Computational Optimization and Applications (2019) 72:413–430
https://doi.org/10.1007/s10589-018-0050-y

# Solving optimal control problems with terminal complementarity constraints via Scholtes' relaxation scheme

**Francisco Benita**[1] · **Patrick Mehlitz**[2]

Received: 2 August 2018 / Published online: 15 December 2018


**Abstract**
We investigate the numerical treatment of optimal control problems of linear ordinary differential equations with terminal complementarity constraints. Therefore, we generalize the well-known relaxation technique of Scholtes to the problem at hand. In principle, any other relaxation approach from finite-dimensional complementarity programming can be adapted in similar fashion. It is shown that the suggested method possesses strong convergence properties under mild assumptions. Finally, some numerical examples are presented.

**Keywords** Complementarity-constrained programming · Optimal control · Relaxation

**Mathematics Subject Classification** 49K15 · 49M20

✉ Patrick Mehlitz
mehlitz@b-tu.de
https://www.b-tu.de/fg-optimale-steuerung/team/dr-patrick-mehlitz

Francisco Benita
francisco_benita@sutd.edu.sg

[1] Singapore University of Technology and Design, Singapore 487372, Singapore

[2] Chair of Optimal Control, Brandenburgische Technische Universität Cottbus-Senftenberg, 03046 Cottbus, Germany

## 1 Introduction

For some positive time $T > 0$ and the associated time interval $I := (0, T)$, we consider the following *optimal control problem with terminal complementarity constraints*

$$\begin{aligned} J(x,u) &\to \min_{x,u} \\ \mathbb{S}(u) - x &= 0 \\ g_i(x(T)) &\leq 0 \qquad i \in \mathcal{L} \\ 0 \leq G_j(x(T)) \perp H_j(x(T)) &\geq 0 \qquad j \in \mathcal{K}. \end{aligned} \tag{OCTCC}$$

Here, we use the index sets $\mathcal{L} := \{1, \ldots, l\}$ and $\mathcal{K} := \{1, \ldots, k\}$. Furthermore, the mapping $\mathbb{S}\colon L^2(I)^m \to H^1(I)^n$ represents the operator which assigns to any $u \in L^2(I)^m$ the uniquely determined solution $x \in H^1(I)^n$ of the ODE-system

$$\begin{aligned} \dot{x}(t) - \mathbf{A}x(t) - \mathbf{B}u(t) &= 0 \quad \text{a.e. on } I \\ x(0) &= 0. \end{aligned} \tag{ODE}$$

It is well known that $\mathbb{S}$ is linear and continuous, see [2, Section 18]. For simplicity, the target-type objective functional $J\colon H^1(I)^n \times L^2(I)^m \to \mathbb{R}$ given by

$$J(x,u) := f(x(T)) + \tfrac{1}{2}\,\|x - \mathbf{x}_\mathrm{d}\|^2_{L^2(I)^n} + \tfrac{\sigma}{2}\,\|u - \mathbf{u}_\mathrm{d}\|^2_{L^2(I)^m}$$

for any $x \in H^1(I)^n$ and $u \in L^2(I)^m$ will be considered throughout this manuscript. However, it is possible to add some integral terms postulating additional assumptions. Note that terminal equality constraints can be easily added to the model as well.

In the recent paper [5], problem (OCTCC) was presented in a more general setting, where mixed control-state constraints were included as well. Some real-world applications from gas balancing on energy markets [15] or multi-agent control [4,7,16] motivate the study of (OCTCC). Clearly, model (OCTCC) belongs to the rich class of so-called *mathematical programs with complementarity constraints*, MPCCs for short.

The precise assumptions on (OCTCC) are stated below.

**Assumption 1.1** The functions $f, g_1, \ldots, g_l, G_1, \ldots, G_k, H_1, \ldots, H_k\colon \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable. Furthermore, $f$ is bounded from below. The matrices $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times m}$ are chosen such that

$$\begin{bmatrix}\mathbf{B} & \mathbf{AB} & \ldots & \mathbf{A}^{n-1}\mathbf{B}\end{bmatrix} \in \mathbb{R}^{n\times nm}$$

possesses full row rank $n$ (i.e. the ODE-system in the constraints of (OCTCC) is *controllable*, see [3]). In order to exclude trivial situations, it is assumed that (OCTCC) possesses at least one feasible point. Finally, the regularization parameter $\sigma > 0$, the desired state $\mathbf{x}_\mathrm{d} \in H^1(I)^n$, and the desired control $\mathbf{u}_\mathrm{d} \in L^2(I)^m$ are fixed.

Under the postulated assumptions, (OCTCC) possesses an optimal solution, see [5, Theorem 5.1]. In [5], the authors derive necessary optimality conditions and constraint

qualifications for (OCTCC) which were motivated by the rich theory on finite-dimensional complementarity programming, see [17,19]. However, the manuscript does not answer the question how problem (OCTCC) can be treated numerically. The aim of this study is to close this gap. Therefore, we follow standard ideas and relax the complementarity constraints as suggested by Scholtes [21]. However, there exist several other relaxation techniques for the computational treatment of complementarity constraints, see [13]. In principle, the findings of this study can be extended to all these relaxation methods doing some obvious changes. Particularly, this manuscript justifies the application of well-known numerical methods for the treatment of finite-dimensional complementarity problems to solve (OCTCC). Note that we do not focus on numerical details of the proposed method's implementation. Here, we abstain from a detailed numerical analysis associated with (OCTCC).

The remaining parts of the paper are structured as follows: Sect. 2 presents fundamental notation and recalls some preliminaries from [5]. Afterwards, we study Scholtes' relaxation scheme for (OCTCC) in Sect. 3. First, a global convergence result is presented. Second, we consider the situation where only Karush–Kuhn–Tucker (KKT for short) points of the surrogate problems are computed. We visualize the proposed approach in Sect. 4 by means of two numerical examples where the second one represents a particular model from multi-agent control with terminal friction constraints. Finally, the paper is closed by some concluding remarks in Sect. 5.

## 2 Preliminaries

In this work, we basically exploit standard notation which has been used in [5] already. However, let us briefly recall that $L^2(I)$ denotes the Banach space of all (equivalence classes of) scalar, measurable function on $I$ whose square is Lebesgue integrable which is equipped with the usual norm $\|\cdot\|_{L^2(I)}$. Furthermore, $H^1(I)$ is the common Sobolev space of all functions $x \in L^2(I)$ which possess a weak derivative $\dot{x}$ which belongs to $L^2(I)$ as well. It is equipped with the norm defined below:

$$\forall x \in H^1(I): \quad \|x\|_{H^1(I)} := \left(\|x\|_{L^2(I)} + \|\dot{x}\|_{L^2(I)}\right)^{1/2}.$$

By $L^2(I)^n$ and $H^1(I)^n$ we denote the $n$-fold Cartesian product of $L^2(I)$ and $H^1(I)$, respectively. Let the mapping $\mathtt{E}\colon H^1(I)^n \to L^2(I)^n$ be the associated natural embedding and denote by $\mathtt{E}_T\colon H^1(I)^n \to \mathbb{R}^n$ the evaluation operator $H^1(I)^n \ni x \mapsto x(T) \in \mathbb{R}^n$. For brevity, we will use the notation $x_T := x(T) = \mathtt{E}_T(x)$ for all $x \in H^1(I)^n$ throughout the remaining parts of the paper. By means of [5, Lemma 4.2], $\mathtt{E}$ and $\mathtt{E}_T$ are compact.

Recall that $\mathtt{S}\colon L^2(I)^m \to H^1(I)^n$ denotes the solution operator of (ODE). We set $\overline{\mathtt{S}} := \mathtt{E} \circ \mathtt{S}$ and $\mathtt{S}_T := \mathtt{E}_T \circ \mathtt{S}$. Note $\overline{\mathtt{S}}\colon L^2(I)^m \to L^2(I)^n$ and $\mathtt{S}_T\colon L^2(I)^m \to \mathbb{R}^n$. The controllability of the system (ODE) ensures that the operator $\mathtt{S}_T$ is surjective. Thus, we already know that the associated adjoint $\mathtt{S}_T^\star$ is injective. Below, the adjoint operators $\overline{\mathtt{S}}^\star$ and $\mathtt{S}_T^\star$ are characterized explicitly. These results are taken from [5, Lemmas 4.3, 4.4].

