psi\rangle_H \qquad \forall \psi \in V^\ell \text{ in } (0,T], \tag{8}$$

$$\langle p_\alpha^\ell(T),\psi\rangle_H = 0 \qquad \forall \psi \in V^\ell. \tag{9}$$

In [20, Theorem 5.38] is the following a-priori convergence result given.

Theorem 2. *Suppose assumptions from Sect. 2 hold. Let the linear, bounded operator $\mathcal{B} : \mathcal{A} \to L^2(0,T;(V^\ell)')$ be given as*

$$\langle \mathcal{B}\alpha)(t),\psi\rangle_{(V^\ell)',V^\ell} = \sum_{j=1}^N \alpha_j(t) \sum_{i=1}^m u_i^j \langle b_i,\psi\rangle_H \quad \textit{for } \alpha \in \mathcal{A} \textit{ in } [0,T] \textit{ a.e.}$$

We assume that $\mathcal{B}$ is injective. For arbitrarily given $\alpha \in \mathcal{A}$ we suppose that the solutions y_α and p_α to (6) and (9), respectively, belong to $H^1(0,T;V) \setminus \{0\}$.

1) *If we compute the POD space* V^ℓ *by solving*

$$\min \sum_{j=1}^{4} \int_0^T \Big\| y^j(t) - \sum_{i=1}^{\ell} \langle y^j(t), \psi_i \rangle_V \, \psi_i \Big\|_V^2 \,\mathrm{d}t \ \ s.t. \ \{\psi_i\}_{i=1}^{\ell} \subset V, \ \langle \psi_i, \psi_j \rangle_V = \delta_{ij}$$

using the snapshots $y^1 = y_\alpha$, $y^2 = (y_\alpha)_t$, $y^3 = p_\alpha$ *and* $y^4 = (p_\alpha)_t$, *then the optimal solution* $\bar{\alpha}$ *of* (P_c^k) *and the optimal solution* $\bar{\alpha}^\ell$ *to the reduced problem* ($\mathrm{P}_c^{\ell,k}$) *satisfy*

$$\lim_{\ell \to \infty} \|\bar{\alpha}^\ell - \bar{\alpha}\|_{\mathcal{A}} = 0.$$

2) *If an optimal POD basis of rank* ℓ *is computed by choosing the snapshots* $y^1 = y_{\bar{\alpha}}$, $y^2 = (y_{\bar{\alpha}})_t$, $y^3 = p_{\bar{\alpha}}$ *and* $y^4 = (p_{\bar{\alpha}})_t$, *then we have*

$$\lim_{\ell \to \infty} \|\bar{\alpha}^\ell - \bar{\alpha}\|_{\mathcal{A}} \leq C \sum_{i=\ell+1}^{\infty} \mu_i,$$

where $\{\mu\}_{i \in \mathbb{N}}$ *are the eigenvalues of the corresponding POD problem satisfying the error formula*

$$\sum_{j=1}^{4} \int_0^T \Big\| y^j(t) - \sum_{i=1}^{\ell} \langle y^j(t), \psi_i \rangle_V \, \psi_i \Big\|_V^2 \,\mathrm{d}t = \sum_{i=\ell+1}^{\infty} \mu_i;$$

cf. [14], *for instance.*

The following a-posteriori error estimate guarantees that the error stays small in the numerical solution method. A proof can be found in [20, Theorem 5.39].

Theorem 3. *Let all assumptions of Theorem 2 hold. For arbitrarily given* $\alpha \in \mathcal{A}$ *we choose the snapshots* $y^1 = y_\alpha$, $y^2 = (y_\alpha)_t$, $y^3 = p_\alpha$ *and* $y^4 = (p_\alpha)_t$. *Define the function* $\zeta^\ell \in \mathcal{A}$ *by*

$$\zeta_i^\ell(t) = \begin{cases} -\min(0, \xi_i^\ell(t)) & a.e. \ in \ \mathfrak{A}_{0,i}^\ell = \{t \in [0,T] \,|\, \bar{\alpha}_i^l(t) = 0\}, \\ \ \ \max(0, \xi_i^\ell(t)) & a.e. \ in \ \mathfrak{A}_{1,i}^\ell = \{t \in [0,T] \,|\, \bar{\alpha}_i^l(t) = 1\}, \\ -\xi_i^\ell(t) & a.e. \ in \ [0,T] \setminus (\mathfrak{A}_{0,i}^\ell \cup \mathfrak{A}_{1,i}^\ell), \end{cases}$$

where $\xi^\ell = \nabla_\alpha \mathcal{L}_c(\alpha^\ell, \lambda)$ *in* $\mathcal{A}$. *Then, for* $c > 0$ *we get the a-posteriori error estimate*

$$\|\bar{\alpha} - \bar{\alpha}^\ell\|_{\mathcal{A}}^2 \leq \frac{1}{c} \|\zeta^\ell\|_{\mathcal{A}}, \tag{10}$$

and in particular, $\lim_{\ell \to \infty} \|\zeta^\ell\|_{\mathcal{A}} = 0$.

With these results, we can solve **(RL)** with the POD method and control the error with the a-posteriori error estimate in our numerical solution approach.

