Constrained and Sparse Switching Times Optimization via Augmented Lagrangian Proximal Methods

Alberto De Marchi

January 16, 2020

This is a preprint version of the paper

A. De Marchi, "Constrained and Sparse Switching Times Optimization via Augmented Lagrangian Proximal Methods," 2020 American Control Conference (ACC), Denver, CO, USA, 2020, pp. 3633–3638.
DOI: 10.23919/ACC45564.2020.9147892

Abstract:
In this paper we reformulate a switching times optimization problem with non-uniform switching costs and dwell-time constraints via direct multiple shooting, sparsity-inducing regularization and semi-continuous variables. The transformed problem has composite smooth/nonsmooth objective function and smooth constraints. Necessary optimality conditions for such problems are derived, resembling results from smooth optimization. A safeguarded, primal-dual, augmented Lagrangian proximal method is proposed for its numerical solution, and the global convergence toward points satisfying the necessary conditions is detailed. Finally, numerical results demonstrate the efficacy and limitations of the method.

Keywords:
Switching time optimization, mixed-integer optimal control, switched systems, switching cost, augmented Lagrangian methods, proximal methods.

Universität der Bundeswehr München, Department of Aerospace Engineering, Institute for Applied Mathematics and Scientific Computing, Werner-Heisenberg-Weg 39, 85577 Neubiberg/Munich, Germany.
Corresponding author: alberto.demarchi@unibw.de

1 Introduction

Switched dynamical systems consist of a collection of continuous subsystems with a switching law defining the active one at each time instant [17]. A similar structure arises in dynamical systems with discrete-valued control inputs [15]. Switching times optimization (STO) problems deal with the choice of the time instants at which the system dynamics change in order to minimize an objective function. In this paper, we consider STO problems for continuous-time autonomous nonlinear switched dynamical systems, with boundary conditions, non-uniform dwell-time constraints, and non-uniform switching (or activation) costs. This work does not adopt the insertion gradient approach [1], but extends the recently proposed cardinality-based formulation [7]. Dwell-time constraints are expressed through feasible sets for the switching intervals, namely the difference between switching times. Also, direct multiple shooting [15, 9] is preferred over single shooting [17, 7], which leads to a problem transcription with more structure and smoothness [8]. The decision variables of the corresponding finite-dimensional problem are the time instants at which the system dynamics switch, called switching times, and the system states at those times. With this formulation, the objective function consists of two, possibly nonconvex, terms: one is smooth and the other has an easy-to-compute proximal mapping. Some constraints are included, extending previous works [2, 7], to satisfy dynamics at switching times, boundary conditions, and final time constraints. A finite-dimensional constrained optimization problem with composite objective function and smooth constraints is obtained. Necessary optimality conditions are investigated for such class of problems, showing the relationship with the corresponding results in constrained smooth optimization. A simple algorithm is proposed based on embedding proximal methods [3, 12, 18] into the augmented Lagrangian framework [5, 4], possibly primal-dual [10, 13], and safeguarded [11]. Under mild assumptions, this method allows to avoid a pure penalty approach [16] and to deal with nonconvex constraints [6].

2 Preliminaries and Notation

$\mathbb{R}$, $\mathbb{R}_+$, and $\mathbb{Z}$ denote the sets of real numbers, non-negative real numbers, and integers, respectively. The identity matrix and the vector of ones are denoted by $\boldsymbol{I}$ and $\mathbf{1}$, respectively, and the extended real line as $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$. $[a, b]$, (a, b), $[a, b)$, and $(a, b]$ stand for closed, open, and half-open intervals, respectively, with end points a and b. $[a; b]$, $(a; b)$, $[a; b)$, and $(a; b]$ stand for discrete intervals, e.g., $[a; b] = [a, b] \cap \mathbb{Z}$. We denote $\|\cdot\|_0 : \mathbb{R} \to \mathbb{R}$ the function which satisfies $\|x\|_0 = 0$ if $x = 0$ and $\|x\|_0 = 1$ otherwise. Given a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a point $\mathbf{x}$ with $f(\mathbf{x})$ finite, a vector $\mathbf{v} \in \mathbb{R}^n$ is a *regular subgradient* of f at $\mathbf{x}$, denoted $\mathbf{v} \in \hat{\partial} f(\mathbf{x})$, if $f(\mathbf{z}) \geq f(\mathbf{x}) + \mathbf{v}^\top(\mathbf{z} - \mathbf{x}) + o(\|\mathbf{z} - \mathbf{x}\|)$ [14]. Given a nonempty closed set $B \subseteq \mathbb{R}^n$, we denote $\chi_B : \mathbb{R}^n \to \overline{\mathbb{R}}$ its *characteristic function*, namely $\chi_B(\mathbf{x}) = 0$ if $\mathbf{x} \in B$ and $\chi_B(\mathbf{x}) = \infty$ otherwise, $\mathrm{dist}_B : \mathbb{R}^n \to \mathbb{R}_+$ its *distance*, namely $\mathbf{x} \mapsto \min_{\mathbf{z} \in B} \|\mathbf{z} - \mathbf{x}\|$, and its *projection* $\mathrm{proj}_B : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, namely $\mathbf{x} \mapsto \arg\min_{\mathbf{z} \in B} \|\mathbf{z} - \mathbf{x}\|$. Given a positive scalar γ and a function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$, let $\mathrm{prox}_{\gamma g} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ denote the *proximal mapping* $\mathbf{x} \mapsto \arg\min_{\mathbf{z}} \left\{ 2\gamma g(\mathbf{z}) + \|\mathbf{z} - \mathbf{x}\|^2 \right\}$.

