Different discretization techniques for solving optimal control problems with control complementarity constraints

Yu Deng[1]

Abstract

There are first-optimize-then-discretize (indirect) and first-discretize-then-optimize (direct) methods to deal with infinite dimensional optimal problems numerically by use of finite element methods. Generally, both discretization techniques lead to different structures. Regarding the indirect method, one derives optimality conditions of the considered infinite dimensional problems in appropriate function spaces firstly and then discretizes them into suitable finite element spaces. One has freedom to chose ansatz spaces for functions. On the contrary, w.r.t. the direct method, one doesn't need to investigate functional properties of the given problem, but transform the overall system into a standard finite dimensional optimal problem. Depending on the situation, each method has its own advantages and disadvantages.

MSC(2020) 49M25, 49M41, 65M22, 90C33.

Keywords Complementarity constraints, Optimal control, Parabolic PDE, Smoothed Fischer Burmeister function, Discretization methods.

1 Introduction

This paper deals with different discretization techniques to solve optimal control problems w.r.t. linear parabolic PDE and control complementarity constraints. We adopt the theoretical results from [2] and solve the considered problem by use of penalty approach with the smoothed Fischer Burmeister function. Finite dimensional nonlinear large scale equation systems of surrogate problems are derived in terms of both discretization techniques. By means of a numerical example both discretization approaches will be quantitatively and demonstratively compared. Explicitly, the following modell of optimal control problem is considered:

$$\begin{aligned} J(y,u,v) = \tfrac{1}{2}\|\mathrm{D}[y(\cdot,T)] - y_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda_1}{2}\|u\|_{\mathcal{U}} + \tfrac{\lambda_2}{2}\|v\|^2_{\mathcal{U}} &\to \min \\ \mathrm{A}(y) - \mathrm{B}(u) - \mathrm{C}(v) &= 0 \\ (u,v) &\in \mathbb{C} \end{aligned} \tag{OC}$$

[1]Technische Universität Bergakademie Freiberg, Faculty of Mathematics and Computer Science, 09596 Freiberg, Germany, yu.deng@math.tu-freiberg.de https://tu-freiberg.de/fakult1/nmo/deng

The following linear parabolic PDE represents the state equation:

$$\begin{aligned} \tfrac{d}{dt} y - \nabla \cdot (\mathrm{C}\nabla y) + \mathrm{a}\, y &= 0 && \text{a.e. on } Q := \Omega \times \mathcal{I} \\ \vec{n} \cdot (\mathrm{C}\nabla y) + \mathrm{q}\, y &= \mathrm{b}u + \mathrm{c}v && \text{a.e. on } \Sigma := \Gamma \times \mathcal{I} \\ y(\cdot, 0) &= 0 && \text{a.e. on } \Omega. \end{aligned} \tag{1}$$

Therein, $Q := \Omega \times \mathcal{I}$ be the underlying space-time cylinder with lateral boundary $\Sigma := \Gamma \times \mathcal{I}$. The domain $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, is nonempty, bounded and possesses a Lipschitz boundary. The time interval $\mathcal{I} = (0, T)$ has the end time $T > 0$. We use ∇ to denote the gradients w.r.t. the spacial variables $x \in \Omega$. The functions $\mathrm{b}, \mathrm{c} \in L^\infty(\Gamma)$ are assumed to stipulate the effective range of the controls and they can be modelled as characteristic functions of certain subsets of the boundary domain Γ. The underlying control space $\mathcal{U}$ is a Hilbert space and the controls are considered to be dependent only on time. It is assumed that the parabolic PDE (1) possesses a unique weak solution y in a suitable state space $\mathcal{Y}$ for any pair of controls $(u, v) \in \mathcal{U}^2$. Furthermore, D is chosen as a linear operator which maps the state at the end time into the observation space $L^2(\Omega)$ and the target $y_\mathrm{d} \in L^2(\Omega)$ is fixed. As a special constraint, Additionally, the objective functional is minimized w.r.t. all controls coming from the complementarity set defined below:

$$\mathbb{C} := \{w, z \in \mathcal{U}^2 : \quad w(x) \geq 0 \wedge z(x) \geq 0 \wedge w(x)z(x) = 0 \quad \text{a.e. on } \mathcal{I}\}. \tag{2}$$

The authors have already considered the control complementarity constrained optimal control problem with a linear PDE as state equation in [1] and realize that $\mathcal{U}$ chosen from at least the first order Sobolev space $H^1(\mathcal{I})$ guarantees that the feasible set $\mathbb{C}$ is weakly sequentially closed, see [1, Lemma 2.3]. The objective functional of (OC) is continuously Fréchet differentiable, convex and $\mathcal{U}$-coercive for positive Tikhonov parameters $\lambda_1, \lambda_2 > 0$. Thus, the problem (OC) possesses an optimal solution, see [1, Lemma 2.4].

In [2], several penalty approaches have been used to deal with the pointwise complementarity of controls, e.g. by use of a single NCP - Fischer Burmeister function, which was introduced by Fischer [3] and is defined by $\phi : \mathbb{R}^2 \to \mathbb{R}$ for $a, b \in \mathbb{R}$:

$$\phi(a, b) := \sqrt{a^2 + b^2} - a - b. \tag{3}$$

As we know, although the squared Fischer Burmeister function is differentiable everywhere, but it isn't second order differentiable at the origin point. Additionally, the Hessian matrix of squared Fischer Burmeister function is singular in the whole complementarity set. One gets difficulty in performing Newton-like method. The sequence of stationary points of surrogate penalized problems fails to converge to a reasonable stationary point of the original considered problem, even by use

of the globalized (damped) Newton method, see [2]. In this paper, we inherit the theoretical results of penalty methods from [2] and the smoothed Fischer Burmeister function is instead applied to deal with the control complementarity constraints. For any $a, b \in \mathbb{R}$, the smoothed Fischer Burmeister function $\phi_\alpha : \mathbb{R}^2 \to \mathbb{R}$ is defined by:

$$\phi_\alpha(a,b) := \sqrt{a^2 + b^2 + \alpha} - a - b. \tag{4}$$

where $\alpha \geq 0$ is the so-called smoothing parameter. It is obvious that

$$\phi_0(a,b) = \phi(a,b) = 0 \Leftrightarrow a \geq 0 \wedge b \geq 0 \wedge ab = 0$$

holds true.

For the setting $\mathcal{U} := H^1(\mathcal{I})$, we consider a sequence of following surrogate problems

$$\begin{aligned} J_{\gamma_k,\alpha_k}(y,u,v) = \tfrac{1}{2}\|\mathrm{D}[y(\cdot,T)] - y_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda_1}{2}\|u\|_{H^1(\mathcal{I})} + \tfrac{\lambda_2}{2}\|v\|^2_{H^1(\mathcal{I})} \\ + \tfrac{\gamma_k}{2}\|\Phi_{\alpha_k}(\mathrm{E}[u],\mathrm{E}[v])\|^2_{L^2(\mathcal{I})} \to \max_{y,u,v} \qquad (\mathrm{P}_{\gamma_k,\alpha_k}) \\ \mathrm{A}(y) - \mathrm{B}(u) - \mathrm{C}(v) = 0 \end{aligned}$$

with positive sequences $\{\gamma_k\}_{k\in\mathbb{N}}$ tending to ∞ and $\{\alpha_k\}_{k\in\mathbb{N}}$ tending to 0 as $k \to \infty$. Therein, $\mathrm{E} : H^1(\mathcal{I}) \hookrightarrow L^2(\mathcal{I})$ is the compact embedding on the given time interval $\mathcal{I}$ and $\Phi_{\alpha_k} : L^2(\mathcal{I}) \times L^2(\mathcal{I}) \to L^2(\mathcal{I})$ is the associated Nemytskii operator with ϕ_{α_k} which is defined by

$$\forall (w,z) \in L^2(\mathcal{I})^2 \ \forall t \in \mathcal{I}: \quad \Phi_{\alpha_k}(w,z)(t) := \phi_{\alpha_k}(w(t), z(t)). \tag{5}$$

Since the presented theoretical results in [2] can be applied for the penalized problem $(\mathrm{P}_{\gamma_k,\alpha_k})$ as well, we will only present some new results w.r.t. the smoothed Fischer Burmeister function and concentrate on the discretization process by use of different discretization techniques.