Table 1. Parameter and function values for the numerical experiments

Symbol	Value	Description
T	1	Final time
Ω	$(0,1)^2$	Spatial domain
$v(x)$	$(1,1)^T$ for all $x \in \Omega$	Convection term
$q(x)$	0.1 on $\partial\Omega$	Isolation of the room
$f(t,x)$	0	No influence from the source term f
$g(t,x)$	see Fig. 1	Outside temperature modeled by a polynomial with degree 3

5 Numerical Experiments

In this section we investigate the mixed-integer optimal control problem numerically by the method introduced in Sect. 4. To recall, we consider the following mixed-integer optimal control problem:

$$\min_{y,u} J(y,u) = \frac{1}{2}\int_0^T \int_\Omega |y(t,\boldsymbol{x}) - y^d(t,\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t + \frac{\gamma}{2}\sum_{i=1}^m \int_0^T |u_i(t)|^2 \mathrm{d}t \tag{11a}$$

subject to a convection-diffusion equation

$$2y_t - \Delta y + v \cdot \nabla y = f + \sum_{i=1}^m u_i b_i \qquad \text{in } \Omega \tag{11b}$$

$$\frac{\partial y}{\partial n} + qy = g \qquad \text{on } \Sigma \tag{11c}$$

$$y(0) = y_0 \qquad \text{in } \Omega \tag{11d}$$

and binary admissibility of $u(\cdot)$

$$u(t) \in \{0,1\}^m \text{ in } [0,T] \text{ a.e. (almost everywhere).} \tag{11e}$$

We will use the parameters and function values which are given in Table 1. Moreover, for the desired temperature we first want a decrease in the temperature to 10° until $t = 0.25$, an increase to 18° until $t = 0.75$ and then again a decrease to 10°. With this we want to avoid simple solutions which are like $u(t) = 1$ or $u(t) = 0$ for all $t \in [0,T]$. Therefore we set for all $x \in \Omega$

$$y^d(t,x) = \begin{cases} 10 & \text{if } t < 0.25 \text{ or } t > 0.75, \\ 18 & \text{otherwise.} \end{cases}$$

In the following we will use all notations from the previous sections. The implementations are done in Python where we use the packages NumPy, SciPy and Matplotlib, see [21] as well as FEniCS, see [2,30]. All computations are done on a standard laptop (Acer Aspire 5, Intel(R) Core(TM) i5-8250U, 1.6 GHz (up to 3.4 GHz), 8 GB DDR4-RAM).

For solving the relaxed optimal control problems **(RL)** we use for the inner optimization in each iteration of the augmented Lagrange method (e.g. solving (P_c^k)) the L-BFGS-B method from SciPy, [7]. The maximum number of iterations for the L-BFGS-B method is set up to 100. For the first time grid we utilize a tolerance of $\varepsilon_0 = 10^{-4}$. After each modification of the time grid we divided ε_0 by 10 until $\varepsilon_k = 10^{-7}$ which give us a sequence of non-negative accuracies $\{\varepsilon_k\}$. To solve the integer problem we use as a tolerance $\varepsilon = 10^{-5}$. The first time discretization is given by an equidistant time grid $\mathcal{G}^0 \subset [0, T]$ with 50 time steps. The first optimal control problem **(RL)** is solved with a (pseudo) random initial control function. After that one we reuse the old solution to get faster convergence results. We stop the algorithm either if

$$\left|J^{BN} - J^{RL}\right| \leq \varepsilon$$

or if the size of the time grid is bigger than 800 grid points. We redefine the time grid in an equidistant way by double it.

In the tests we consider a tow dimensional control and impose a floor heating in the subdomains $\Omega_{b_1} = (0, 0.25) \times (0, 0.5) \subset \Omega$ and $\Omega_{b_2} = (0.25, 0.5) \times (0, 0.5) \subset \Omega$. Therefore we set $b_1(\boldsymbol{x}) = 1$ for all $\boldsymbol{x} \in \Omega_{b_1}$ and $b_2(\boldsymbol{x}) = 1$ for all $\boldsymbol{x} \in \Omega_{b_2}$. Set $u_1^1 = 0$, $u_1^2 = 1$, $u_2^1 = 0$ and $u_2^2 = 1$ the convexification and the relaxation leads to **(RL)** which has the form

$$\min_{y,\alpha} J(y, \alpha) = \frac{1}{2} \int_0^T \int_\Omega |y(t, \boldsymbol{x}) - y^d(t, \boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} \mathrm{d}t + \frac{\gamma}{2} \int_0^T (\alpha_2(t) + \alpha_3(t) + 2\alpha_4(t)) \mathrm{d}t$$

subject to

$$\begin{aligned}
y_t - \Delta y + v \cdot \nabla y &= f + b_1\alpha_2 + b_2\alpha_3 + (b_1 + b_2)\alpha_4, && \text{in } \Omega,\\
\frac{\partial y}{\partial n} + qy &= g, && \text{on } \Sigma,\\
y(0) &= y_0, && \text{in } \Omega,\\
\alpha(t) &\in [0, 1]^4, && \text{in } [0, T] \text{ a.e.},\\
\sum_{i=1}^4 \alpha_i(t) &= 1, && \text{in } [0, T] \text{ a.e..}
\end{aligned}$$

After finding an optimal control α we use the sum-up rounding strategy as described in Sect. 4.2 to get a binary admissible β and set

$$u(t) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \beta_1(t) + \begin{pmatrix} 1 \\ 0 \end{pmatrix} \beta_2(t) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \beta_3(t) + \begin{pmatrix} 1 \\ 1 \end{pmatrix} \beta_4(t)$$

for all $t \in [0, T]$.

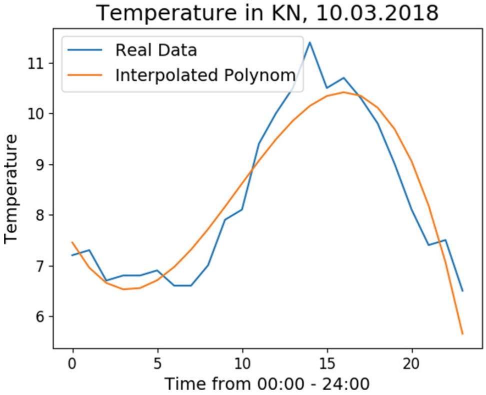

Fig. 1. Temperature outside given by real data and the interpolated polynom.