**Lemma 2.1** *1. For any $v \in L^2(I)^n$, we have $\overline{\mathrm{S}}^\star(v) = \mathbf{B}^\top p_1$ where $p_1 \in H^1(I)^n$ is the unique solution of the ODE-system*

$$\begin{aligned} \dot{p}_1(t) + \mathbf{A}^\top p_1(t) + v(t) &= 0 \quad \text{a.e. on } I \\ p_1(T) &= 0. \end{aligned}$$

*2. For any $b \in \mathbb{R}^n$, we have $\mathrm{S}_T^\star(b) = \mathbf{B}^\top p_2$ where $p_2 \in H^1(I)^n$ is the unique solution of the ODE-system*

$$\begin{aligned} \dot{p}_2(t) + \mathbf{A}^\top p_2(t) &= 0 \quad \text{a.e. on } I \\ p_2(T) &= b. \end{aligned}$$

The subsequent corollary follows easily from the above lemma.

**Corollary 2.1** *For $v \in L^2(I)^n$, $b \in \mathbb{R}^n$, and $w \in L^2(I)^m$, the function $p \in H^1(I)^n$ may solve the system*

$$\begin{aligned} \dot{p}(t) + \mathbf{A}^\top p(t) + v(t) &= 0 \quad \text{a.e. on } I \\ w(t) + \mathbf{B}^\top p(t) &= 0 \quad \text{a.e. on } I \\ p(T) &= b. \end{aligned}$$

*Then, $0 = \overline{\mathrm{S}}^\star(v) + \mathrm{S}_T^\star(b) + w$ holds true.*

## 3 Scholtes' relaxation technique

The terminal complementarity constraints appearing in our model (OCTCC) make a direct numerical treatment difficult due to two main issues: First, the feasible set of (OCTCC) is almost disconnected. Second, the appearance of the terminal complementarity constraint implies that standard regularity assumptions from nonlinear programming are generally violated at the feasible points of (OCTCC). In order to overcome these difficulties, for a sequence $\{\theta^r\}_{r\in\mathbb{N}} \subset \mathbb{R}^+$ of positive relaxation parameters converging to zero, we consider the relaxed problem

$$\begin{aligned} J(x,u) &\to \min_{x,u} \\ \mathrm{S}(u) - x &= 0 \\ g_i(x_T) &\le 0 \qquad i \in \mathcal{L} \\ G_j(x_T) &\ge 0 \qquad j \in \mathcal{K} \\ H_j(x_T) &\ge 0 \qquad j \in \mathcal{K} \\ G_j(x_T)H_j(x_T) &\le \theta^r \qquad j \in \mathcal{K}. \end{aligned} \tag{OCTCC($\theta^r$)}$$

This idea was introduced in [21] to solve standard complementarity problems in the finite-dimensional context. Qualitative results associated with this relaxation approach

are presented in [13] as well. Noting that the complementarity requirement enters problem (OCTCC) only in terms of the terminal conditions, it is reasonable to think that some of these results can be generalized to the optimal control setting we are considering here.

Mimicking the proof of [5, Theorem 5.1], we easily see that (OCTCC($\theta^r$)) possesses an optimal solution for any $r \in \mathbb{N}$. Clearly, the feasible set of (OCTCC($\theta^r$)) is a superset of the feasible set of (OCTCC). Consequently, if a local minimizer of (OCTCC($\theta^r$)) is already feasible to (OCTCC), then this point is a local minimizer of the latter.

Forthwith, we will present two types of convergence results. First, we investigate the situation where (OCTCC($\theta^r$)) can be solved globally for any $r \in \mathbb{N}$. This might be possible if the structure of the terminal constraints appearing in (OCTCC) is not too difficult. Here, we show the natural result that the global minimizers of (OCTCC($\theta^r$)) converge strongly (along a subsequence) to a global minimizer of (OCTCC). Second, we examine the case where only KKT points of the surrogate problems (OCTCC($\theta^r$)) can be computed. Due to the nonconvexity of (OCTCC($\theta^r$)), this assumption is much more natural than the first one. Clearly, one cannot hope that the computed sequence converges to a local or even global minimizer of (OCTCC). However, we prove that under reasonable assumptions a sequence of KKT points associated with (OCTCC($\theta^r$)) converges (along a subsequence) to a so-called Clarke-stationary point of (OCTCC), see Definition 3.2 below. This seems to be a natural extension of a similar convergence result for finite-dimensional MPCCs obtained in [13, Theorem 3.1].

We start with the promised global convergence result.

**Theorem 3.1** *Let $\{\theta^r\}_{r\in\mathbb{N}} \subset \mathbb{R}^+$ be a sequence of positive relaxation parameters converging to zero. For any $r \in \mathbb{N}$, let $(\bar{x}^r, \bar{u}^r) \in H^1(I)^n \times L^2(I)^m$ be a globally optimal solution of* (OCTCC($\theta^r$)). *Then, $\{(\bar{x}^r, \bar{u}^r)\}_{r\in\mathbb{N}}$ possesses a convergent subsequence (without relabeling) whose limit point $(\bar{x}, \bar{u}) \in H^1(I)^n \times L^2(I)^m$ is a globally optimal solution of* (OCTCC).

***Proof*** Fix an arbitrary feasible point $(x, u) \in H^1(I)^n \times L^2(I)^m$ of (OCTCC). Clearly, $(x, u)$ is feasible to (OCTCC($\theta^r$)) as well which yields

$$\begin{aligned} f(\bar{x}^r_T) + \tfrac{1}{2}\left\|\bar{x}^r - \mathbf{x}_\mathrm{d}\right\|^2_{L^2(I)^n} + \tfrac{\sigma}{2}\left\|\bar{u}^r - \mathbf{u}_\mathrm{d}\right\|^2_{L^2(I)^m} \\ \leq f(x_T) + \tfrac{1}{2}\left\|x - \mathbf{x}_\mathrm{d}\right\|^2_{L^2(I)^n} + \tfrac{\sigma}{2}\left\|u - \mathbf{u}_\mathrm{d}\right\|^2_{L^2(I)^m} \end{aligned}$$

for any $r \in \mathbb{N}$ by definition of $J$. Noting that $f$ is bounded from below while $\sigma > 0$ holds true, $\{\bar{u}^r\}_{r\in\mathbb{N}}$ must be bounded in $L^2(I)^m$. Thus, it possesses a weakly convergent subsequence (without relabeling) with weak limit $\bar{u} \in L^2(I)^m$. Set $\bar{x} := \mathrm{S}(\bar{u})$. Due to the continuity of $\mathrm{S}$, $\{\bar{x}^r\}_{r\in\mathbb{N}}$ converges weakly in $H^1(I)^n$ to $\bar{x}$. Since $\mathrm{E}$ is compact, we obtain $\bar{x}^r \to \bar{x}$ in $L^2(I)^n$. Noting that $\mathrm{E}_T$ is a compact operator as well, we have $\bar{x}^r_T \to \bar{x}_T$ in $\mathbb{R}^n$. The continuity of $g$, $G$, and $H$ as well as $\theta^r \to 0$ guarantee the feasibility of $(\bar{x}, \bar{u})$ to (OCTCC).

Next, we exploit the above convergences, the continuity of $f$, as well as the weak lower semicontinuity of norms in order to obtain

$$\begin{aligned} J(\bar{x},\bar{u}) &= f(\bar{x}_T) + \tfrac{1}{2}\|\bar{x} - \mathbf{x}_\mathrm{d}\|^2_{L^2(I)^n} + \tfrac{\sigma}{2}\|\bar{u} - \mathbf{u}_\mathrm{d}\|^2_{L^2(I)^m} \\ &\leq \lim_{r\to\infty} f(\bar{x}^r_T) + \lim_{r\to\infty} \tfrac{1}{2}\left\|\bar{x}^r - \mathbf{x}_\mathrm{d}\right\|^2_{L^2(I)^n} + \liminf_{r\to\infty} \tfrac{\sigma}{2}\left\|\bar{u}^r - \mathbf{u}_\mathrm{d}\right\|^2_{L^2(I)^m} \\ &= \liminf_{r\to\infty} \left( f(\bar{x}^r_T) + \tfrac{1}{2}\left\|\bar{x}^r - \mathbf{x}_\mathrm{d}\right\|^2_{L^2(I)^n} + \tfrac{\sigma}{2}\left\|\bar{u}^r - \mathbf{u}_\mathrm{d}\right\|^2_{L^2(I)^m} \right) \\ &\leq \limsup_{r\to\infty} \left( f(\bar{x}^r_T) + \tfrac{1}{2}\left\|\bar{x}^r - \mathbf{x}_\mathrm{d}\right\|^2_{L^2(I)^n} + \tfrac{\sigma}{2}\left\|\bar{u}^r - \mathbf{u}_\mathrm{d}\right\|^2_{L^2(I)^m} \right) \\ &\leq f(x_T) + \tfrac{1}{2}\|x - \mathbf{x}_\mathrm{d}\|^2_{L^2(I)^n} + \tfrac{\sigma}{2}\|u - \mathbf{u}_\mathrm{d}\|^2_{L^2(I)^m} = J(x,u) \end{aligned}$$