2 Penalized problem with squared smoothed Fischer Burmeister function

For all $(w,z) \in L^2(\mathcal{I})$, the operator $\Phi_\alpha : L^2(\mathcal{I})^2 \to L^2(\mathcal{I})$ is well defined, see [1]. We denote the penalty term by $F_\alpha : H^1(\mathcal{I})^2 \to \mathbb{R}^+_0$ with

$$\forall (u,v) \in H^1(\mathcal{I})^2: \quad F_\alpha(u,v) := \tfrac{1}{2}\|\Phi_\alpha(\mathrm{E}(u),\mathrm{E}(v))\|^2_{L^2(\mathcal{I})}. \tag{6}$$

Let $\alpha > 0$ be fixed, the operator Φ_α as well as F_α are Fréchet differentiable everywhere, since ϕ_α is smooth, see [4]. Here, we present the Fréchet derivative of

$F_\alpha(u,v)$ without proof. At an arbitrary point $(\bar{u},\bar{v}) \in H^1(\mathcal{I})^2$ and a fixed $\alpha > 0$, the Fréchet derivative of F_α is given by

$$\forall (\delta^u,\delta^v) \in H^1(\mathcal{I})^2 : \quad F'_\alpha(\bar{u},\bar{v})[\delta^u,\delta^v] = \int_\Omega \phi_\alpha(\bar{u}(x),\bar{v}(x))\,(\eta_\alpha(\bar{u}(x),\bar{v}(x))\delta^u(x) + \eta_\alpha(\bar{v}(x),\bar{u}(x))\delta^v(x))\,\mathrm{d}x, \tag{7}$$

where η_α is a real valued function defined by

$$\forall (a,b) \in \mathbb{R}^2 : \quad \eta_\alpha(a,b) = \frac{a}{\sqrt{a^2+b^2+\alpha}} - 1. \tag{8}$$

Obviously, for any penalty parameter $\gamma > 0$ and smoothing parameter $\alpha \geq 0$, the problem $(\mathrm{P}_{\gamma_k,\alpha_k})$ possesses an optimal solution. The similar argument can be found in [1, Proposition 4.3].

Proposition 2.1. *Let two positive sequences $\{\gamma_k\}_{k\in\mathbb{N}}$ and $\{\alpha_k\}_{k\in\mathbb{N}}$ satisfy $\gamma_k \to +\infty$, $\alpha_k \to 0$ and $\gamma_k\alpha_k \to 0$ as $k \to \infty$. For any $k \in \mathbb{N}$, let $(y_k,u_k,v_k) \in \mathcal{Y} \times H^1(\mathcal{I}) \times H^1(\mathcal{I})$ be a global solution of $(\mathrm{P}_{\gamma_k,\alpha_k})$. Then, $\{(u_k,v_k)\}_{k\in\mathbb{N}}$ contains a subsequence converging strongly in $H^1(\mathcal{I})^2$ to a point $(\bar{u},\bar{v}) \in \mathbb{C}$ and $\bar{y} \in \mathcal{Y}$ be the associated state with $(\bar{u},\bar{v})$ w.r.t. the state equation such that $(\bar{y},\bar{u},\bar{v})$ is an optimal solution of* (OC).

Proof. For any fixed $k \in \mathbb{N}$, we can obtain an upper bound for the objective functional

$$J_{\gamma_k,\alpha_k}(y_k,u_k,v_k) \leq \tfrac{1}{2}\|y_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\gamma_k}{2}\alpha_k|\mathcal{I}|$$

by considering the feasible point $(0,0,0)$ for $(\mathrm{P}_{\gamma_k,\alpha_k})$. Here, $|\mathcal{I}|$ denotes the length of the time interval $\mathcal{I}$. It follows that the sequence of optimal solutions $\{(u_k,v_k)\}_{k\in\mathbb{N}} \subset H^1(\mathcal{I})^2$ is bounded and therefore contains a weakly convergent subsequence (without relabelling) which converges strongly to $(\bar{u},\bar{v}) \in L^2(\mathcal{I})^2$ due to the compact embedding $H^1(\mathcal{I}) \hookrightarrow L^2(\mathcal{I})$. Furthermore, from the estimation

$$0 \leq \|\Phi_{\alpha_k}(\mathrm{E}(u_k),\mathrm{E}(v_k))\|_{L^2(\mathcal{I})} \leq \sqrt{\tfrac{1}{\gamma_k}}\|y_\mathrm{d}\|_{L^2(\Omega)} + \sqrt{\alpha_k|\mathcal{I}|} \to 0$$

we observe that as $k \to \infty$, at least along a subsequence of $\{(u_k,v_k)\}_{k\in\mathbb{N}}$, the sequence $\{\Phi_{\alpha_k}(\mathrm{E}(u_k),\mathrm{E}(v_k))\}_{k\in\mathbb{N}}$ converges pointwise a.e. to 0, i.e. $(\bar{u},\bar{v}) \in \mathbb{C}$.

Let $(u,v) \in \mathbb{C}$ be arbitrarily chosen, for arbitrary positive parameters γ_k und α_k, the estimation

$$\begin{aligned}
\tfrac{\gamma_k}{2}\|\Phi_{\alpha_k}(\mathrm{E}(u),\mathrm{E}(v))\|^2_{L^2(\mathcal{I})} &= \tfrac{\gamma_k}{2}\|\sqrt{\mathrm{E}(u)^2+\mathrm{E}(v)^2+\alpha_k} - \mathrm{E}(u) - \mathrm{E}(v)\|^2_{L^2(\mathcal{I})} \\
&= \tfrac{\gamma_k}{2}\|\sqrt{\mathrm{E}(u)^2+\mathrm{E}(v)^2+\alpha_k} - \sqrt{\mathrm{E}(u)^2+\mathrm{E}(v)^2}\|^2_{L^2(\mathcal{I})} \\
&\leq \tfrac{\gamma_k}{2}\|\sqrt{\alpha_k}\|^2_{L^2(\mathcal{I})} \\
&= \tfrac{\gamma_k}{2}\alpha_k|\mathcal{I}|
\end{aligned}$$

holds true. From the assumption that $\gamma_k \alpha_k \to 0$ as $k \to \infty$, it follows

$$0 \leq \lim_{k\to\infty} \tfrac{\gamma_k}{2} \|\Phi_{\alpha_k}(u,v)\|^2_{L^2(\mathcal{I})} \leq \lim_{k\to\infty} \tfrac{\gamma_k}{2} \alpha_k |\mathcal{I}| = 0,$$

i.e. $\lim_{k\to\infty} \frac{\gamma_k}{2} \|\Phi_{\alpha_k}(u,v)\|^2_{L^2(\mathcal{I})} = 0$. Furthermore, let $y \in \mathcal{Y}$ be the associated state with (u,v) w.r.t. the state equation. Since $(y_k, u_k, v_k) \in H^1(\mathcal{I})^2$ solves the problem $(\mathrm{P}_{\gamma_k,\alpha_k})$, by use of the weak lower semicontinuity of the functionals, we obtain the following estimate:

$$\begin{aligned}
&\tfrac{1}{2}\|\mathrm{D}(y) - y_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda_1}{2}\|u\|_{H^1(\mathcal{I})} + \tfrac{\lambda_2}{2}\|v\|^2_{H^1(\mathcal{I})} \\
&\quad = \tfrac{1}{2}\|\mathrm{D}(y) - y_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda_1}{2}\|u\|_{H^1(\mathcal{I})} + \tfrac{\lambda_2}{2}\|v\|^2_{H^1(\mathcal{I})} + \lim_{k\to\infty} \tfrac{\gamma_k}{2}\|\Phi_{\alpha_k}(u,v)\|^2_{L^2(\mathcal{I})} \\
&\quad = \lim_{k\to\infty} \left(\tfrac{1}{2}\|\mathrm{D}(y) - y_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda_1}{2}\|u\|^2_{H^1(\mathcal{I})} + \tfrac{\lambda_2}{2}\|v\|^2_{H^1(\mathcal{I})} + \tfrac{\gamma_k}{2}\|\Phi_{\alpha_k}(u,v)\|^2_{L^2(\mathcal{I})} \right) \\
&\quad \geq \limsup_{k\to\infty} \left(\tfrac{1}{2}\|\mathrm{D}(y_k) - y_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda_1}{2}\|u_k\|^2_{H^1(\mathcal{I})} + \tfrac{\lambda_2}{2}\|v_k\|^2_{H^1(\mathcal{I})} + \tfrac{\gamma_k}{2}\|\Phi_{\alpha_k}(u_k,v_k)\|^2_{L^2(\mathcal{I})} \right) \\
&\quad \geq \limsup_{k\to\infty} \left(\tfrac{1}{2}\|\mathrm{D}(y_k) - y_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda_1}{2}\|u_k\|^2_{H^1(\mathcal{I})} + \tfrac{\lambda_2}{2}\|v_k\|^2_{H^1(\mathcal{I})} \right) \\
&\quad \geq \liminf_{k\to\infty} \left(\tfrac{1}{2}\|\mathrm{D}(y_k) - y_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda_1}{2}\|u_k\|^2_{H^1(\mathcal{I})} + \tfrac{\lambda_2}{2}\|v_k\|^2_{H^1(\mathcal{I})} \right) \\
&\quad \geq \tfrac{1}{2}\|\mathrm{D}(\bar{y}) - y_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda_1}{2}\|\bar{u}\|^2_{H^1(\mathcal{I})} + \tfrac{\lambda_2}{2}\|\bar{v}\|^2_{H^1(\mathcal{I})}.
\end{aligned}$$

It was shown that $(\bar{y}, \bar{u}, \bar{v}) \in \mathcal{Y} \times H^1(\mathcal{I}) \times H^1(\mathcal{I})$ is a global minimizer of the original optimal control problem (OC). Choosing $y := \bar{y}, u := \bar{u}$ and $v := \bar{v}$, we have

$$\tfrac{1}{2}\|\mathrm{D}(y_k) - y_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda_1}{2}\|u_k\|^2_{H^1(\mathcal{I})} + \tfrac{\lambda_2}{2}\|v_k\|^2_{H^1(\mathcal{I})} \to \tfrac{1}{2}\|\mathrm{D}(\bar{y}) - y_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda_1}{2}\|\bar{u}\|^2_{H^1(\mathcal{I})} + \tfrac{\lambda_2}{2}\|\bar{v}\|^2_{H^1(\mathcal{I})}.$$

It follows

$$\|u_k\|^2_{H^1(\mathcal{I})} + \|v_k\|^2_{H^1(\mathcal{I})} \to \|\bar{u}\|^2_{H^1(\mathcal{I})} + \|\bar{v}\|^2_{H^1(\mathcal{I})}$$

and

$$\|u_k\|_{H^1(\Omega)} \to \|\bar{u}\|_{H^1(\mathcal{I})}, \quad \|v_k\|_{H^1(\Omega)} \to \|\bar{v}\|_{H^1(\mathcal{I})}$$

by use of [1, Lemma A.1]. Due to the weak convergences of $u_k \rightharpoonup \bar{u}$ and $v_k \rightharpoonup \bar{v}$ in the Hilbert space $H^1(\mathcal{I})$, the strong convergences $u_k \to \bar{u}$ and $v_k \to \bar{v}$ follow. □

3 Adjoint state

The finite element method will be used to solve the optimal PDE control problems numerically and it is based on the weak theory of PDEs. In order to ensure that there exists a unique weak solution of state $y \in W_2^{1,0}(Q)$ for any controls $u, v \in H^1(\mathcal{I})$, see [7, Section 26], we give the following assumption on the parabolic PDE (1).

Assumption 3.1. *The coefficient function* $\mathrm{C} \in L^\infty(Q, S^d(\mathbb{R}))$ *(where* $S^d(\mathbb{R})$ *denotes the set of symmetric* $d \times d$ *matrices) satisfies the condition of uniform parabolic, i.e.*

$$\exists c_0 > 0 \; \forall \epsilon \in \mathbb{R}^d : \quad \epsilon^\top \mathrm{C}(x,t)\epsilon \geq c_0 |\epsilon|_2^2 \quad \textit{f.a.a. } (x,t) \in Q.$$

The functions $\mathrm{a} \in L^\infty(Q)$ *and* $\mathrm{q} \in L^\infty(\Sigma)$ *with* $\mathrm{q}(x,t) \geq 0$ *f.a.a. in* Σ *are fixed.*

Let $v \in C^\infty(\overline{Q})$ be an arbitrary test function. We test the first and second line of (1) with v and integrate them over Q. By use of Green's formula w.r.t. the spacial variable x we obtain

$$\begin{aligned} \forall v \in C^\infty(\overline{Q}): \qquad & \iint_Q \frac{d}{dt} y v \mathrm{d}x\mathrm{d}t + \iint_Q ((\mathrm{C}\nabla y)\nabla v + \mathrm{a} y v)\mathrm{d}x\mathrm{d}t + \iint_\Sigma \mathrm{q} y v \mathrm{d}s\mathrm{d}t \\ & = \iint_\Sigma (\mathrm{b}(s)u(t) + \mathrm{c}(s)v(t))v \mathrm{d}s\mathrm{d}t \end{aligned} \tag{9}$$

Due to the third condition in (1) and employing the Green's formula w.r.t. the temporal variable t, it follows

$$\begin{aligned} \forall v \in C^\infty(\overline{Q}): \qquad & \int_\Omega y(x,T)v(x,T)\mathrm{d}x - \iint_Q y \frac{d}{dt} v \mathrm{d}x\mathrm{d}t + \iint_Q ((\mathrm{C}\nabla y)\nabla v + \mathrm{a} y v)\mathrm{d}x\mathrm{d}t \\ & + \iint_\Sigma \mathrm{q} y v \mathrm{d}s\mathrm{d}t = \iint_\Sigma (\mathrm{b}(s)u(t) + \mathrm{c}(s)v(t))v \mathrm{d}s\mathrm{d}t. \end{aligned} \tag{10}$$

It is possible to find a solution of state y from the space $W_2^{1,0}(Q)$ for all test functions $v \in C^\infty(\overline{Q})$. Note that the space $W_2^{1,0}(Q)$ contains all functions from $L^2(Q)$ that are first order weakly differentiable w.r.t. the spacial variables from Ω and their spacial gradient belongs to $L^2(Q; \mathbb{R}^d)$. The existence statement of $y \in W_2^{1,0}(Q)$ of the parabolic PDE (1) can be found in [7, Section 26]. According to [6, Theorem 3.13], the weak solution $y \in W_2^{1,0}(Q)$ belongs, possibly after a suitable modification on a set of measure zero, to the function space $W(0,T)$, i.e. $y \in L^2([0,T]; H^1(\Omega))$ as well as $\frac{d}{dt} y \in L^2([0,T]; H^1(\Omega)^\star)$ hold true. On this account, we consider the operator D as a natural embedding $\mathrm{E}_\Omega : H^1(\Omega) \to L^2(\Omega)$, which maps the abstract function value $y(T)$ from $H^1(\Omega)$ into $L^2(\Omega)$. Furthermore, due to the density of $C^\infty(\overline{Q})$ in $W_2^{1,1}(Q)$, which comprises all functions from $W_2^{1,0}(Q)$ that are first order weakly differentiable w.r.t. time variable from $\mathcal{I}$ and their time derivative $\partial_t y$ belongs to $L^2(Q)$. Note that $W_2^{1,1}(Q)$ is isomorphic to the space $H^1(Q)$, the equations (9) and (10) are valid for all $v \in W_2^{1,1}(Q)$. The affiliation of $v \in W_2^{1,1}(Q)$ can be extended to $v \in W(0,T)$ by means of a density argument as well, see [6].