5.1 Full Finite Element Method Model

In our numerical experiments we compare the behavior of the algorithm with different regularization parameters γ. Notice, that for $\gamma = 0$ the problem leads to a bang-bang solution. Therefore, as bigger the regularization parameter gets, as more complicated is the integer problem. But on the other hand, big regularization parameters should make the relaxed problem easier to handle.

5.1.1 Case $\gamma = 0.01$

The first test is done with $\gamma = 0.01$. Here we use the initial condition that represents a constant temperature of 16° in the whole room, i.e. $y_0(\boldsymbol{x}) = 16$ for all $\boldsymbol{x} \in \Omega$. After four times redefining the grid, the algorithm has found a solution with a difference between $J^{BN} = J^{BL}$ and J^{RL} around $6 \cdot 10^{-5}$. The convergence behavior is given in the left subfigure of Fig. 2. Notice, that in the same figure we see also the difference between $J^{BN} = J^{BL}$ and J^{RL} (blue line) as well as the difference between J^{BN} and J^{RN} (orange line) which is close to the other one, caused of the small regularization parameter γ. Notice as well that we always have that $J^{RL} < J^{RN}$ which we could expect from the theoretical results. In Fig. 3 are the optimal control functions u_1 and u_2 and the corresponding binary functions β_1, β_2, β_3 and β_4. Notice that all control functions are close to bang-bang. Notice as well that to guarantee the SOS-1 property for the binary control functions β_i we need in β_2 an additional 1 at time step $t \approx 0.08$ which we don't see in the relaxed control α_2.

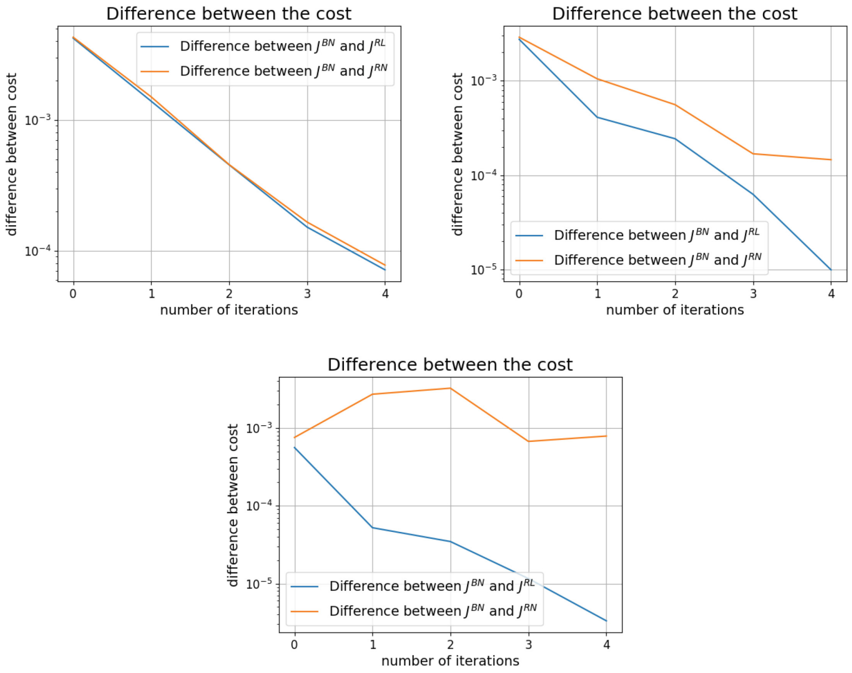

Fig. 2. Difference between J^{BN} and J^{RL} as well as between J^{BN} and J^{RN}. Case $\gamma = 0.01$ on the top left, $\gamma = 0.1$ on the top right and $\gamma = 1$ on the bottom middle.

5.1.2 Case $\gamma = 0.1$

Next we increase the regularization parameter and set therefore $\gamma = 0.1$. The algorithm modifies the time grid four times to reach a tolerance around 10^{-6} for the difference of $J^{BN} = J^{BL}$ and J^{RL}. The convergence behavior is given in the right subfigure of Fig. 2. Notice that the difference of $J^{BN} = J^{BL}$ and J^{RN} is bigger compared to the case $\gamma = 0.01$ and would not converge to a solution which is as close as that one we have found. Or in other words, just relax the integer problem leads to big duality gaps which do not close by just redefining the time grid.

5.1.3 Case $\gamma = 1$

Finally, we do the same test with $\gamma = 1$. To avoid solutions which are zero everywhere, we set $y_0(\boldsymbol{x}) = 14$ for all $\boldsymbol{x} \in \Omega$. The algorithm is done after four times redefinition of the time grid and a difference between $J^{BN} = J^{BL}$ and J^{RL} which is $\approx 10^{-6}$. The convergence behavior is given in the bottom subfigure of Fig. 2. Notice that there is a big difference between J^{BN} and J^{RN}. Here we see again the nice benefit of the convexification. Without that, it would be impossible to get such a small difference as for the solution, and we could say nothing how good our solution would be. Again we see that the duality gab