Thus, $(\bar{x},\bar{u})$ solves (OCTCC) globally. Choosing $x := \bar{x}$ as well as $u := \bar{u}$ and exploiting $f(\bar{x}^r_T) \to f(\bar{x}_T)$ as well as $\bar{x}^r \to \bar{x}$ in $L^2(I)^n$, we obtain $\|\bar{u}^r - \mathbf{u}_\mathrm{d}\|_{L^2(I)^m} \to \|\bar{u} - \mathbf{u}_\mathrm{d}\|_{L^2(I)^m}$. We combine this with the weak convergence $\bar{u}^r \rightharpoonup \bar{u}$ in $L^2(I)^m$ and the property of $L^2(I)^m$ to be a Hilbert space in order to get the strong convergence $\bar{u}^r \to \bar{u}$ in $L^2(I)^m$. Finally, the continuity of $\mathrm{S}$ already yields $\bar{x}^r \to \bar{x}$ in $H^1(I)^n$ which completes the proof. □

In practice, the computation of a globally optimal solution of the nonconvex relaxed surrogate problem (OCTCC($\theta^r$)) might be difficult. Instead, it is a nearby presumption that only KKT points of the surrogate problems can be identified. In the following definition, the KKT conditions of (OCTCC($\theta^r$)) are presented. The derivation of this system is omitted here since the necessary arguments mainly parallel those ones used in [5, Section 6].

**Definition 3.1** For fixed $r \in \mathbb{N}$, let $(\bar{x}^r, \bar{u}^r) \in H^1(I)^n \times L^2(I)^m$ be feasible to problem (OCTCC($\theta^r$)). Then, $(\bar{x}^r, \bar{u}^r)$ is a KKT point of (OCTCC($\theta^r$)) if and only if there are an adjoint state $p^r \in H^1(I)^n$ and multipliers $\lambda^r \in \mathbb{R}^l$ as well as $\alpha^r, \beta^r, \xi^r \in \mathbb{R}^k$ which solve the following system:

$$\begin{aligned} 0 &= \dot{p}^r(t) + \mathbf{A}^\top p^r(t) + \bar{x}^r(t) - \mathbf{x}_\mathrm{d}(t) \qquad \text{a.e. on } I, \\ p^r_T &= \nabla f(\bar{x}^r_T) + \sum_{i\in\mathcal{L}} \lambda^r_i \nabla g_i(\bar{x}^r_T) \\ &\quad - \sum_{j\in\mathcal{K}} \left[\alpha^r_j - \xi^r_j H_j(\bar{x}^r_T)\right] \nabla G_j(\bar{x}^r_T) \\ &\quad - \sum_{j\in\mathcal{K}} \left[\beta^r_j - \xi^r_j G_j(\bar{x}^r_T)\right] \nabla H_j(\bar{x}^r_T), \\ 0 &= \sigma(\bar{u}^r(t) - \mathbf{u}_\mathrm{d}(t)) + \mathbf{B}^\top p^r(t) \qquad \text{a.e. on } I, \\ \lambda^r &\geq 0, \quad \forall i \notin \mathcal{I}^g_r \colon \lambda^r_i = 0, \\ \alpha^r &\geq 0, \quad \forall j \notin \mathcal{I}^G_r \colon \alpha^r_j = 0, \\ \beta^r &\geq 0, \quad \forall j \notin \mathcal{I}^H_r \colon \beta^r_j = 0, \\ \xi^r &\geq 0, \quad \forall j \notin \mathcal{I}^{GH}_r \colon \xi^r_j = 0. \end{aligned}$$

Here, the index sets $\mathcal{I}_r^g$, $\mathcal{I}_r^G$, $\mathcal{I}_r^H$, and $\mathcal{I}_r^{GH}$ are defined as stated below:

$$\begin{aligned}\mathcal{I}_r^g &:= \{i \in \mathcal{L} \mid g_i(\bar{x}_T^r) = 0\},\\ \mathcal{I}_r^G &:= \{j \in \mathcal{K} \mid G_j(\bar{x}_T^r) = 0\},\\ \mathcal{I}_r^H &:= \{j \in \mathcal{K} \mid H_j(\bar{x}_T^r) = 0\},\\ \mathcal{I}_r^{GH} &:= \{j \in \mathcal{K} \mid G_j(\bar{x}_T^r)H_j(\bar{x}_T^r) - \theta^r = 0\}.\end{aligned}$$

As mentioned above, the second convergence result of this work shows that a sequence $\{(\bar{x}^r, \bar{u}^r)\}_{r\in\mathbb{N}} \subset H^1(I)^n \times L^2(I)^m$ of KKT points associated with (OCTCC($\theta^r$)) contains a convergent subsequence whose limit point is a so-called Clarke-stationary point of (OCTCC) provided $\{\bar{u}^r\}_{r\in\mathbb{N}}$ is bounded. This observation parallels classical results from [13,21] for standard complementarity problems. Below, we state an appropriate generalization of Clarke-stationarity for (OCTCC). Other reasonable stationarity notions for (OCTCC), namely weak, Mordukhovich- and strong stationarity, are introduced in [5, Section 6].

**Definition 3.2** Let $(\bar{x}, \bar{u}) \in H^1(I)^n \times L^2(I)^m$ be a feasible point of (OCTCC). Then, $(\bar{x}, \bar{u})$ is called Clarke-stationary, C-stationary for short, if and only if there exist an adjoint state $p \in H^1(I)^n$ as well as multipliers $\lambda \in \mathbb{R}^l$ and $\mu, \nu \in \mathbb{R}^k$ which satisfy the following conditions:

$$0 = \dot{p}(t) + \mathbf{A}^\top p(t) + \bar{x}(t) - \mathbf{x}_\mathrm{d}(t) \qquad \text{a.e. on } I, \tag{1a}$$

$$p_T = \nabla f(\bar{x}(T)) + \sum_{i\in\mathcal{L}} \lambda_i \nabla g_i(\bar{x}_T) - \sum_{j\in\mathcal{K}} \big[\mu_j \nabla G_j(\bar{x}_T) + \nu_j \nabla H_j(\bar{x}_T)\big], \tag{1b}$$

$$0 = \sigma(\bar{u}(t) - \mathbf{u}_\mathrm{d}(t)) + \mathbf{B}^\top p(t) \qquad \text{a.e. on } I, \tag{1c}$$

$$\lambda \geq 0, \quad \forall i \notin \mathcal{I}^g\colon\ \lambda_i = 0, \tag{1d}$$

$$\forall j \in \mathcal{I}^{+0}\colon\ \mu_j = 0, \tag{1e}$$

$$\forall j \in \mathcal{I}^{0+}\colon\ \nu_j = 0, \tag{1f}$$

$$\forall j \in \mathcal{I}^{00}\colon\ \mu_j \nu_j \geq 0. \tag{1g}$$

Therein, the index sets $\mathcal{I}^g$, $\mathcal{I}^{+0}$, $\mathcal{I}^{0+}$, and $\mathcal{I}^{00}$ are defined as stated below:

$$\begin{aligned}\mathcal{I}^g &:= \{i \in \mathcal{L} \mid g_i(\bar{x}_T) = 0\},\\ \mathcal{I}^{0+} &:= \{j \in \mathcal{K} \mid G_j(\bar{x}_T) = 0 \ \wedge\ H_j(\bar{x}_T) > 0\},\\ \mathcal{I}^{+0} &:= \{j \in \mathcal{K} \mid G_j(\bar{x}_T) > 0 \ \wedge\ H_j(\bar{x}_T) = 0\},\\ \mathcal{I}^{00} &:= \{j \in \mathcal{K} \mid G_j(\bar{x}_T) = 0 \ \wedge\ H_j(\bar{x}_T) = 0\}.\end{aligned}$$

Next, we postulate our standing assumptions for further theoretical investigations of the relaxation technique.