For the sake of convenience we define a function $a[t;\cdot,\cdot] : W(0,T)^2 \to \mathbb{R}$ in a bilinear form:

$$a[t; y, v] = \int_\Omega ((\mathrm{C}(t)\nabla y(t))\nabla v(t) + \mathrm{a}(t) y(t) v(t))\mathrm{d}x + \int_\Gamma \mathrm{q}(t) y(t) v(t) \mathrm{d}s. \tag{11}$$

To obtain an optimality system which can be solved numerically, we use the formal Lagrange techniques and consider the Lagrange function with the adjoint state $p \in W(0,T)$:

$$\begin{aligned} L(y,u,v,p) = {} & \tfrac{1}{2}\|\mathrm{E}_\Omega[y(\cdot,T)] - y_\mathrm{d}\|^2_{L^2(\Omega)} + \tfrac{\lambda_1}{2}\|u\|^2_{H^1(\mathcal{I})} + \tfrac{\lambda_1}{2}\|v\|^2_{H^1(\mathcal{I})} \\ & + \tfrac{\gamma_k}{2}\|\Phi_{\alpha_k}(\mathrm{E}[u],\mathrm{E}[v])\|^2_{L^2(\mathcal{I})} - \int_0^T \left(\tfrac{d}{dt}y(t), p(t)\right)_{H^1(\Omega)^\star, H^1(\Omega)} \mathrm{d}t - \int_0^T a[t;y,p]\mathrm{d}t \\ & + \iint_\Sigma (b(s)u(t) + c(s)v(t))p(s,t)\mathrm{d}s\mathrm{d}t. \end{aligned} \tag{12}$$

The initial condition of the state will be treated explicitly. For all $\xi \in W(0,T)$ with $\xi(\cdot,0) = 0$, the following necessary optimality condition is valid

$$\langle \mathrm{E}_\Omega[y(\cdot,T)] - y_\mathrm{d}, \mathrm{E}_\Omega[\xi(\cdot,T)]\rangle_{L^2(\Omega)} - \int_0^T \left(\tfrac{d}{dt}\xi(t), p(t)\right)_{H^1(\Omega)^\star, H^1(\Omega)} \mathrm{d}t - \int_0^T a[t;\xi,p]\mathrm{d}t = 0$$

By use of the Green's formula w.r.t. the temporal integral, we can write

$$\begin{aligned} \int_\Omega \xi(x,T)p(x,T)\mathrm{d}x - \int_\Omega \xi(x,0)p(x,0)\mathrm{d}x - \iint_Q \xi\tfrac{d}{dt}p\mathrm{d}x\mathrm{d}t + \int_0^T a[t;\xi,p]\mathrm{d}t \\ = \int_\Omega (y(x,T) - y_\mathrm{d})\xi(x,T)\mathrm{d}x, \end{aligned}$$

which corresponds to the weak formulation of the so-called adjoint PDE:

$$\begin{aligned} -\tfrac{d}{dt}p - \nabla\cdot(\mathrm{C}\nabla p) + \mathrm{a}p = 0 \qquad & \text{a.e. on } Q := \Omega\times\mathcal{I} \\ \vec{n}\cdot(\mathrm{C}\nabla p) + \mathrm{q}p = 0 \qquad & \text{a.e. on } \Sigma := \Gamma\times\mathcal{I} \\ p(\cdot,T) = y(\cdot,T) - y_\mathrm{d} \qquad & \text{a.e. on } \Omega. \end{aligned} \tag{13}$$

4 First-optimize-then-discretize approach (Indirect method)

The following equations together with the systems (1) and (13) build the necessary conditions for the penalized problem $(\mathrm{P}_{\gamma_k,\alpha_k})$.

$$\begin{aligned} \lambda_1\langle u,\varrho\rangle_{H^1(\mathcal{I})} + \gamma_k\langle\Psi_{\alpha_k}(\mathrm{E}[u],\mathrm{E}[v]),\mathrm{E}[\varrho]\rangle_{L^2(\mathcal{I})} + \iint_\Sigma b(s)\varrho(t)p(s,t)\mathrm{d}s\mathrm{d}t = 0 \\ \lambda_2\langle v,\varrho\rangle_{H^1(\mathcal{I})} + \gamma_k\langle\Psi_{\alpha_k}(\mathrm{E}[v],\mathrm{E}[u]),\mathrm{E}[\varrho]\rangle_{L^2(\mathcal{I})} + \iint_\Sigma c(s)\varrho(t)p(s,t)\mathrm{d}s\mathrm{d}t = 0, \end{aligned} \tag{14}$$

where $\Psi_{\alpha_k} : L^2(\mathcal{I})\times L^2(\mathcal{I}) \to L^2(\mathcal{I})$ is defined by

$$\forall (w,z)\in L^2(\mathcal{I})^2\ \forall x\in\mathcal{I}: \quad \Psi_{\alpha_k}(w,z) = \psi_{\alpha_k}(w(t),z(t))$$

with

$$\forall (a,b) \in \mathbb{R}^2 : \quad \psi_{\alpha_k}(a,b) = \phi_{\alpha_k}(a,b)\eta_{\alpha_k}(a,b),$$

and $\phi_{\alpha_k}, \eta_{\alpha_k}$ are given in (4) and (8), respectively. Note that the equations (14) are valid for all functions $\varrho \in H^1(\mathcal{I})$.

The derived conditions (1), (13) and (14) will be discretized into a nonlinear system by considering the combination of (semi-)discretization w.r.t. the spacial dimension and finite difference methode w.r.t. the temporal interval. We test (1) with an arbitrary test function $\zeta \in H^1(\Omega)$ and integrate over the domain Ω on a fixed time point t. The weak formulation reads

$$\begin{aligned} \int_\Omega \frac{d}{dt} y(x,t)\zeta(x)\mathrm{d}x + a[t;y,\zeta] &= \int_\Gamma \mathrm{b}(s)u(t)\zeta(s)\mathrm{d}s + \int_\Gamma \mathrm{c}(s)v(t)\zeta(s)\mathrm{d}s \\ \int_\Omega y(x,0)\zeta(x)\mathrm{d}x &= 0. \end{aligned} \tag{15}$$

Note that all functions except test functions are considered to be abstract functions, which map from the time interval to suitable function spaces. Let Ω_Δ and Γ_Δ denote the discrete domain of Ω and its boundary Γ. The number of nodes and elements on Ω_Δ is denoted by n_p and n_e, respectively. The finite element space $\mathcal{P}^0(\Omega_\Delta)$ with piecewise constant functions is cognizant for the discretized coefficient functions $\vec{\mathrm{C}}(t), \vec{\mathrm{a}}(t), \vec{\mathrm{b}}(t)$, $\vec{\mathrm{c}}(t)$ and $\vec{\mathrm{q}}(t)$. The discretized state $\vec{y}(t)$ is considered in the finite element space $\mathcal{P}^1(\Omega_\Delta)$ with piecewise affine linear functions, which possesses the stiffness matrix $K_{\Omega_\Delta}(\cdot)$ and mass matrix $M^1_{\Omega_\Delta}(\cdot)$ associated with considering coefficient matrices. W.r.t. the boundary condition, the matrix $Q_{\Gamma_\Delta}(\vec{\mathrm{q}}(t))$ represents the boundary integral of involving $\vec{\mathrm{q}}(t)$. Additionally, $G_{\Gamma_\Delta}(\vec{\mathrm{b}})$, $G_{\Gamma_\Delta}(\vec{\mathrm{c}})$ are vectors depending on the discrete condition $\vec{\mathrm{b}}$ and $\vec{\mathrm{c}}$. The semi-discretization of the equation (15) is obtained:

$$\begin{aligned} \frac{d}{dt} M^1_{\Omega_\Delta}(1)\vec{y}(t) + \left(K_{\Omega_\Delta}(\vec{\mathrm{C}}(t)) + M^1_{\Omega_\Delta}(\vec{\mathrm{a}}(t)) + Q_{\Gamma_\Delta}(\vec{\mathrm{q}}(t))\right)\vec{y}(t) &= G_{\Gamma_\Delta}(\vec{\mathrm{b}})\vec{u}(t) + G_{\Gamma_\Delta}(\vec{\mathrm{c}})\vec{v}(t) \\ M^1_{\Omega_\Delta}(1)\vec{y}(0) &= 0 \end{aligned} \tag{16}$$