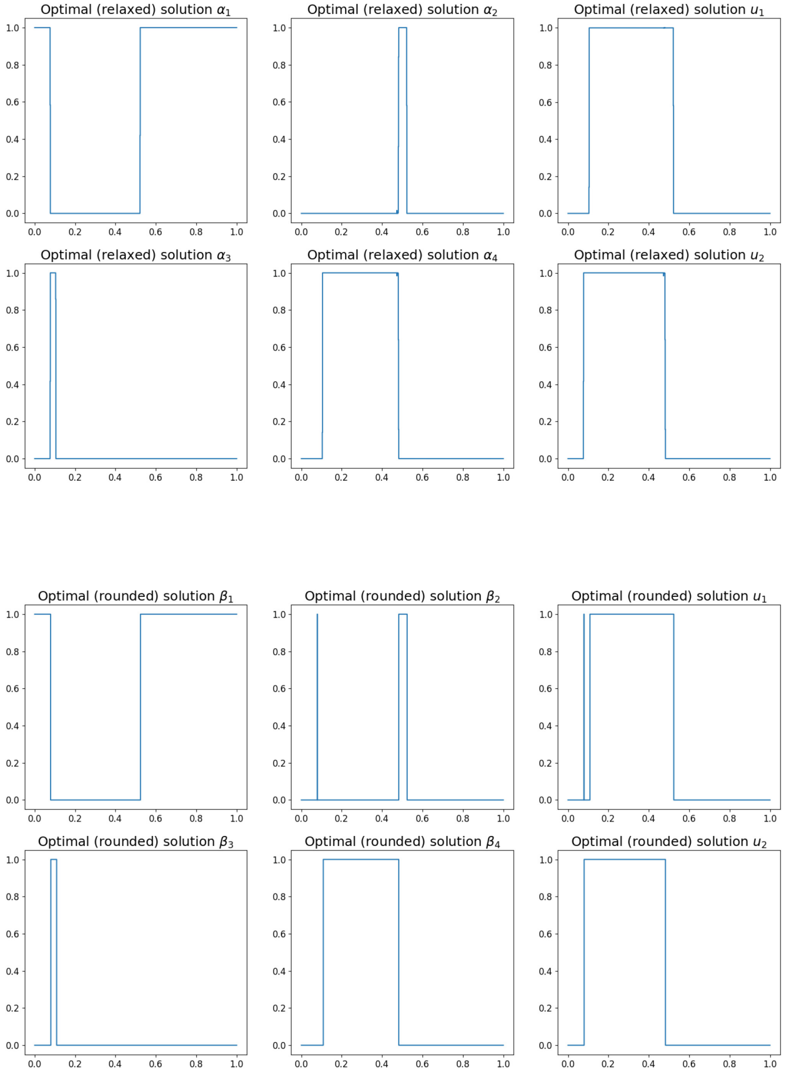

Fig. 3. Optimal control functions (relaxed and rounded) and corresponding optimal binary functions β_i for $\gamma = 0.01$.

between the integer solution and the relaxed solution without convexification is huge compared to the gap of the confexified problem.

If we have a look on Fig. 4 we can see the influence of the convection term in the control functions. The second control function is on the right side in the domain and from the convection therefore more expansive in the costs, although we do not weight our control functions. Therefore we get a zero control function for the second control. For the first one we have a little (but less than in the

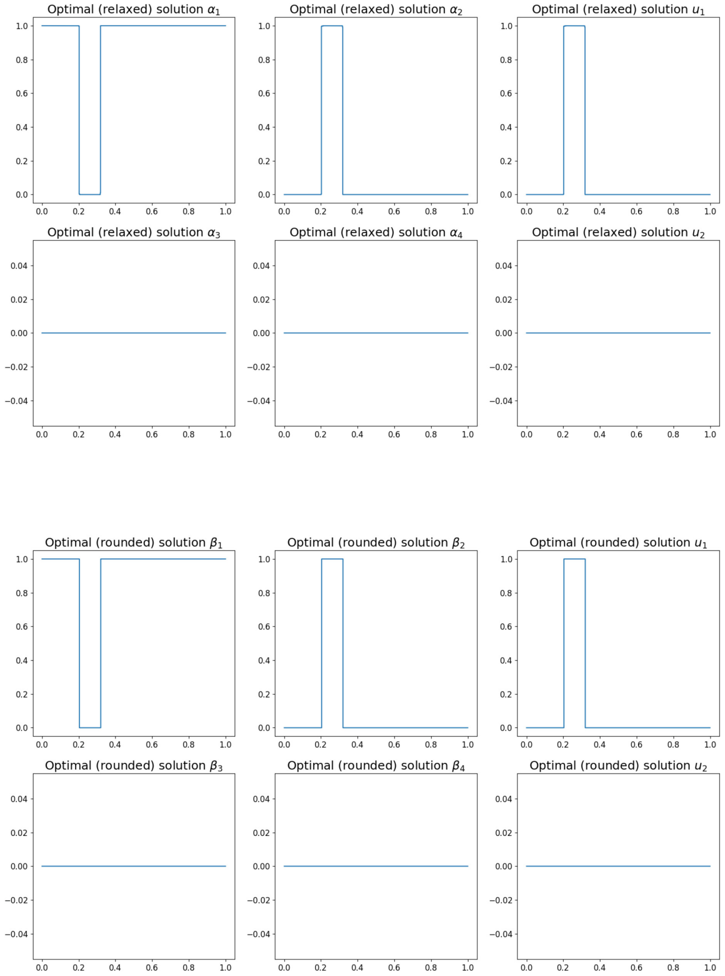

Fig. 4. Optimal control functions (relaxed and rounded) and corresponding optimal binary functions β_i for $\gamma = 1$.

case without convexification) chattering behavior. Summarized we see for a two-dimensional problem that the convexification leads to good solutions and the gap between integer and relaxed closes nicely. We also see, that the estimate between J^{BN} and J^{RL} is sharp, in contrast to J^{RN} which we could expect from the theoretical results. Moreover, we see a nice linear convergence behavior for the solutions by redefining the time grid in an equidistant way by doubling it.