**Assumption 3.1** We fix a sequence $\{\theta^r\}_{r\in\mathbb{N}} \subset \mathbb{R}^+$ converging to zero. For any $r \in \mathbb{N}$, let $(\bar{x}^r, \bar{u}^r) \in H^1(I)^n \times L^2(I)^m$ be a KKT point of (OCTCC($\theta^r$)). Furthermore, let

$p^r \in H^1(I)^n$, $\lambda^r \in \mathbb{R}^l$, and $\alpha^r, \beta^r, \xi^r \in \mathbb{R}^k$ be the associated Lagrange multipliers, see Definition 3.1. We assume that $\{\bar{u}^r\}_{r\in\mathbb{N}}$ is bounded and, thus, possesses a weakly convergent subsequence with weak limit $\bar{u} \in L^2(I)^m$. For simplicity of notation, we do not relabel this subsequence but use the expression $\{\bar{u}^r\}_{r\in\mathbb{N}}$ again. Finally, $\bar{x} := \mathrm{S}(\bar{u})$ is the state function associated with $\bar{u}$.

Let us briefly comment on the assumption that $\{\bar{u}^r\}_{r\in\mathbb{N}}$ is bounded.

***Remark 3.1*** The boundedness of $\{\bar{u}^r\}_{r\in\mathbb{N}} \subset L^2(I)^m$ is not very restrictive and can be guaranteed under a mild assumption: If the sequence $\{J(\bar{x}^r, \bar{u}^r)\}_{r\in\mathbb{N}}$ of objective values associated with the KKT points $(\bar{x}^r, \bar{u}^r)$ of (OCTCC($\theta^r$)) is bounded in $\mathbb{R}$, then $\{\bar{u}^r\}_{r\in\mathbb{N}}$ needs to be bounded in $L^2(I)^m$ since the function $f$ is bounded from below. Note that this observation has been exploited in the proof of Theorem 3.1 as well.

For the proof of our next convergence result, we need to study the behavior of the index sets $\mathcal{I}_r^g$, $\mathcal{I}_r^G$, and $\mathcal{I}_r^H$ as $r \to \infty$. Some corresponding observations which directly follow from Assumption 3.1 are stated in the lemma below.

**Lemma 3.1** *For sufficiently large $r \in \mathbb{N}$, the following relations hold true:*

$$\begin{aligned} \mathcal{I}_r^g &\subset \mathcal{I}^g, \\ \mathcal{I}_r^G &\subset \mathcal{I}^{0+} \cup \mathcal{I}^{00}, \\ \mathcal{I}_r^H &\subset \mathcal{I}^{+0} \cup \mathcal{I}^{00}. \end{aligned}$$

***Proof*** Due to $\bar{u}^r \rightharpoonup \bar{u}$ in $L^2(I)^m$ and the continuity of the solution operator $\mathrm{S}$ associated with (ODE), we obtain $\bar{x}^r \rightharpoonup \bar{x}$ in $H^1(I)^n$. Recalling that $\mathrm{E}_T$ is compact, $\bar{x}_T^r \to \bar{x}_T$ holds true in $\mathbb{R}^n$. Thus, the validity of the presented inclusions is guaranteed by continuity of $g$, $G$, and $H$. □

Now, we are in position to state our second convergence result. For its validation, we exploit ideas used in the proof of [13, Theorem 3.1]. However, some essentials of infinite-dimensional programming have to be taken into account as well in order to show the following theorem.

**Theorem 3.2** *Suppose that the following constraint qualification is valid:*

$$\left.\begin{aligned} &0 = \sum_{i\in\mathcal{L}} \lambda_i \nabla g_i(\bar{x}_T) - \sum_{j\in\mathcal{K}} \big[\mu_j \nabla G_j(\bar{x}_T) + \nu_j \nabla H_j(\bar{x}_T)\big], \\ &\lambda \geq 0, \quad \forall i \notin \mathcal{I}^g\colon\ \lambda_i = 0, \\ &\forall j \in \mathcal{I}^{+0}\colon\ \mu_j = 0, \\ &\forall j \in \mathcal{I}^{0+}\colon\ \nu_j = 0 \end{aligned}\right\} \implies \left\{\begin{aligned} &\lambda = 0, \\ &\mu = 0, \\ &\nu = 0. \end{aligned}\right. \quad \text{(CQ)}$$

*Then, $(\bar{x}, \bar{u})$ is a C-stationary point of* (OCTCC). *Furthermore, the following convergences hold along a subsequence:*

$$\bar{x}^r \to \bar{x} \ \text{ in } H^1(I)^n, \qquad \bar{u}^r \to \bar{u} \ \text{ in } L^2(I)^m, \qquad p^r \to p \ \text{ in } H^1(I)^n.$$

*Here, $p \in H^1(I)^n$ is the adjoint state which appears in the system (1).*

*If $\mathbf{u}_\mathrm{d} \in H^1(I)^m$ is valid, then $\{\bar{u}^r\}_{r\in\mathbb{N}} \subset H^1(I)^m$ and $\bar{u} \in H^1(I)^m$ hold. Furthermore, we obtain $\bar{u}^r \to \bar{u}$ in $H^1(I)^m$.*

***Proof*** From $\bar{u}^r \rightharpoonup \bar{u}$ in $L^2(I)^m$, we obtain $\bar{x}^r \rightharpoonup \bar{x}$ in $H^1(I)^n$ and $\bar{x}_T^r \to \bar{x}_T$ in $\mathbb{R}^n$ since $\mathtt{S}$ is continuous while $\mathtt{S}_T$ is compact and continuous.

We note that for any $r \in \mathbb{N}$ and any $j \in \mathcal{K}$, the multipliers $\alpha_j^r$ and $\xi_j^r$ ($\beta_j^r$ and $\xi_j^r$, respectively) cannot be positive at the same time since $\mathcal{I}_r^G \cap \mathcal{I}_r^{GH} = \varnothing$ ($\mathcal{I}_r^H \cap \mathcal{I}_r^{GH} = \varnothing$) is valid by definition. For any $r \in \mathbb{N}$, let us introduce $\mu^r, \nu^r \in \mathbb{R}^k$ as stated below for any $j \in \mathcal{K}$:

$$\mu_j^r := \begin{cases} \alpha_j^r & \text{if } \alpha_j^r > 0, \\ -\xi_j^r H_j(\bar{x}_T^r) & \text{if } \xi_j^r > 0 \ \wedge \ j \notin \mathcal{I}^{+0}, \\ 0 & \text{otherwise,} \end{cases}$$

$$\nu_j^r := \begin{cases} \beta_j^r & \text{if } \beta_j^r > 0, \\ -\xi_j^r G_j(\bar{x}_T^r) & \text{if } \xi_j^r > 0 \ \wedge \ j \notin \mathcal{I}^{0+}, \\ 0 & \text{otherwise.} \end{cases}$$

Then, we obtain

$$\begin{aligned} p_T^r = \nabla f(\bar{x}_T^r) + \sum_{i\in\mathcal{L}} \lambda_i^r \nabla g_i(\bar{x}_T^r) - \sum_{j\in\mathcal{K}} \big[\mu_j^r \nabla G_j(\bar{x}_T^r) + \nu_j^r \nabla H_j(\bar{x}_T^r)\big] \\ + \sum_{j\in\mathcal{I}^{+0}} \xi_j^r H_j(\bar{x}_T^r) \nabla G_j(\bar{x}_T^r) + \sum_{j\in\mathcal{I}^{0+}} \xi_j^r G_j(\bar{x}_T^r) \nabla H_j(\bar{x}_T^r) \end{aligned}$$

from Definition 3.1. Note that the appearing index sets $\mathcal{I}^{+0}$ and $\mathcal{I}^{0+}$ correspond to the limit point $\bar{x}$ and not to the KKT points of (OCTCC($\theta^r$)). Now, we apply Corollary 2.1 and Definition 3.1 to deduce

$$\begin{aligned} 0 = \overline{\mathrm{S}}^\star(\mathrm{E}(\bar{x}^r) - \mathbf{x}_\mathrm{d}) + \sigma(\bar{u}^r - \mathbf{u}_\mathrm{d}) \\ + \mathrm{S}_T^\star \Bigg[ \nabla f(\bar{x}_T^r) + \sum_{i\in\mathcal{L}} \lambda_i^r \nabla g_i(\bar{x}_T^r) - \sum_{j\in\mathcal{K}} \big[\mu_j^r \nabla G_j(\bar{x}_T^r) + \nu_j^r \nabla H_j(\bar{x}_T^r)\big] \\ + \sum_{j\in\mathcal{I}^{+0}} \xi_j^r H_j(\bar{x}_T^r) \nabla G_j(\bar{x}_T^r) + \sum_{j\in\mathcal{I}^{0+}} \xi_j^r G_j(\bar{x}_T^r) \nabla H_j(\bar{x}_T^r) \Bigg]. \end{aligned} \tag{2}$$