We use the implicit Euler method to deal with the ODE w.r.t. the time interval $\mathcal{I}$. The temporal discretization is realized as the family $\mathcal{I}_\Delta := \{[t_i, t_{i+1}]\}_{i=0}^{n-1}$ with n subintervals. The vectors $\vec{y} = (\vec{y}^0, \ldots, \vec{y}^n)^\top$, $\vec{u} = (\vec{u}^0, \ldots, \vec{u}^n)^\top$ and $\vec{v} = (\vec{v}^0, \ldots, \vec{v}^n)^\top$ are discretized state and controls. The elements $\vec{y}^i, \vec{u}^i, \vec{v}^i$ correspond to their approximation at each time t_i for $i = 0, \ldots, n$. The temporal discretization of coefficient functions is in a similar fashion. By use of the forward differences to consider the derivative w.r.t. time, we obtain a linear system

$$\mathcal{A}\vec{y} - \mathcal{B}(\vec{\mathrm{b}})\vec{u} - \mathcal{B}(\vec{\mathrm{c}})\vec{v} = 0 \tag{17}$$

where $\mathcal{A} \in \mathbb{R}^{(n+1)n_p \times (n+1)n_p}$ is given by

$$\mathcal{A} = \begin{pmatrix} M^1_{\Omega_\Delta}(1) & \mathrm{N} & \dots & \dots & \dots & \mathrm{N} \\ -M^1_{\Omega_\Delta}(1) & \Pi^0 & \ddots & & & \vdots \\ \mathrm{N} & -M^1_{\Omega_\Delta}(1) & \Pi^1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & -M^1_{\Omega_\Delta}(1) & \Pi^{n-2} & \mathrm{N} \\ \mathrm{N} & \dots & \dots & \mathrm{N} & -M^1_{\Omega_\Delta}(1) & \Pi^{n-1} \end{pmatrix}$$

with $\Pi^i := M^1_{\Omega_\Delta}(1) + (t_{i+1} - t_i)\left(K_{\Omega_\Delta}(\vec{\mathrm{C}}^{i+1}) + M^1_{\Omega_\Delta}(\vec{\mathrm{a}}^{i+1}) + Q_{\Gamma_\Delta}(\vec{\mathrm{q}}^{i+1})\right), i = 0, \dots, n-1$; the matrix $\mathcal{B}(\rho) \in \mathbb{R}^{(n+1)n_p \times (n+1)}$ is given by

$$\mathcal{B}(\rho) = \begin{pmatrix} \mathrm{n} & \mathrm{n} & \dots & \dots & \mathrm{n} \\ \mathrm{n} & (t_1 - t_0)G_{\Gamma_\Delta}(\rho) & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & (t_{n-1} - t_{n-2})G_{\Gamma_\Delta}(\rho) & \mathrm{n} \\ \mathrm{n} & \dots & \mathrm{n} & & (t_n - t_{n-1})G_{\Gamma_\Delta}(\rho) \end{pmatrix}$$

with $\rho \in \{\vec{\mathrm{b}}, \vec{\mathrm{c}}\}$. Furthermore, $\mathrm{N} \in \mathbb{R}^{n_p \times n_p}$, $\mathrm{n} \in \mathbb{R}^{n_p}$ are all-zero matrix and vector, respectively.

Next, we discretize the adjoint PDE (13). Let us additionally denote the vector $\vec{p} = (\vec{p}^0, \dots, \vec{p}^n)^\top$ as the discretized adjoint. Applying the Green's formula on the first two lines of (13) w.r.t. the spacial variable x, the weak formulation reads

$$\begin{aligned} -\int_\Omega \frac{d}{dt} p(x,t)\zeta(x)\mathrm{d}x + a[t; p, \zeta] = 0 \\ \int_\Omega p(x,T)\zeta(x)\mathrm{d}x = \int_\Omega (y(x,T) - y_\mathrm{d}(x))\zeta(x)\mathrm{d}x \end{aligned} \tag{18}$$

and the associated semi-discretization is

$$\begin{aligned} -\frac{d}{dt} M^1_{\Omega_\Delta}(1)\vec{p}(t) + \left(K_{\Omega_\Delta}(\vec{\mathrm{C}}(t)) + M^1_{\Omega_\Delta}(\vec{\mathrm{a}}(t)) + Q_{\Gamma_\Delta}(\vec{\mathrm{q}}(t))\right)\vec{p}(t) = 0 \\ M^1_{\Omega_\Delta}(1)\vec{p}(T) = M^1_{\Omega_\Delta}(1)\vec{y}(T) - M^1_{\Omega_\Delta}(1)E^{01}_{\Omega_\Delta}\vec{y}_\mathrm{d}. \end{aligned} \tag{19}$$

Here, $E^{01}_{\Omega_\Delta}$ transforms the approximation of the desired state y_d from $\mathcal{P}^0(\Omega_\Delta)$ into $\mathcal{P}^1(\Omega_\Delta)$. To deal with the discretization w.r.t. the temporal dimension, we have to take the reverse time orientation into consideration. We define

$$\Lambda^i =: M^1_{\Omega_\Delta}(1) + (t_{i+1} - t_i)\left(K_{\Omega_\Delta}(\vec{\mathrm{C}}^i) + M^1_{\Omega_\Delta}(\vec{\mathrm{a}}^i) + Q_{\Gamma_\Delta}(\vec{\mathrm{q}}^i)\right), \quad i = 0, \dots, n-1$$

and obtain

$$-\tilde{\mathcal{A}}\vec{p} + \mathcal{L}\vec{y} + b = 0, \tag{20}$$

where $\tilde{\mathcal{A}}, \mathcal{L} \in \mathbb{R}^{(n+1)n_p \times (n+1)n_p}$, $b \in \mathbb{R}^{(n+1)n_p}$ are given by

$$\tilde{\mathcal{A}} = \begin{pmatrix} \Lambda^0 & -M^1_{\Omega_\Delta}(1) & \dots & \dots & \dots & \mathrm{N} \\ \mathrm{N} & \Lambda^1 & -M^1_{\Omega_\Delta}(1) & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \Lambda^{n-1} & -M^1_{\Omega_\Delta}(1) \\ \mathrm{N} & \dots & \dots & \dots & \dots & M^1_{\Omega_\Delta}(1). \end{pmatrix},$$

$$\mathcal{L} = \begin{pmatrix} \mathrm{N} & \dots & \dots & \mathrm{N} \\ \vdots & \ddots & \ddots & \vdots \\ \mathrm{N} & \dots & \dots & M^1_{\Omega_\Delta}(1) \end{pmatrix}, \qquad b = \begin{pmatrix} \mathrm{n} \\ \vdots \\ \mathrm{n} \\ -M^1_{\Omega_\Delta}(1) E^{01}_{\Omega_\Delta} \vec{y}_{\mathrm{d}}. \end{pmatrix}$$