5.2 The Reduced POD Model

In the following we do the same tests as in Sect. 5.1 but apply the POD method on the corresponding relaxed problems. We investigated the quality of the POD approximated solutions with the a-posteriori error estimate from Theorem 3 and compare the computational time. Since we are solving an integer problem we expect similar solutions as before and compare therefore the values of the cost functions corresponding to **(RL)** as well as **(RN)**.

To generate a POD basis of rank l we use the snapshots $y^1 = S\alpha_0$ and $y^2 = \mathcal{A}\alpha_0$, where α_0 is a (pseudo) random initial control function which we use for both problems as initial one e.g. we start both optimal control problems with this one. We compute the POD basis as describes in [20, Section 4] using trapezoidal weights and solve the eigenvalue problem corresponding to the POD problem with the SVD method. As weighting matrix we use the mass matrix and we use $l = 10$ snapshots. The offline phase, which mean calculating the POD ansatz functions need in all three cases around 0.025 s.

5.2.1 Case $\gamma = 0.01$

For the FEM method, the algorithm needs to redefine the grid four times and found a solution after 1994.2 s. The algorithm spends the most time (around 1400 s) in the last time grid. This could come from a bad initial condition since in all other cases, the reuse of the old solution works pretty well. The difference between $J^{BN} = J^{BL}$ and J^{RL} is $\approx 7 * 10^{-5}$.

The POD method needs 154.2 s and is therefore more than 12 times faster, although we have use the same initial condition. In this test for u_2, at $t \approx 0.5$ the POD method has found a solution where $u_2^{POD}(t) = 1$ and $u_2^{FEM}(t) = 0$. The rest is equal. The convergence behavior of the full problem and the POD problem is given in Fig. 5 as well as the difference between J_{FEM}^{RL} and J_{FEM}^{RL}. In Table 2 we have given all values of the cost in the different time grids for the full problem as well as for the POD problem. Notice the interesting behavior of the a-posteriori error functions ζ.

Finally we have a look of the difference of the controls in the final grid. Here we have

$$\left\|u_{FEM}^{Rel} - u_{POD}^{Rel}\right\|_A = 0.02171, \quad \left\|u_{FEM}^{Int} - u_{POD}^{Int}\right\|_A = 0.02.$$

Therefore we can conclude that the POD method founds a similar solution and is much faster than the full method. Moreover, we can see that the POD method has found a solution in each time grid which is equal or slightly bigger than that one from the full model.

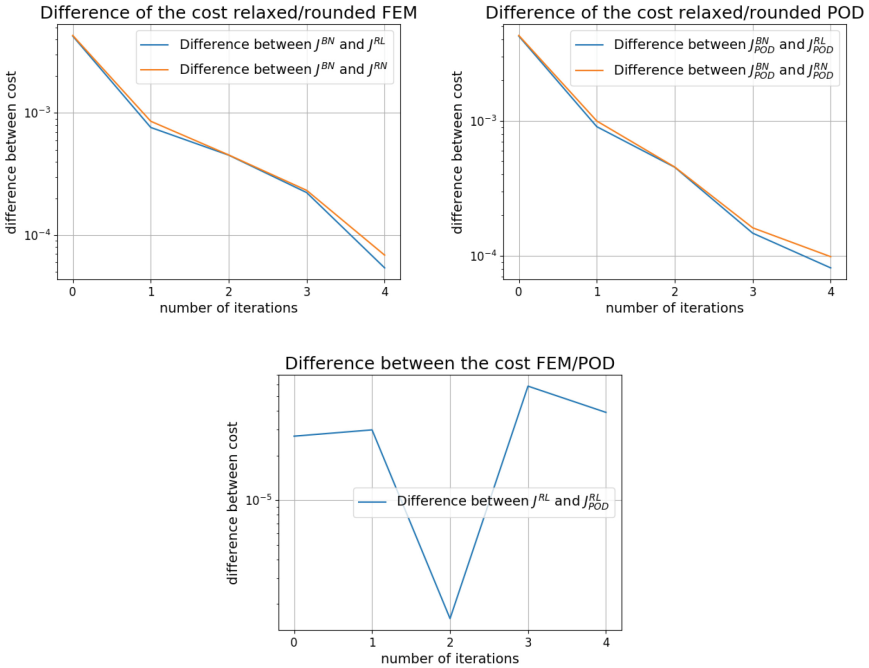

Fig. 5. Convergence behavior of the cost for $\gamma = 0.01$. On the top left, difference between the relaxed and the binary cost for the FEM. On the top right the difference for the POD method and on the bottom middle the difference between the convexified relaxed cost for the FEM and the POD method.

Table 2. Summarized values in the different time grids for $\gamma = 0.01$.

Time instances	50	100	200	400	800
$\frac{1}{c}\|\zeta\|$	$6 \cdot 10^{-8}$	0.00011	10^{-8}	$6 \cdot 10^{-5}$	$8 \cdot 10^{-6}$
$J^{BN} = J^{BL}$	11.5186	11.4813	11.4924	11.5009	11.5027
J^{RN}	11.5143	11.4805	11.4919	11.5006	11.5026
J^{RL}	11.5144	11.4806	11.4919	11.5006	11.5026
$J^{BN}_{POD} = J^{BL}_{POD}$	11.5186	11.4815	11.4924	11.5009	11.5028
J^{RN}_{POD}	11.5143	11.4805	11.4919	11.5007	11.5027
J^{RL}_{POD}	11.5144	11.4806	11.4919	11.5007	11.5027

5.2.2 Case $\gamma = 0.1$

Like before both algorithms need to redefine the time grid four times. For the full model the algorithm needs 1789.4 s. Again, the augmented Lagrange method in the final time grid needs the most time of the whole process. The difference between $J^{BN} = J^{BL}$ and J^{RL} is $\approx 10^{-6}$ and therefore reached our accuracy.