Suppose that the sequence $\{(\lambda^r, \mu^r, \nu^r, \xi^r_{\mathcal{I}^{+0}\cup\mathcal{I}^{0+}})\}_{r\in\mathbb{N}}$ is not bounded, define the constants $\kappa^r := \left|(\lambda^r, \mu^r, \nu^r, \xi^r_{\mathcal{I}^{+0}\cup\mathcal{I}^{0+}})\right|_2$ for all $r \in \mathbb{N}$, and set

$$(\tilde{\lambda}^r, \tilde{\mu}^r, \tilde{\nu}^r, \tilde{\xi}^r_{\mathcal{I}^{+0}\cup\mathcal{I}^{0+}}) := \frac{1}{\kappa^r}(\lambda^r, \mu^r, \nu^r, \xi^r_{\mathcal{I}^{+0}\cup\mathcal{I}^{0+}})$$

for all $r \in \mathbb{N}$. Then, $\{(\tilde{\lambda}^r, \tilde{\mu}^r, \tilde{\nu}^r, \tilde{\xi}^r_{\mathcal{I}^{+0} \cup \mathcal{I}^{0+}})\}_{r \in \mathbb{N}}$ is a bounded sequence and converges w.l.o.g. to a nonvanishing multiplier $(\tilde{\lambda}, \tilde{\mu}, \tilde{\nu}, \tilde{\xi}_{\mathcal{I}^{+0} \cup \mathcal{I}^{0+}})$. Hence, dividing (2) by the positive number $\kappa^r$, taking the limit $r \to \infty$, and observing that $\{\bar{x}^r\}_{r \in \mathbb{N}}$ and $\{\bar{u}^r\}_{r \in \mathbb{N}}$ are bounded in $L^2(I)^n$ and $L^2(I)^m$, respectively, yields

$$0 = \mathbb{S}_T^\star \left[ \sum_{i \in \mathcal{L}} \tilde{\lambda}_i \nabla g_i(\bar{x}_T) - \sum_{j \in \mathcal{K}} \left[ \tilde{\mu}_j \nabla G_j(\bar{x}_T) + \tilde{\nu}_j \nabla H_j(\bar{x}_T) \right] \right]$$

as well as $\tilde{\lambda} \geq 0$, $\tilde{\lambda}_i = 0$ $(i \notin \mathcal{I}^g)$, $\tilde{\mu}_j = 0$ $(j \in \mathcal{I}^{+0})$, and $\tilde{\nu}_j = 0$ $(j \in \mathcal{I}^{0+})$, see Lemma 3.1. We note that the validity of the constraint qualification (CQ) implies $\tilde{\lambda} = 0, \tilde{\mu} = 0$, and $\tilde{\nu} = 0$ since $\mathbb{S}_T^\star$ is injective. Thus, due to the above observation, $\tilde{\xi}_{\mathcal{I}^{+0}}$ or $\tilde{\xi}_{\mathcal{I}^{0+}}$ possesses a nonvanishing component. Assume w.l.o.g. that there is $j_0 \in \mathcal{I}^{+0}$ such that $\tilde{\xi}_{j_0}$ does not vanish. Then, $\xi^r_{j_0} > \varepsilon \kappa^r$ holds true for sufficiently large $r \in \mathbb{N}$ and some $\varepsilon > 0$. By construction, $\nu^r_{j_0} < -\varepsilon \kappa^r G_{j_0}(\bar{x}^r_T)$ is valid for sufficiently large $r \in \mathbb{N}$, i.e. taking the limit $r \to \infty$ yields

$$\tilde{\nu}_{j_0} = \lim_{r \to \infty} \frac{\nu^r_{j_0}}{\kappa^r} \leq -\varepsilon \lim_{r \to \infty} G_{j_0}(\bar{x}^r_T) = -\varepsilon G_{j_0}(\bar{x}_T) < 0$$

due to $j_0 \in \mathcal{I}^{+0}$. This, however, contradicts $(\tilde{\lambda}, \tilde{\mu}, \tilde{\nu}) = (0, 0, 0)$.

Thus, $\{(\lambda^r, \mu^r, \nu^r, \xi^r_{\mathcal{I}^{+0} \cup \mathcal{I}^{0+}})\}_{r \in \mathbb{N}}$ is bounded and converges (along a subsequence without relabeling) to a multiplier $(\lambda, \mu, \nu, \xi_{\mathcal{I}^{+0} \cup \mathcal{I}^{0+}})$. Due to $\bar{x}^r \to \bar{x}$ in $L^2(I)^n$, $\bar{x}^r_T \to \bar{x}_T$ in $\mathbb{R}^n$, and the continuity of $f$, $g$, $G$, and $H$, we infer $\bar{u}^r \to \bar{u}$ in $L^2(I)^m$, (1d), (1e), (1f), as well as

$$\begin{aligned} 0 = {} & \overline{\mathbb{S}}^\star(\mathbb{E}(\bar{x}) - \mathbf{x}_\mathrm{d}) + \sigma(\bar{u} - \mathbf{u}_\mathrm{d}) \\ & + \mathbb{S}_T^\star \left[ \nabla f(\bar{x}_T) + \sum_{i \in \mathcal{L}} \lambda_i \nabla g_i(\bar{x}_T) - \sum_{j \in \mathcal{K}} \left[ \mu_j \nabla G_j(\bar{x}_T) + \nu_j \nabla H_j(\bar{x}_T) \right] \right] \end{aligned}$$

from (2), see Lemma 3.1. Now, we apply Lemma 2.1 in order to obtain the conditions (1a)–(1c).

Fix $j \in \mathcal{I}^{00}$ and suppose that $\mu_j \nu_j < 0$ holds true. If $\mu_j < 0$ and $\nu_j > 0$ are valid, then $\mu^r_j < 0$ and $\nu^r_j > 0$ must be satisfied for sufficiently large $r \in \mathbb{N}$. On the other hand, $\mu^r_j < 0$ implies $\xi^r_j > 0$ which contradicts $\beta^r_j = \nu^r_j > 0$. Similarly, we obtain a contradiction from $\mu_j > 0$ and $\nu_j < 0$. Therefore, condition (1g) holds as well, i.e. $(\bar{x}, \bar{u})$ is a C-stationary point of (OCTCC).

Noting that we have the convergences $\bar{x}^r \to \bar{x}$ in $L^2(I)^n$, $\bar{x}^r_T \to \bar{x}_T$ in $\mathbb{R}^n$, and $(\lambda^r, \mu^r, \nu^r, \xi^r_{\mathcal{I}^{+0} \cup \mathcal{I}^{0+}}) \to (\lambda, \mu, \nu, \xi_{\mathcal{I}^{+0} \cup \mathcal{I}^{0+}})$ while the solution operator of the ODE-system

$$\begin{aligned} \dot{p}(t) + \mathbf{A}^\top p(t) + v(t) &= 0 \qquad \text{a.e. on } I \\ p_T &= b \end{aligned}$$

is continuous as a mapping $L^2(I)^n \times \mathbb{R}^n \ni (v, b) \mapsto p \in H^1(I)^n$, see e.g. [2, Chapter 18], we obtain $p^r \to p$ in $H^1(I)^n$ from Definition 3.1 and (1a), (1b). Combining Definition 3.1 and (1c), we have

$$\bar{u}^r = \mathbf{u}_\mathrm{d} - \tfrac{1}{\sigma}\mathbf{B}^\top p^r \to \mathbf{u}_\mathrm{d} - \tfrac{1}{\sigma}\mathbf{B}^\top p = \bar{u}$$

in $L^2(I)^m$. Thus, if $\mathbf{u}_\mathrm{d}$ is a function from $H^1(I)^m$, then the same holds true for $\bar{u}^r$, $r \in \mathbb{N}$, and $\bar{u}$ and the above convergence can be extended to $H^1(I)^m$. This completes the proof. □

Let us present some brief remarks regarding the regularity condition (CQ).

***Remark 3.2*** Assume that the computed limit point $(\bar{x}, \bar{u})$ satisfies the constraint qualification (CQ). If $(\bar{x}, \bar{u})$ is a local minimizer of (OCTCC), then it is already a Mordukhovich-stationary point, i.e. it satisfies the C-stationarity conditions (1) where (1g) is strengthened to

$$\forall j \in \mathcal{I}^{00}\colon\ \mu_j \nu_j = 0 \ \vee\ (\mu_j > 0 \ \wedge\ \nu_j > 0),$$

see [5, Theorem 7.5]. Staying close to the notion of finite-dimensional complementarity programming, the constraint qualification (CQ) might be referred to as MPCC-MFCQ, see [13, Definition 2.4].