Now, we discretize the conditions (14) related to controls. Firstly, we consider finite element spaces $\mathcal{P}^0(\mathcal{I}_\Delta)$ and $\mathcal{P}^1(\mathcal{I}_\Delta)$ on the discretized time interval $\mathcal{I}_\Delta$, which possess mass matrices $M^0_{\mathcal{I}_\Delta}$ and $M^1_{\mathcal{I}_\Delta}$, respectively. The matrix $K_{\mathcal{I}_\Delta}$ is the stiffness matrix in the finite element space $\mathcal{P}^1(\mathcal{I}_\Delta)$. Obviously, the discretized interval $\mathcal{I}_\Delta$ possesses $n+1$ nodes and n elements. The discrete counterparts of (14) read

$$\begin{aligned} &\lambda_1(K_{\mathcal{I}_\Delta} + M^1_{\mathcal{I}_\Delta})\vec{u} + \gamma_k {E^{10}_{\mathcal{I}_\Delta}}^\top M^0_{\mathcal{I}_\Delta} \vec{\psi}_{\alpha_k}(E^{10}_{\mathcal{I}_\Delta}\vec{u}, E^{10}_{\mathcal{I}_\Delta}\vec{v}) + M^1_{\mathcal{I}_\Delta} \mathrm{diag}(G_{\Gamma_\Delta}(\vec{\mathrm{b}})^\top)\vec{p} = 0 \\ &\lambda_2(K_{\mathcal{I}_\Delta} + M^1_{\mathcal{I}_\Delta})\vec{v} + \gamma_k {E^{10}_{\mathcal{I}_\Delta}}^\top M^0_{\mathcal{I}_\Delta} \vec{\psi}_{\alpha_k}(E^{10}_{\mathcal{I}_\Delta}\vec{v}, E^{10}_{\mathcal{I}_\Delta}\vec{u}) + M^1_{\mathcal{I}_\Delta} \mathrm{diag}(G_{\Gamma_\Delta}(\vec{\mathrm{c}})^\top)\vec{p} = 0, \end{aligned} \tag{21}$$

therein, $E^{10}_{\mathcal{I}_\Delta}$ transforms the discrete controls from $\mathcal{P}^1(\mathcal{I}_\Delta)$ into $\mathcal{P}^0(\mathcal{I}_\Delta)$, which corresponds the natural embedding operator $\mathrm{E} : H^1(\mathcal{I}) \hookrightarrow L^2(\mathcal{I})$. For $\rho \in \{\vec{\mathrm{b}}, \vec{\mathrm{c}}\}$, the matrix $\mathrm{diag}(G_{\Gamma_\Delta}(\rho)^\top) \in \mathbb{R}^{(n+1)\times(n+1)n_p}$ consists of transposed $G_{\Gamma_\Delta}(\rho)$ as blocks in diagonal. The partial derivative of squared smoothed Fischer Burmeister function w.r.t. vectors of discretized controls is calculated by use of $\psi_{\alpha_k}(a,b)$ in componentenwise fashion. An advantage of the "first-optimize-then-discretize" approach for solving this surrogate problem ($\mathrm{P}_{\gamma_k,\alpha_k}$) is the easy realization of first and second derivation of the penalized (smoothed) Fischer Burmeister function.

Finally, the following nonlinear equation system which is comprised of discrete equations (17), (20) and (21), w.r.t. vector-formed discretized variables $\vec{z} = (\vec{y}, \vec{u}, \vec{v}, \vec{p})$ has to be solved:

$$\mathcal{F}_{OD}\vec{z} + f_{OD} + \mathcal{K}_{OD} = 0, \tag{22}$$

where $\mathcal{F}_{OD} \in \mathbb{R}^{(2(n+1)n_p+2(n+1))\times(2(n+1)n_p+2(n+1))}$ is the linear part in this equation system which is given by

$$\mathcal{F}_{OD} = \begin{pmatrix} \mathcal{L} & \mathcal{N}_2^\top & \mathcal{N}_2^\top & -\tilde{\mathcal{A}} \\ \mathcal{N}_2 & \lambda_1(K_{I_\Delta} + M^1_{I_\Delta}) & \mathcal{N}_3 & M^1_{I_\Delta}\operatorname{diag}(G_{\Gamma_\Delta}(\vec{\mathrm{b}}^\top) \\ \mathcal{N}_2 & \mathcal{N}_3 & \lambda_2(K_{I_\Delta} + M^1_{I_\Delta}) & M^1_{I_\Delta}\operatorname{diag}(G_{\Gamma_\Delta}(\vec{\mathrm{c}}^\top) \\ -\mathcal{A} & \mathcal{B}(\vec{\mathrm{b}}) & \mathcal{B}(\vec{\mathrm{c}}) & \mathcal{N}_1 \end{pmatrix}. \tag{23}$$

Note that the matrix $\mathcal{F}_{OD}$ is not symmetric. The vectors $f_{OD}, \mathcal{K}_{OD} \in \mathbb{R}^{2(n+1)n_p+2(n+1)}$ are nonlinear part and constant part, respectively, i.e.

$$f_{OD} = \begin{pmatrix} \mathbf{n} \\ \gamma_k {E^{10}_{I_\Delta}}^\top M^0_{I_\Delta} \vec{\psi}_{\alpha_k}(E^{10}_{I_\Delta}\vec{u}, E^{10}_{I_\Delta}\vec{v}) \\ \gamma_k {E^{10}_{I_\Delta}}^\top M^0_{I_\Delta} \vec{\psi}_{\alpha_k}(E^{10}_{I_\Delta}\vec{v}, E^{10}_{I_\Delta}\vec{u}) \\ \mathbf{n} \end{pmatrix}, \qquad \mathcal{K}_{OD} = \begin{pmatrix} b \\ \mathbf{n} \\ \mathbf{n} \\ \mathbf{n} \end{pmatrix}. \tag{24}$$

The notations $\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3, \mathbf{n}$ stand for zero matrices and vectors in suitable dimensions.

5 First-discretize-then-optimize approach (Direct method)

The first-discretize-then-optimize approach is also called direct method. At first, one discretizes the objective functional and constraint conditions in suitable finite element spaces and needs to solve a standard finite dimensional problem. W.r.t. the penalized problem ($\mathrm{P}_{\gamma_k,\alpha_k}$), the adjoint state only exists in a discrete form and corresponds the multiplier associated with the linear discretized state equation. By use of the same notations for discretized variables as used in the previous section, the discretized counterpart of the penalized problem ($\mathrm{P}_{\gamma_k,\alpha_k}$) reads:

$$\begin{aligned} \vec{J}_{\gamma_k,\alpha_k}(\vec{y},\vec{u},\vec{v}) = {} & \tfrac{1}{2}\left(E^{10}_{\Omega_\Delta}\vec{y}^n - \vec{y}_{\mathrm{d}}\right)^\top M^0_{\Omega_\Delta}(1)\left(E^{10}_{\Omega_\Delta}\vec{y}^n - \vec{y}_{\mathrm{d}}\right) + \tfrac{\lambda_1}{2}\vec{u}^\top(K_{I_\Delta} + M^1_{I_\Delta})\vec{u} \\ & + \tfrac{\lambda_2}{2}\vec{v}^\top(K_{I_\Delta} + M^1_{I_\Delta})\vec{v} + \tfrac{\gamma_k}{2}\left(\vec{\phi}_{\alpha_k}(E^{10}_{I_\Delta}\vec{u}, E^{10}_{I_\Delta}\vec{v})\right)^\top M^0_{I_\Delta}(1)\left(\vec{\phi}_{\alpha_k}(E^{10}_{I_\Delta}\vec{u}, E^{10}_{I_\Delta}\vec{v})\right) \to \max_{\vec{y},\vec{u},\vec{v}} \\ & \mathcal{A}\vec{y} - \mathcal{B}(\vec{\mathrm{b}})\vec{u} - \mathcal{B}(\vec{\mathrm{c}})\vec{v} = 0, \end{aligned} \tag{25}$$

therein, in addition to the notations which have been introduced in Section 4, $M^0_{\Omega_\Delta}(1)$ means the mass matrix in finite element space $\mathcal{P}^0(\Omega_\Delta)$ and $E^{10}_{\Omega_\Delta}$ transforms the approximation of functions from $\mathcal{P}^1(\Omega_\Delta)$ into $\mathcal{P}^0(\Omega_\Delta)$.