For the reduced POD model the algorithm needs 144.9 s and is therefore again more than 12 times faster. But the difference between $J_{POD}^{BN} = J_{POD}^{BL}$ and J_{POD}^{RL} is $\approx 10^{-5}$ and therefore a bit worse than for the full model. The convergence behavior of the costs for the full model and the POD model is given in Fig. 6. Table 3 shows again the different values of the cost and the behavior of the error functions ζ.

Again we have a look at the difference of the control in the final grid. This time we have

$$\left\| u_{FEM}^{Rel} - u_{POD}^{Rel} \right\|_A = 0.039, \quad \left\| u_{FEM}^{Int} - u_{POD}^{Int} \right\|_A = 0.141.$$

The discrete control u_1 of the full model and the POD model differs for

$$t \in \left\{ 0.13625, 0.48875, 0.49, 0.49125 \right\}$$

and the discrete control u_2 differs for

$$t \in \left\{ 0.13375, 0.135, 0.41, 0.4125, 0.49125 \right\}$$

which causes the difference of the binary control functions. Notice, again the value of the cost function J_{POD}^{BN} is equal at every time grid except in the case where the grid is of the size 200. Here the value of J_{POD}^{BN} is slightly bigger.

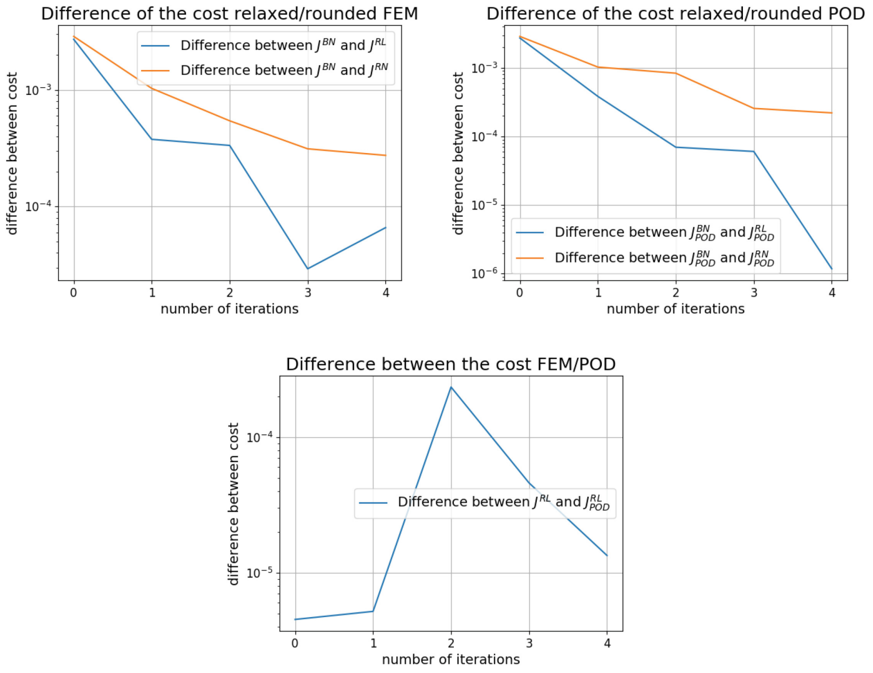

Fig. 6. Convergence behavior of the cost for $\gamma = 0.1$. On the top left, difference between the relaxed and the binary cost for the FEM. On the top right the difference for the POD method and on the bottom middle the difference between the convexified relaxed cost for the FEM and the POD method.

Table 3. Summarized values in the different time grids for $\gamma = 0.1$.

Time instances	50	100	200	400	800
$\frac{1}{c}\|\zeta\|$	$1 \cdot 10^{-9}$	$8 \cdot 10^{-5}$	$6 \cdot 10^{-5}$	$5 \cdot 10^{-7}$	$4 \cdot 10^{-6}$
$J^{BN} = J^{BL}$	11.5494	11.5126	11.5246	11.5338	11.5355
J^{RN}	11.5465	11.5115	11.5241	11.5335	11.5353
J^{RL}	11.5467	11.5122	11.5243	11.5337	11.5355
$J^{BN}_{POD} = J^{BL}_{POD}$	11.5494	11.5126	11.5247	11.5338	11.5355
J^{RN}_{POD}	11.5465	11.5116	11.5238	11.5335	11.5353
J^{RL}_{POD}	11.5466	11.5122	11.5246	11.5337	11.5352

5.2.3 Case $\gamma = 1$

The full model redefines the time grid three times and needed 277.1 s. The difference between $J^{BN} = J^{BL}$ and J^{RL} is $\approx 10^{-6}$, so we reach our tolerance what is the reason why the algorithm needs much less time and one time grid less than in the cases of $\gamma = 0.01$ and $\gamma = 0.1$. This could come from a good initial control u_0. Notice in the test of Subsect. 5.1 the algorithm needed one time grid more to reach this tolerance. Using the reduced POD model the algorithm needs 42.8 s and redefined the time grids three times. The plots of the convergence behavior are given in Fig. 7 and in Table 4 we have summarized our findings. Notice this time the value of J^{BL}_{POD} is a bit smaller or equal than the cost J^{BL} for the full model. The difference between $J^{BN}_{POD} = J^{BL}_{POD}$ and J^{RL}_{POD} is $\approx 10^{-6}$.

Having a look at the difference of the control functions in the final time grid we get

$$\left\|u^{Rel}_{FEM} - u^{Rel}_{POD}\right\|_A = 0.0073,$$
$$\left\|u^{Int}_{FEM} - u^{Int}_{POD}\right\|_A = 0.02.$$

The only difference in the binary control functions is for u_1 at $t = 0.205$.