In order to ensure that locally or even globally optimal solutions of the surrogate problem (OCTCC($\theta^r$)) satisfy the KKT conditions stated in Definition 3.1, the validity of a constraint qualification is necessary. Here, we rely on Robinson's constraint qualification which is the fundamental regularity condition in the context of Banach space programming. It has been introduced by Robinson [20] in order to study the stability properties of parameterized nonlinear systems in Banach spaces. Later, Kurcyusz and Zowe exploited this condition in order to derive necessary optimality conditions of KKT-type in Banach space programming, see [24]. Further information on Robinson's constraint qualification can be found in the monograph [6].

Here, we want to emphasize that the validity of the constraint qualification (CQ) at the limit point $(\bar{x}, \bar{u})$ implies that Robinson's constraint qualification holds at the iterates $(\bar{x}^r, \bar{u}^r)$ w.r.t. the surrogate problem (OCTCC($\theta^r$)) for sufficiently large $r$. Consequently, the assumption that $(\bar{x}^r, \bar{u}^r)$ is a KKT point of (OCTCC($\theta^r$)) would not be restrictive anymore provided $(\bar{x}^r, \bar{u}^r)$ is a locally optimal solution of the relaxed surrogate.

Note that the upcoming result can be seen as a natural extension of [13, Theorem 3.2] which shows that the validity of MPCC-MFCQ at the limit point produced by Scholtes' relaxation scheme (applied to standard MPCCs) implies that MFCQ holds in a neighborhood of this point for all the relaxed surrogate problems where the relaxation parameter is sufficiently small. The proof of this result is omitted since it follows easily reprising the arguments in [13] while recalling that the linear operator $(\bar{\mathrm{S}}, \mathrm{S}_T)\colon L^2(I)^m \to L^2(I)^n \times \mathbb{R}^n$ is surjective, see [5, Lemmas 7.1 and 7.4].

**Lemma 3.2** *Let the constraint qualification* (CQ) *be valid. Then, Robinson's constraint qualification is valid at the iterates* $(\bar{x}^r, \bar{u}^r)$ *for sufficiently large* $r \in \mathbb{N}$. *Particularly, if* $(\bar{x}^r, \bar{u}^r)$ *is locally optimal for* (OCTCC($\theta^r$)), *then it satisfies the KKT conditions from Definition* 3.1.

In our above considerations, we only commented on Scholtes' relaxation scheme. Investigating the presented proofs which, from the infinite-dimensional point of view, only exploit the controllability of (ODE) and the present function space setting but not the precise geometry of the relaxed feasible set, similar results are likely to be satisfied for other relaxation schemes, see [13].

## 4 Numerical experiments

### 4.1 Discretization strategy

Throughout the section, we assume $\mathbf{u}_\mathrm{d} \in H^1(I)^m$ which is non-restrictive in the setting of the aforementioned underlying real-world applications where $\mathbf{u}_\mathrm{d}$ generally vanishes, see [4,7,15,16]. Invoking Theorem 3.2, we now can restrict our consideration to the situation where the control function in (OCTCC) is chosen from $H^1(I)^m$. Noting that $H^1(I)$ is continuously embedded into $C(\bar{I})$, see [1, Theorem 6.3], the pointwise evaluation of state *and* control functions is reasonable in this setting.

For the computational treatment of (OCTCC) via a sequence of surrogate problems of the form (OCTCC($\theta^r$)), we rely on a *first-discretize-then-optimize*-approach. Therefore, we decompose $I$ into $N \in \mathbb{N}$ equidistant intervals of length $h := T/N$. The discretized variables are given by $x_s := x(hs)$ and $u_s := u(hs)$ for $s = 0, \ldots, N$. Note that the data functions $\mathbf{x}_\mathrm{d}$ and $\mathbf{u}_\mathrm{d}$ are discretized similarly. Now, we need to choose an appropriate strategy to represent the discretized linear system associated with (ODE). Therefore, we exploit the *trapezoidal rule* which is a certain *Runge–Kutta method*, see [8]. Particularly, we set

$$x_{s+1} = x_s + \tfrac{h}{2}\Big[\mathbf{A}\big[x_s + x_{s+1}\big] + \mathbf{B}\big[u_s + u_{s+1}\big]\Big] \qquad s = 0, \ldots, N-1.$$

For any function $w \in H^1(I)$, we have

$$\int_0^T w(t)\mathrm{d}t \approx \sum_{s=0}^{N-1} \frac{w(sh) + w((s+1)h)}{2N} = \frac{1}{N}\left(\tfrac{1}{2}w(0) + \sum_{s=1}^{N-1} w(sh) + \tfrac{1}{2}w(T)\right).$$

This observation provides the following discretization scheme for the norms in the objective functional of (OCTCC):

$$\begin{aligned}
&\tfrac{1}{2}\,\|x-\mathbf{x}_{\mathrm{d}}\|^2_{L^2(I)^n}\\
&\quad=\frac{1}{2}\sum_{i=1}^{n}\int_0^T (x_i(t)-\mathbf{x}_{\mathrm{d},i}(t))^2\mathrm{d}t\\
&\quad\approx\frac{1}{2N}\sum_{i=1}^{n}\left(\tfrac{1}{2}(x_{0,i}-\mathbf{x}_{\mathrm{d},0,i})^2+\sum_{s=1}^{N-1}(x_{s,i}-\mathbf{x}_{\mathrm{d},s,i})^2+\tfrac{1}{2}(x_{T,i}-\mathbf{x}_{\mathrm{d},T,i})^2\right).
\end{aligned}$$

Similarly, the term $\frac{\sigma}{2}\,\|u-\mathbf{u}_{\mathrm{d}}\|^2_{L^2(I)^m}$ is discretized.

Note that whenever $\mathbf{u}_{\mathrm{d}}\notin H^1(I)^m$ holds true, we cannot rely on the pointwise evaluation of the control which only possesses $L^2$-regularity in general. Thus, the above discretization strategy has to be changed slightly. One possible way in order to proceed might be the approximation of the control by piecewise constant functions, see e.g. [10, Section 5].

The resulting finite-dimensional programs are (due to the relaxation of the terminal complementarity constraint) standard nonlinear problems which can be solved using e.g. the interior-point-solver `IPOPT`, see [23].

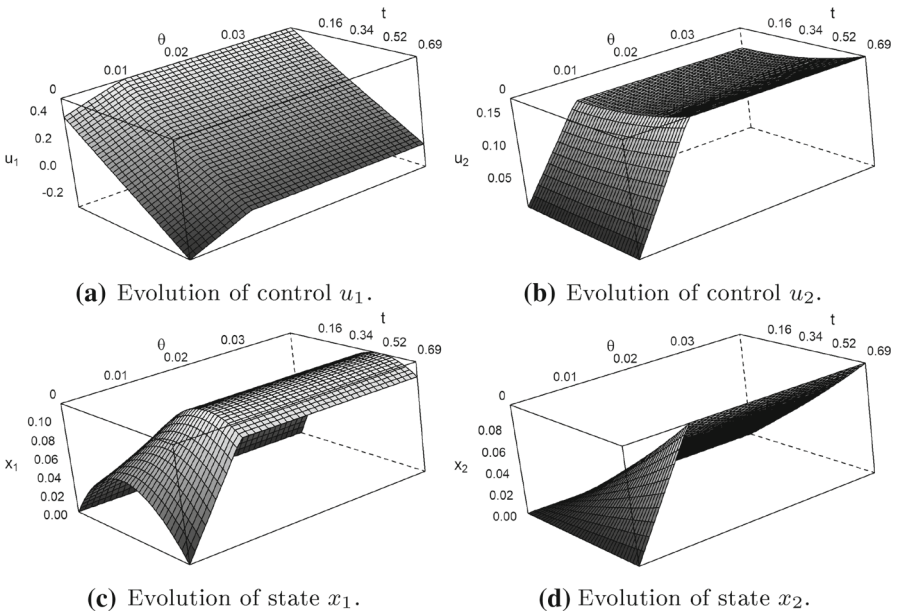
In this study, we focus on the theoretical details of the suggested relaxation method which is why the derivation of error estimates for the suggested Runge–Kutta method is clearly beyond of the paper's scope. However, let us briefly mention that related considerations regarding the numerical analysis of optimal control problems of ordinary differential equations can be found in [8–10,22]. Another idea for the derivation of error estimates for a different discretization strategy would be to couple the well-known method of *variational discretization*, where a discrete counterpart of the solution operator to (ODE) depending on $h$ and its convergence in a suitable operator topology are considered, see [11,12], with already available error estimates for the numerical handling of linear ordinary differential equations via higher order Runge–Kutta schemes, see [8]. Noting that the terminal constraints in (ODE) and (OCTCC($\theta^r$)) are not convex, the foreshadowed considerations may turn out to be technically challenging.