At this point, we need to solve the first order optimality system of (25)

$$\mathcal{F}_{DO}\vec{z} + f_{DO} + \mathcal{K}_{DO} = 0, \tag{26}$$

which is totally different from (22) and explicitly,

$$\mathcal{F}_{DO} = \begin{pmatrix} \tilde{\mathcal{L}} & \mathcal{N}_2^\top & \mathcal{N}_2^\top & -\mathcal{A}^\top \\ \mathcal{N}_2 & \lambda_1(K_{\mathcal{I}_\Delta} + M^1_{\mathcal{I}_\Delta}) & \mathcal{N}_3 & \mathcal{B}(\vec{\mathrm{b}})^\top \\ \mathcal{N}_2 & \mathcal{N}_3 & \lambda_2(K_{\mathcal{I}_\Delta} + M^1_{\mathcal{I}_\Delta}) & \mathcal{B}(\vec{\mathrm{c}})^\top \\ -\mathcal{A} & \mathcal{B}(\vec{\mathrm{b}}) & \mathcal{B}(\vec{\mathrm{c}}) & \mathcal{N}_1 \end{pmatrix}, \tag{27}$$

with

$$\tilde{\mathcal{L}} = \begin{pmatrix} \mathrm{N} & \dots & \dots & \mathrm{N} \\ \vdots & \ddots & \ddots & \vdots \\ \mathrm{N} & \dots & \dots & {E^{10}_{\Omega_\Delta}}^\top M^0_{\Omega_\Delta}(1) E^{10}_{\Omega_\Delta}. \end{pmatrix}$$

Note that the matrix $\mathcal{F}_{OD}$ is symmetric. On one side, we don't need to consider the adjoint state equation (13), on the other side, it is strenuous to calculate the gradient of the discrete penalty term

$$\vec{F}_{\alpha_k}(\vec{u}, \vec{v}) = \tfrac{1}{2} \left(\vec{\phi}_{\alpha_k}(E^{10}_{\mathcal{I}_\Delta}\vec{u}, E^{10}_{\mathcal{I}_\Delta}\vec{v}) \right)^\top M^0_{\mathcal{I}_\Delta}(1) \left(\vec{\phi}_{\alpha_k}(E^{10}_{\mathcal{I}_\Delta}\vec{u}, E^{10}_{\mathcal{I}_\Delta}\vec{v}) \right)$$

w.r.t. the vector variables $\vec{u}$ and $\vec{v}$. The vectors $f_{DO}, \mathcal{K}_{DO} \in \mathbb{R}^{2(n+1)n_p+2(n+1)}$ are given as

$$\begin{aligned} f_{DO} &= \begin{pmatrix} \mathbf{n} \\ \gamma_k {E^{10}_{\mathcal{I}_\Delta}}^\top \mathrm{diag}(\eta(E^{10}_{\mathcal{I}_\Delta}\vec{u}, E^{10}_{\mathcal{I}_\Delta}\vec{v}) M^0_{\mathcal{I}_\Delta} \phi_{\alpha_k}(E^{10}_{\mathcal{I}_\Delta}\vec{u}, E^{10}_{\mathcal{I}_\Delta}\vec{v}) \\ \gamma_k {E^{10}_{\mathcal{I}_\Delta}}^\top \mathrm{diag}(\eta(E^{10}_{\mathcal{I}_\Delta}\vec{v}, E^{10}_{\mathcal{I}_\Delta}\vec{u}) M^0_{\mathcal{I}_\Delta} \phi_{\alpha_k}(E^{10}_{\mathcal{I}_\Delta}\vec{u}, E^{10}_{\mathcal{I}_\Delta}\vec{v}) \\ \mathbf{n} \end{pmatrix}, \\ \mathcal{K}_{DO} &= \begin{pmatrix} \tilde{b} \\ \mathbf{n} \\ \mathbf{n} \\ \mathbf{n} \end{pmatrix} \quad \text{with} \quad \tilde{b} = \begin{pmatrix} \mathrm{n} \\ \vdots \\ \mathrm{n} \\ -{E^{10}_{\Omega_\Delta}}^\top M^0_{\Omega_\Delta}(1)\vec{y}_\mathrm{d} \end{pmatrix}. \end{aligned} \tag{28}$$

6 Numerical experiment

The following example will be performed with the object oriented MATLAB class library OOPDE, see [5]. Obviously, the equation system (26) obtained by first-discretize-then-optimize approach is totally different from the system (22) obtained by first-optimize-then-discretize approach. For solving both nonlinear equations, we make use of damped Newton method. Since $M^1_{\Omega_\Delta}(1)$ and ${E^{10}_{\Omega_\Delta}}^\top M^0_{\Omega_\Delta}(1) E^{10}_{\Omega_\Delta}$, $M^1_{\Omega_\Delta}(1) E^{01}_{\Omega_\Delta}$ and ${E^{10}_{\Omega_\Delta}}^\top M^0_{\Omega_\Delta}(1)$ don't have great difference, respectively, the main distinction in the linear parts of both equation systems is in the discrete adjoint

operator. The order of consideration for the penalty term of control complementarity constraints leads to different nonlinear parts in (22) and (26), respectively. One expects dissimilar solutions for a same example. But actually, both solutions are correct with reservations of respective discretization inaccuracy. We will evaluate them by means of performance profile of associated objective functional values and complementarity feasibilities, which was introduced in [2].

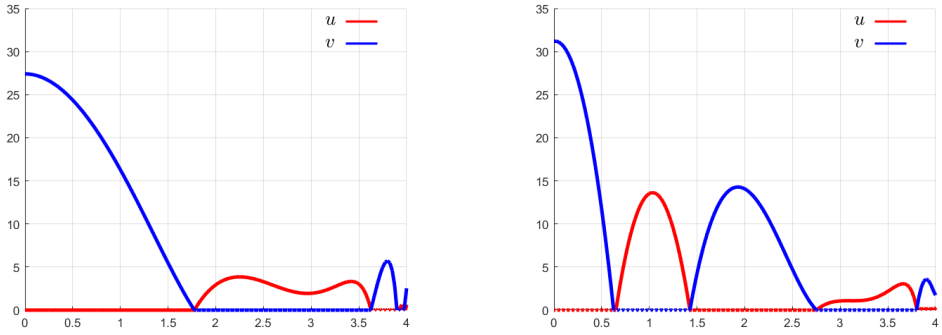
Let $\Omega := (0,1)$ be a one dimensional interval. We have a desired state

$$\forall x \in \Omega: \quad y_{\mathrm{d}}(x) := \begin{cases} 1 & 0 < x \leq \frac{1}{2} \\ 2 & \frac{1}{2} < x < 1, \end{cases}$$

which is discontinuous on Ω. The controls u, v are active on the time interval $\mathcal{I} := (0,4)$ at the respective boundary points of Ω. Explicitly, the associated coefficient functions b, c satisfy

$$\forall s \in \{0,1\}: \quad b(s) := \begin{cases} 1 & s = 0 \\ 0 & s = 1, \end{cases} \qquad c(s) := \begin{cases} 0 & s = 0 \\ 1 & s = 1. \end{cases}$$

Furthermore, we have $\mathrm{C} \equiv 0.0125$, $\mathrm{a} \equiv 0$, and $\mathrm{q} \equiv 1$ in the parabolic state equation. The Tikhonov parameters $\lambda_1 = \lambda_2 := 10^{-5}$ are fixed. We use the same equidistant stepsize $h = 0.025$ to discretize the spacial interval Ω and the temporal interval $\mathcal{I}$, i.e. Ω_Δ and $\mathcal{I}_\Delta$ have 40 and 160 equidistant subintervals, respectively. For the sake of simplicity, we set $\alpha_k = 1/\gamma_k^2$ for all $k \in \mathbb{N}$. One lets $\gamma_0 = 1$ and $\gamma_{k+1} = 1.2 * \gamma_k$ for implementing the Newton iteration process. The starting point is set to be $\vec{z}_0 = 1$. Finally, we obtain the solutions of controls, which are illustrated in the following Figure 1.