Table 4. Summarized values in the different time grids for $\gamma = 1$.

Time instances	50	100	200	400
$\frac{1}{c}\|\zeta\|$	$2 \cdot 10^{-9}$	$1 \cdot 10^{-6}$	$6 \cdot 10^{-6}$	$1 \cdot 10^{-6}$
$J^{BN} = J^{BL}$	12.2703	12.2492	12.2561	12.2936
J^{RN}	12.2697	12.2466	12.2529	12.2903
J^{RL}	12.2697	12.2491	12.2566	12.2936
$J^{BN}_{POD} = J^{BL}_{POD}$	12.2703	12.2492	12.2561	12.2936
J^{RN}_{POD}	12.2697	12.2466	12.2579	12.2903
J^{RL}_{POD}	12.2697	12.2491	12.2561	12.2936

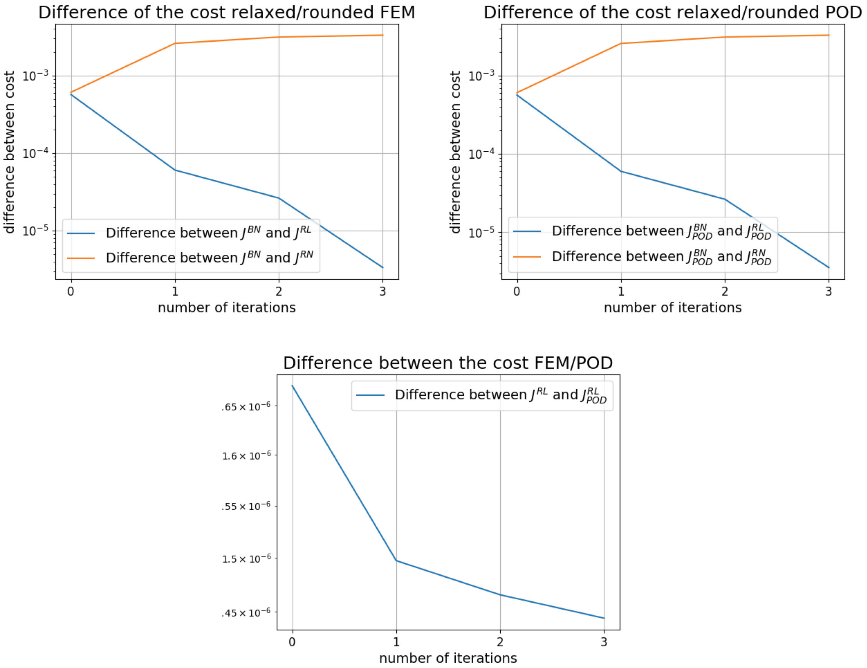

Fig. 7. Convergence behavior of the cost for $\gamma = 1$. On the top left, difference between the relaxed and the binary cost for the FEM. On the top right the difference for the POD method and on the bottom middle the difference between the convexified relaxed cost for the FEM and the POD method.

Table 5. Different computational times for the FEM method and the POD method for different regularization parameter.

Computational time			
			FEM/POD
$\gamma = 0.01$	FEM	1994.2 s	12, 9
	POD	154.2 s	
$\gamma = 0.1$	FEM	1789.4 s	12.3
	POD	144.9 s	
$\gamma = 1$	FEM	277.1 s	6.5
	POD	42.8 s	

Summarized we can definitely say that working with a reduced POD model instead of the full model gives a huge improvement of the computational time (see Table 5 for summarizing the speed up). Moreover, we get with this approach similar solutions which can lead incidental to smaller values of the cost function, therefore even better solutions. We have also seen, that a very small number l of ansatz functions for the POD basis are enough to reach this good solutions.

6 Conclusions and Outlook

In this chapter the authors have dealt with the application of relaxation methods combined with proper orthogonal decomposition (POD) methods for model order reduction to solve mixed-integer optimal control problems governed by linear convection-diffusion equations. After adopting the algorithm of [17] and verifying that this problem satisfies the assumptions of Theorem 1 in [17] to guarantee convergence a detailed description of the numerical solution method was given. Since the finite element method to discretize the state and adjoint equations from the optimization procedure leads to huge systems which have to be solved frequently, the POD method was introduced. This reduced the time-consuming optimization process and leads to a significant acceleration of the CPU times while the error remains small. The functionality of the algorithm and this behavior was verified by numerical experiments.

References

1. Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems. SIAM, Philadelphia (2005)
2. Alnæs, M., Blechta, J., Hake, J., Johansson, A., Kehlet, B., Logg, A., Richardson, C., Ring, J., Rognes, M., Wells, G.: The FEniCS project version 1.5. Arch. Numer. Softw. **3**, 9–23 (2015)
3. Bachmann, F.: A branch-and-bound approach to mixed-integer optimal control using POD. Master's thesis, Department of Mathematics and Statistics, University of Konstanz (2017). http://nbn-resolving.de/urn:nbn:de:bsz:352-0-408645
4. Bachmann, F., Beermann, B., Lu, J., Volkwein, S.: POD-based mixed-integer optimal control of the heat equation. J. Sci. Comput. **81**, 48–75 (2019)
5. Belotti, P., Kirches, Ch., Leyffer, S., Linderoth, J., Luedtke, J., Mahajan, A.: Mixed-integer nonlinear optimization. Acta Numerica **22**, 1–131 (2013)
6. Bertsekas, D.: Nonlinear Programming. Athena Scientific Optimization and Computation Series. Athena Scientific, Belmont (1999)
7. Byrd, R., Lu, P., Nocedal, J., Zhu, C.: A limited memory algorithm for bound constrained optimization. SIAM J. Sci. Comput. **16**, 1190–1208 (1995)
8. Dautray, R., Lions, J.-L.: Mathematical Analysis and Numerical Methods for Science and Technology. Volume 5: Evolution Problems I. Springer, Berlin (2000)
9. Evans, L.C.: Partial Differential Equations. American Mathematical Society, Providence (2002)
10. Fügenschuh, A., Geißler, B., Martin, A., Morsi, A.: The transport PDE and mixed-integer linear programming. In: Models and Algorithms for Optimization in Logistics, 21–26 June 2009 (2009)
11. Gerdts, M.: Optimal Control of ODEs and DAEs. De Gruyter, Berlin (2011)
12. Göttlich, S., Hante, F., Potschka, A., Schewe, L.: Penalty alternating direction methods for mixed-integer optimal control with combinatorial constraints. Preprint (2019). arXiv:1905.13554v2
13. Grepl, M.A., Maday, Y., Nguyen, N.C., Patera, A.T.: Efficient reduced-basis treatment of nonaffine and nonlinear partial differential equations. ESAIM: Math. Model. Numer. Anal. **41**, 575–605 (2007)