### 4.2 Example 1

Here, we first review [5, Example 7.9] in terms of the presented numerical method. For $n=m:=2$ and $T:=\ln 2$, we consider the problem

$$\begin{aligned}
\tfrac{1}{2}\big(\|x_1-1\|^2_{L^2(I)}+\|x_2\|^2_{L^2(I)}\big)+\tfrac{1}{2}\big(\|u_1\|^2_{L^2(I)}+\|u_2\|^2_{L^2(I)}\big)\;&\to\;\min_{x,u}\\
\dot{x}_1(t)-u_1(t)\;=\;\dot{x}_2(t)-u_2(t)\;&=\;0 \qquad \text{a.e. on } I\\
x_1(0)\;=\;x_2(0)\;&=\;0\\
x_1(T)-x_2(T)\;&\le\;0\\
0\;\le\;x_1(T)\;\perp\;x_2(T)\;&\ge\;0.
\end{aligned}\tag{3}$$

From [5] we know that the unique global solution of (3) is given by




**(a)** Evolution of control $u_1$. **(b)** Evolution of control $u_2$.

**(c)** Evolution of state $x_1$. **(d)** Evolution of state $x_2$.

**Fig. 1** Surface plots of the solutions to the discretized, relaxed surrogate problems

$$\bar{x}_1(t) = \tfrac{1}{3}\sinh(t) - \cosh(t) + 1, \qquad \bar{u}_1(t) = \tfrac{1}{3}\cosh(t) - \sinh(t), \qquad \bar{x}_2(t) = \bar{u}_2(t) = 0$$

for all $t \in I$. First, we analytically calculate the optimal solution of the associated relaxed problem in the given function space setting. For this purpose, we introduce

$$x_1^c(t) := c\sinh t - \cosh h + 1 \qquad x_2^d(t) := d\sinh t$$
$$u_1^c(t) := c\cosh t - \sinh t \qquad u_2^d(t) := d\cosh t$$

for all $t \in I$ and constants $c, d \in \mathbb{R}$. Using standard methods, see [14], one can easily check that the relaxation of problem (3) possesses two KKT points which depend on $\theta > 0$. For $\theta \geq 1/25$, we set $c_1 := 7/15$ and $d_1 := 2/15$ in order to obtain the KKT point $P = (\bar{x}^1, \bar{u}^1) := ((x_1^{c_1}, x_2^{d_1}), (u_1^{c_1}, u_2^{d_1}))$. Furthermore, for any $\theta \leq \frac{1}{3}\sqrt{5}$, $c_2(\theta) := \frac{1}{3}(2\sqrt{\theta}+1)$, and $d_2(\theta) := \frac{2}{3}\sqrt{\theta}$, another KKT point $Q(\theta) = (\bar{x}^2(\theta), \bar{u}^2(\theta))$ is given by $((x_1^{c_2(\theta)}, x_2^{d_2(\theta)}), (u_1^{c_2(\theta)}, u_2^{d_2(\theta)}))$. One can easily check that both KKT points coincide for $\theta = 1/25$. The objective value of $P$ equals $19/30$, while the objective value of $Q(\theta)$ is given by $\frac{5}{6}\theta - \frac{1}{3}\sqrt{\theta} + \frac{2}{3}$. Comparing both, the global minimizer of the $\theta$-relaxation is given by $Q(\theta)$ for any $\theta \in (0, 1/25)$ and $P$ for $\theta \in [1/25, \infty)$.

Figure 1 shows the evolution of controls and states for $\theta \in [0, 0.04]$ and $N = 500$. Note that `IPOPT` computes the global minimizer of the relaxed problems for sufficiently small values of $\theta$.

### 4.3 Example 2

The following example reflects the situation of multi-agent-control with terminal friction conditions. Let us consider $n_a \in \mathbb{N}$ agents whose position at time $t$ is denoted by $x_i(t), i = 1, \ldots, n_a$. All agents need to start their movements at time $t = 0$ at the point 0. It is desirable that agent $i$ moves as close to the given path $\mathbf{x}_{\mathrm{d},i}$ as possible for all $i = 1, \ldots, n_a$. By $x_{i+n_a} := \dot{x}_i$, we denote the speed of agent $i, i = 1, \ldots, n_a$, which can be controlled by $u_i$. At terminal time $T$, the agents need to satisfy given frictional constraints in order to allow the sharing of goods while their respective speed must be zero. Thus, we have $n = 2n_a$ and $m = n_a$ in this situation and it can be easily checked that the associated ODE-system satisfies the proposed controllability assumption. An overall formulation of the problem is given by

$$
\begin{aligned}
&\tfrac{1}{2}\sum_{i=1}^{n_a}\left(\left\|x_i - \mathbf{x}_{\mathrm{d},i}\right\|^2_{L^2(I)} + \left\|x_{i+n_a}\right\|^2_{L^2(I)}\right) \\
&\qquad + \tfrac{\sigma}{2}\sum_{i=1}^{n_a}\|u_i\|^2_{L^2(I)} \;\to\; \min_{x,u} \\
&\qquad \dot{x}_i(t) - x_{i+n_a}(t) = 0 \qquad \text{a.e. on } I,\ i = 1, \ldots, n_a \\
&\qquad \dot{x}_{i+n_a}(t) - u_i(t) = 0 \qquad \text{a.e. on } I,\ i = 1, \ldots, n_a \\
&\qquad x_i(0) = 0 \qquad i = 1, \ldots, 2n_a \\
&\qquad x_{i+n_a}(T) = 0 \qquad i = 1, \ldots, n_a \\
&\qquad 0 \le G_j(x_T) \perp H_j(x_T) \ge 0 \qquad j \in \mathcal{K}.
\end{aligned}
\tag{4}
$$

For our numerical calculation, we choose $T := 1$, $n_a := 2$, $\sigma := 10^{-2}$, and $k := 1$ with

$$G_1(x_T) := 1 - x_1(T) \qquad H_1(x_T) := 1 - x_2(T)$$

for all $x \in H^1(I)^4$. Furthermore, let us set $\mathbf{x}_{\mathrm{d},1}(t) := t - t^2$ and $\mathbf{x}_{\mathrm{d},2}(t) := t^2 - t^3$ for all $t \in I$. Note that we have $\mathbf{x}_{\mathrm{d},1}(1) = \mathbf{x}_{\mathrm{d},2}(1) = 0$ for this choice of the desired states which clearly antagonizes the complementarity requirement on $x_1$ and $x_2$.

It is important to highlight that `IPOPT` allows to compute the numerical solution of the relaxed problem associated with (4) without any relaxation. However, the algorithm solves a sequence of barrier problems that may lead to infeasible solutions of the original problem if the feasible set is too small. Furthermore, it turns out that regarding different starting points, the relaxation approach is much more robust than a direct treatment of (4) when it comes down to identifying global minimizers. Thus, it is reasonable to rely on the relaxation approach.

To start, we note that due to $n_a = 2$, the discretized problem associated with (4) can be decomposed into two convex programs by replacing the terminal complementarity constraint by $x_{T,1} = 1$ and $x_{T,2} \le 1$ in the first or $x_{T,1} \le 1$ and $x_{T,2} = 1$ in the second case. Thus, by comparing the solution of both programs one can find the global minimizer and its corresponding objective value $\mathrm{Obj}(N)^*$, where $N$ denotes the total

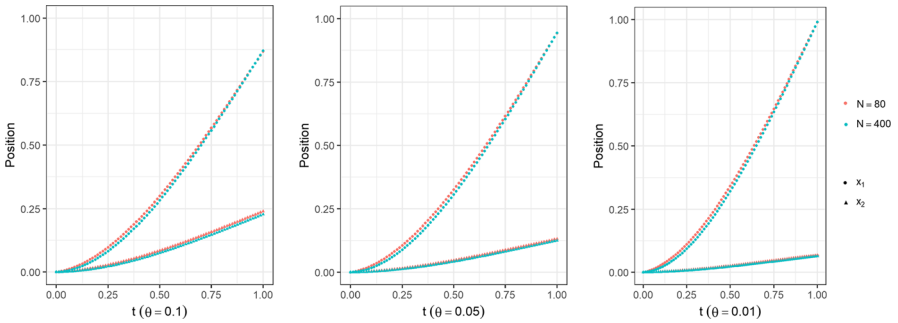



**Fig. 2** Position of the agents $x_1$ (dots) and $x_2$ (triangles) for $(\theta, N) \in \{0.1, 0.05, 0.01\} \times \{80, 400\}$

number of (equidistant) sampling intervals. We emphasize that the global minimizer, which clearly depends on $N$, satisfies $x_{T,1} = 1$ and $x_{T,2} < 1$ for sufficiently large $N$. The strategy helps to determine the relative gap $\mathcal{G}(\theta, N)$ between the objective value of the relaxed problem $\mathrm{Obj}(\theta, N)$ at the solution determined by `IPOPT` and $\mathrm{Obj}(N)^*$ which computes as stated below:

$$\mathcal{G}(\theta, N) := \left| \frac{\mathrm{Obj}(\theta, N) - \mathrm{Obj}(N)^*}{\mathrm{Obj}(N)^*} \right|.$$

We note that $\mathrm{Obj}(\theta, N)$ does not necessary equal the optimal objective value of the associated $\theta$-relaxation with $N$ sampling intervals since the latter is nonconvex and it cannot be guaranteed that `IPOPT` finds its global minimizer.