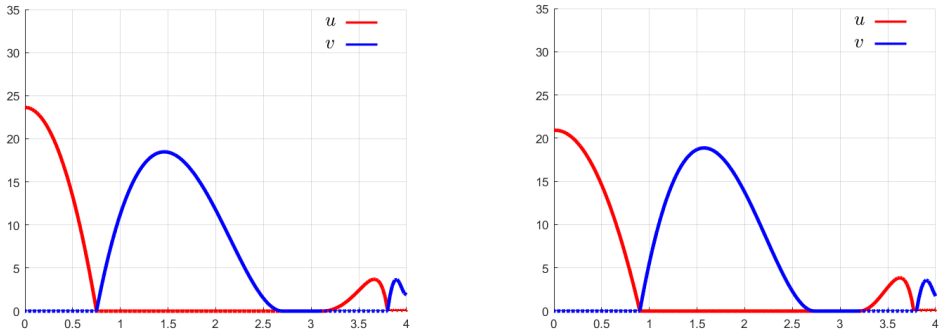


(a) Controls obtained by indirect method (b) Controls obtained by direct method

Figure 1: Solutions of controls with constant 1 as starting point.

It is already known from the study in [2], that one may yield reasonable "good" solutions by use of the optimal solution of the optimal control problem (OC) with

only nonnegative constraints on controls, i.e. $(u, v) \in H^1_+(\mathcal{I})^2$ as starting points. Here, "good" means that the obtained solutions may lie in the neighborhood of the potential global optimal solutions due to the local convergence property of the Newton-type method. The solutions of (22) and (26) by use of positive controls as starting point in Newton-type solving process are illustrated in Figure 2. There isn't great difference any more, at least visually.



(a) Controls obtained by indirect method (b) Controls obtained by direct method

Figure 2: Solutions of controls with positive controls as starting point.

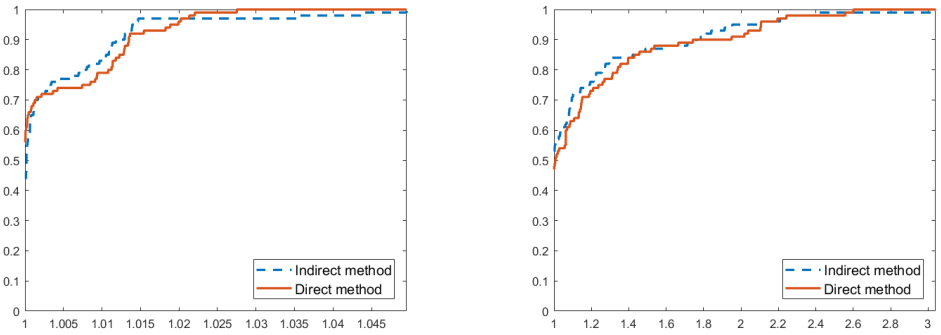
To compare the robustness of both discretization approaches, we arbitrarily choose 100 vectors whose components from $[0, 20]$ serve as staring points in Newton-type solving process. Due to the regularity of Hessian matrix of squared smoothed Fischer Burmeister for a reasonable positive parameter α_k, the applied damped Newton method always deliver an accumulation point from any starting point. For any pair of discrete controls $(\vec{u}, \vec{v})$, we compare the associated objective function values and feasibility of control complementarity associated with all the outputs, which are defined by

$$\vec{J}(\vec{y}, \vec{u}, \vec{v}) = \tfrac{1}{2}\left(E^{10}_{\Omega_\Delta}\vec{y}^n - \vec{y}_\mathrm{d}\right)^\top M^0_{\Omega_\Delta}(1)\left(E^{10}_{\Omega_\Delta}\vec{y}^n - \vec{y}_\mathrm{d}\right) + \tfrac{\lambda_1}{2}\vec{u}^\top(K_{\mathcal{I}_\Delta} + M^1_{\mathcal{I}_\Delta})\vec{u} + \tfrac{\lambda_2}{2}\vec{v}^\top(K_{\mathcal{I}_\Delta} + M^1_{\mathcal{I}_\Delta})\vec{v}$$

and

$$\Theta(\vec{u}, \vec{v}) = \left(\left(\sqrt{(E^{10}_{\mathcal{I}_\Delta}\vec{u})^2 + (E^{10}_{\mathcal{I}_\Delta}\vec{v})^2} - E^{10}_{\mathcal{I}_\Delta}\vec{u} - E^{10}_{\mathcal{I}_\Delta}\vec{v}\right)^\top M^0_{\mathcal{I}_\Delta}\left(\sqrt{(E^{10}_{\mathcal{I}_\Delta}\vec{u})^2 + (E^{10}_{\mathcal{I}_\Delta}\vec{v})^2} - E^{10}_{\mathcal{I}_\Delta}\vec{u} - E^{10}_{\mathcal{I}_\Delta}\vec{v}\right)\right)^{\frac{1}{2}},$$

respectively. In the same line as the computation of derivative of discrete squared (smoothed) Fischer Burmeister function, all the powers and square roots are implemented componentwise. The quantitative comparison of these values is realized by illustration of associated performance profiles, see Figure 3. Here, the comparative values are set to be the minimum of all 200 associated function values for Figure 3a and zero as the absolute feasibility of controls for Figure 3b, respectively. A detailed description for calculating the performance profiles can be found in [2].



(a) Function value

(b) Feasibility of complementarity

Figure 3: Comparison of indirect and direct methods via performance profiles w.r.t. function values and feasibility of control complementarity.

It is visually significant that first-optimize-then-discretize and first-discretize-then-optimize methods used to solve the infinite dimensional optimization problem are not indistinguishable regarding to resulted function values and feasibility of obtained controls. The results of this example answer our expectation as well.

References

[1] C. Clason et al. “Optimal control problems with control complementarity constraints: existence results, optimality conditions, and a penalty method”. In: *Optimization Methods and Software* 35.1 (2020), pp. 142–170. DOI: 10.1080/10556788.2019.1604705.

[2] Y. Deng, P. Mehlitz, and U. Prüfert. “Coupled versus decoupled penalization of control complementarity constraints”. In: *ESAIM: Control, Optimisation and Calculus of Variations* (2021). DOI: 10.1051/cocv/2021022.

[3] A. Fischer. “A special Newton-type optimization method”. In: *Optimization* 24.3-4 (1992), pp. 269–284. DOI: 10.1080/02331939208843795.

[4] H. Goldberg, W. Kampowsky, and F. Tröltzsch. “On Nemytskij operators in L_p-spaces of abstract functions”. In: *Mathematische Nachrichten* 155 (1992), pp. 127–140. DOI: 10.1002/mana.19921550110.

[5] U. Prüfert. *OOPDE: An object oriented toolbox for finite elements in Matlab*. TU Bergakademie Freiberg, 2015. URL: http://www.mathe.tu-freiberg.de/files/personal/255/oopde-quickstart-guide-2015.pdf.

[6] F. Tröltzsch. *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*. Providence, RI: American Mathematical Society, 2010. DOI: 10.1090/gsm/112.

[7] J. Wolka. *Partielle Differentialgleichungen*. Leipzig: Teubner, 1982. ISBN: 3-519-02225-7.