14. Gubisch, M., Volkwein, S.: Proper orthogonal decomposition for linear-quadratic optimal control (Chap. 1). In: Model Reduction and Approximation. Computational Science and Engineering, pp. 3–63 (2017)
15. Gugat, M., Leugering, G., Martin, A., Schmidt, M., Sirvent, M., Wintergerst, D.: MIP-based instantaneous control of mixed-integer PDE-constrained gas transport problems (2017)
16. Hante, F.: Relaxation methods for hyperbolic PDE mixed-integer optimal control problems. Optim. Control Appl. Methods **38**, 1103–1110 (2017)
17. Hante, F., Sager, S.: Relaxation methods for mixed-integer optimal control of partial differential equations. Comput. Optim. Appl. **55**, 197–225 (2013)
18. Hinze, M., Pinnau, R., Ulbrich, M., Ulbrich, S.: Optimization with PDE Constraints. Mathematical Modelling: Theory and Applications, vol. 23. Springer, Berlin (2009)
19. Holmes, P., Lumley, J., Berkooz, G., Rowley, C.: Turbulence, Coherent Structures, Dynamical Systems and Symmetry, Cambridge Monographs on Mechanics, 2nd edn. Cambridge University Press, Cambridge (2012)
20. Jäkle, C.: POD-based mixed-integer optimal control of convection-diffusion equations. Master thesis, Department of Mathematics and Statistics, University of Konstanz (2019). http://nbn-resolving.de/urn:nbn:de:bsz:352-2-qgy7rxfhsbp89
21. Jones, E., Oliphant, T., Peterson, P., et al.: SciPy: Open source scientific tools for Python (2001)
22. Kirches, C.: Fast Numerical Methods for Mixed-Integer Nonlinear Model-Predictive Control. Advances in Numerical Mathematics. Vieweg+Teubner Verlag (2011)
23. Kunisch, K., Volkwein, S.: Optimal snapshot location for computing POD basis functions. ESAIM: Math. Model. Numer. Anal. **44**, 503–529 (2010)
24. Leyffer, S., Cay, P., Kouri, D., van Bloemen Waanders, B.: Mixed-integer PDE-constrained optimization. Technical report, Oberwolfach Report (2015)
25. Lions, J.L.: Optimal Control of Systems Governed by Partial Differential Equations. Springer-Verlag, Grundlehren der mathematischen Wissenschaften (1971)
26. Manns, P., Bestehorn, F., Hansknecht, C., Kirches, C., Lenders, F.: Approximation properties of sum-up rounding and consequences for mixed-integer PDE-constrained optimization. In: Liberti, L., Sager, S., Wiegele, A. (eds.) Mixed-Integer Nonlinear Optimization: A Hatchery for Modern Mathematics, Oberwolfach Reports **26**, 40-42 (2019)
27. Manns, P., Kirches, C., Lenders, F.: A linear bound on the integrality gap for Sum-Up Rounding in the presence of vanishing constraints. Submitted (2018). http://www.optimization-online.org/DB_FILE/2018/04/6580.pdf
28. Nittka, R.: Regularity of solutions of linear second order elliptic and parabolic boundary value problems on Lipschitz domains. J. Differ. Equ. **251**, 860–880 (2011)
29. Nittka, R.: Elliptic and parabolic problems with robin boundary conditions on Lipschitz domains. Ph.D. thesis, Faculty of Mathematics and Economics, Ulm University (2010)
30. Ølgaard, K., Logg, A., Wells, G.: Automated code generation for discontinuous Galerkin methods. SIAM J. Sci. Comput. **31**, 849–864 (2008)
31. Sager, S.: Numerical Methods for Mixed-Integer Optimal Control Problems. Der andere Verlag, Tönning (2005). https://mathopt.de/PUBLICATIONS/Sager2005.pdf
32. Sager, S., Bock, H., Diehl, M.: The integer approximation error in mixed-integer optimal control. Math. Program. **133**, 1–23 (2012)

33. Sager, S., Reinelt, G., Bock, H.G.: Direct methods with maximal lower bound for mixed-integer optimal control problems. Math. Program. **18**, 109–149 (2009)
34. Schilders, W., van der Vorst, H., Rommes, J.: Model Order Reduction: Theory. The European Consortium for Mathematics in Industry. Research Aspects and Applications. Springer, Heidelberg (2008)
35. Tröltzsch, F., Volkwein, S.: POD a-posteriori error estimates for linear-quadratic optimal control. Comput. Optim. Appl. **44**, 83–115 (2009)
36. Tröltzsch, F.: Optimal Control of Partial Differential Equations: Theory, Methods, and Applications. American Mathematical Society, Providence (2010)