We solve the discretized surrogate problem for all

$$(\theta, N) \in \{0.01, 0.05, 0.1\} \times \{80, 160, 240, 320, 400\}$$

in order to test the performance of the relaxations. Here, we follow the approach of [13] by starting with large values of $\theta$, i.e., $\theta = 0.1$ and decreasing its value while using the previously found solution as a starting vector. This approach allows to improve the quality of the solution as $\theta$ is reduced. In order to start the algorithm, we use the global minimizer of the program (4) where the complementarity constraint is weakened to the nonnegativity requirements $G_j(x_T), H_j(x_T) \geq 0$ for all $j \in \mathcal{K}$. We note that the resulting program is convex and its global minimizer is easily computed using `IPOPT`.

To illustrate our findings, Fig. 2 displays the position $x_1$ and $x_2$ of the agents for different combinations of $\theta$ and $N$. One easily sees that feasibility for the complementarity problem is achieved as $\theta$ tends to zero.

We note that `IPOPT` solves the $\theta$-relaxations globally in this example and by means of Theorem 3.1, the associated sequence of solutions tends to the global minimizer of the underlying complementarity program (4) as $\theta$ falls to zero. Particularly, the consideration of the relative gap $\mathcal{G}(\theta, N)$ is meaningful and we report on the development of $\mathcal{G}(\theta, N)$ in Table 1. As expected, for fixed $N$, $\mathcal{G}(\theta, N)$ tends to zero as $\theta$ falls to zero.

**Table 1** Development of relative gap $\mathcal{G}(\theta, N)$

| $\mathcal{G}(\theta, N)$ | | $N$ | | | | |
|---|---|---|---|---|---|---|
| | | 80 | 160 | 240 | 320 | 400 |
| $\theta$ | 0.1 | 0.1750 | 0.1773 | 0.1781 | 0.1785 | 0.1788 |
| | 0.05 | 0.0901 | 0.0909 | 0.0912 | 0.0913 | 0.0914 |
| | 0.01 | 0.0182 | 0.0183 | 0.0184 | 0.0184 | 0.0184 |

## 5 Conclusions

Terminal complementarity constraints appear frequently in the context of optimal control of linear dynamical systems. Common examples arise in the fields of multi-agent control, satellite clustering [7], flocking [18], spacecraft formation [4], or natural gas balancing [15]. In this work, we have shown that relaxation methods which are well known from finite-dimensional complementarity programming, see [13], can be used to treat optimal control problems with terminal complementarity constraints. Exemplary, we analyzed Scholtes' relaxation technique but similar results are likely to hold for different relaxation approaches. On the one hand, it has been shown that global solutions of the relaxed surrogate problems converge in norm to a global minimizer of the complementarity problem. On the other hand, we demonstrated that a sequence of KKT points associated with the relaxed surrogate problems converges (under reasonable assumptions) to a C-stationary point of the complementarity problem. Numerical examples were presented to illustrate the method and visualize possible applications in multi-agent control.

There exist many interesting topics deserving further investigation like the computation of error estimates resulting from discretization or the practical solution of real-world problems exploiting the proposed numerical method.

**Acknowledgements** The authors would like to thank the anonymous reviewers for some valuable comments which helped us to improve the presentation of our results.

## References

1. Adams, R.A., Fournier, J.J.F.: Sobolev Spaces. Elsevier, Kidlington (2003)
2. Agarwal, R.P., O'Regan, D.: An Introduction to Ordinary Differential Equations. Springer, New York (2008)
3. Barnett, S., Cameron, R.G.: Introduction to Mathematical Control Theory. Oxford University Press, New York (1990)
4. Beard, R.W., Lawton, J., Hadaegh, F.Y.: A coordination architecture for spacecraft formation control. IEEE Trans. Control Syst. Technol. **9**(6), 777–790 (2001). https://doi.org/10.1109/87.960341
5. Benita, F., Mehlitz, P.: Optimal control problems with terminal complementarity constraints. SIAM J. Optim. **28**(4), 3079–3104 (2018). https://doi.org/10.1137/16M107637X
6. Bonnans, J.F., Shapiro, A.: Perturbation Analysis of Optimization Problems. Springer, New York (2002)
7. Bonuccelli, M.A., Martelli, F., Pelagatti, S.: Optimal packet scheduling in tree-structured LEO satellite clusters. Mobile Netw. Appl. **9**(4), 289–295 (2004). https://doi.org/10.1145/1023663.1023674

8. Butcher, J.C.: Numerical Methods for Ordinary Differential Equations. Wiley, Chichester (2016)
9. Dontchev, A.L.: Discrete approximations in optimal control. In: Mordukhovich, B.S., Sussmann, H.J. (eds.) Nonsmooth Analysis and Geometric Methods in Deterministic Optimal Control, pp. 59–80. Springer, New York (1996)
10. Gerdts, M.: Optimal Control of ODEs and DAEs. De Gruyter, Berlin (2012)
11. Hinze, M.: A variational discretization concept in control constrained optimization: the linear-quadratic case. Comput. Optim. Appl. **30**(1), 45–61 (2005). https://doi.org/10.1007/s10589-005-4559-5
12. Hinze, M., Rösch, A.: Discretization of Optimal Control Problems, pp. 391–430. Springer, Basel (2012). https://doi.org/10.1007/978-3-0348-0133-1_21
13. Hoheisel, T., Kanzow, C., Schwartz, A.: Theoretical and numerical comparison of relaxation methods for mathematical programs with complementarity constraints. Math. Program. Ser. A **137**(1), 257–288 (2013). https://doi.org/10.1007/s10107-011-0488-5
14. Jahn, J.: Introduction to the Theory of Nonlinear Optimization. Springer, Berlin (1996)
15. Kalashnikov, V.V., Benita, F., Mehlitz, P.: The natural gas cash-out problem: a bilevel optimal control approach. Math. Probl. Eng. (2015). https://doi.org/10.1155/2015/286083
16. Löber, J.: Optimal Trajectory Tracking of Nonlinear Dynamical Systems. Springer, Berlin (2017)
17. Luo, Z.Q., Pang, J.S., Ralph, D.: Mathematical Programs with Equilibrium Constraints. Cambridge University Press, Cambridge (1996)
18. Olfati-Saber, R.: Flocking for multi-agent dynamic systems: algorithms and theory. IEEE Trans. Autom. Control. **51**(3), 401–420 (2006). https://doi.org/10.1109/TAC.2005.864190
19. Outrata, J.V., Kočvara, M., Zowe, J.: Nonsmooth Approach to Optimization Problems with Equilibrium Constraints. Kluwer, Dordrecht (1998)
20. Robinson, S.M.: Stability theory for systems of inequalities, part II: differentiable nonlinear systems. SIAM J. Numer. Anal. **13**(4), 497–513 (1976). https://doi.org/10.1137/0713043
21. Scholtes, S.: Convergence properties of a regularization scheme for mathematical programs with complementarity constraints. SIAM J. Optim. **11**(4), 918–936 (2001). https://doi.org/10.1137/S1052623499361233
22. Schwartz, A., Polak, E.: Consistent approximations for optimal control problems based on Runge–Kutta integration. SIAM J. Control Optim. **34**(4), 1235–1269 (1996). https://doi.org/10.1137/S0363012994267352
23. Wächter, A., Biegler, L.T.: On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. Math. Program. Ser. A **106**(1), 25–57 (2006). https://doi.org/10.1007/s10107-004-0559-y
24. Zowe, J., Kurcyusz, S.: Regularity and stability for the mathematical programming problem in Banach spaces. Appl. Math. Optim. **5**(1), 49–62 (1979). https://doi.org/10.1007/BF01442543