3 Problem Reformulation

Let Mayer cost $m : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}$, boundary conditions $\mathbf{b} : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_b}$, and time interval $[0, T]$, $T > 0$, be given; herein n_x is the state dimension. Let us consider a switched autonomous dynamical system with N modes, namely dynamics $\mathbf{f}_i : \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$ for $i \in [1; N]$. Switching times $\boldsymbol{\tau} \in \mathbb{R}^{N+1}$, with $\tau_1 = 0$ and $\tau_{N+1} = T$, refer to the time instants at which the system dynamics change [17]. Switching intervals $\mathbf{d} \in \mathbb{R}^N$ are defined by $d_i := \tau_{i+1} - \tau_i$, $i \in [1; N]$. The problem of interest is to

$$\begin{aligned} \text{find} \quad & \mathbf{x} \in W^{1,\infty}([0, T], \mathbb{R}^{n_x}),\ \mathbf{d} \in \mathbb{R}^N \\ \text{minimizing} \quad & m(\mathbf{x}(0), \mathbf{x}(T)) + s(\mathbf{d}) \\ \text{such that} \quad & \dot{\mathbf{x}}(t) = \mathbf{f}_i(\mathbf{x}(t)),\ \text{for } t \in [\tau_i, \tau_{i+1}),\ i \in [1; N] \\ & \mathbf{b}(\mathbf{x}(0), \mathbf{x}(T)) = \mathbf{0} \\ & \mathbf{d} \in D,\ \mathbf{1}^\top \mathbf{d} = T. \end{aligned} \tag{P1}$$

Herein, the feasible set $D := D_1 \times \cdots \times D_N \subseteq \mathbb{R}_+^N$ models dwell-time constraints on switching intervals $\mathbf{d}$ and the cost term $s : \mathbb{R}^N \to \mathbb{R}$ is defined as

$$s(\mathbf{d}) := \sum_{i=1}^{N} \sigma_i \|d_i\|_0 \tag{1}$$

for a given vector $\boldsymbol{\sigma}$ of non-negative switching costs, extending [7]. The Mayer cost m can account for running costs by augmenting the system state [9]. For free final time problems, the simplex constraint coupling the switching intervals $\mathbf{d}$ can be dropped from (P1).

3.1 Direct Multiple Shooting

Let us reformulate (P1) via direct multiple shooting, considering switching times as shooting nodes [9]; this unusual choice gives the problem a simpler structure. Then, a finite-dimensional problem is obtained, with switching intervals $\mathbf{d} \in D$ and states at switching times, namely $\boldsymbol{\xi} := \{\boldsymbol{\xi}_k | k \in [1; N+1]\}$, with $\boldsymbol{\xi}_k := \mathbf{x}(\tau_k) \in \mathbb{R}^{n_x}$, as decision variables. We consider functions $\boldsymbol{\Phi}_i : \mathbb{R}^{n_x} \times \mathbb{R} \to \mathbb{R}^{n_x}$, $i \in [1; N]$, as $\boldsymbol{\Phi}_i(\mathbf{x}_0, d) = \mathbf{x}^i(d)$, being $\mathbf{x}^i : \mathbb{R} \to \mathbb{R}^{n_x}$ the solution to the initial value problem

$$\dot{\mathbf{x}}(t) = \mathbf{f}_i(\mathbf{x}(t)),\ \mathbf{x}(0) = \mathbf{x}_0,\ t \in [0, d] \tag{2}$$

for any $\mathbf{x}_0 \in \mathbb{R}^{n_x}$, $d \in \mathbb{R}$. The constraint involving the ordinary differential equation in (P1) is substituted, with functions $\boldsymbol{\Phi}_i$, by equalities induced by the shooting intervals [15, 9]. In order to deal with possibly disconnected feasible set D, we embed the constraint $\mathbf{d} \in D$ in (P1) into the nonsmooth term via characteristic function χ_D, defining $s_D := s + \chi_D$.

Hence, the problem is to

$$\begin{aligned} \text{find} \quad & \boldsymbol{\xi} \in \mathbb{R}^{n_x \times (N+1)},\ \mathbf{d} \in \mathbb{R}^N \\ \text{minimizing} \quad & m(\boldsymbol{\xi}_1, \boldsymbol{\xi}_{N+1}) + s_D(\mathbf{d}) \\ \text{such that} \quad & \boldsymbol{\xi}_{i+1} = \boldsymbol{\Phi}_i(\boldsymbol{\xi}_i, d_i) \qquad\qquad i \in [1; N] \\ & \mathbf{b}\left(\boldsymbol{\xi}_1, \boldsymbol{\xi}_{N+1}\right) = \mathbf{0} \\ & \mathbf{1}^\top \mathbf{d} = T. \end{aligned} \tag{P2}$$

Functions $\boldsymbol{\Phi}_i$ are evaluated via numerical integrators, thus discretizing the time domain and applying suitable integration schemes. Computing the sensitivity of $\boldsymbol{\Phi}_i(\boldsymbol{\xi}_i, d_i)$ to $\boldsymbol{\xi}_i$ is a standard task in multiple shooting methods; see [15, 9]. The sensitivity to switching interval d_i comes from dynamics at final state. Notice that there is often a gap between continuous- and discrete-time dynamics, possibly converging with finer discretizations [9]. We consider the sensitivities for the discrete-time dynamics, matching the numerical integration scheme. Moreover, by decoupling dynamics $\mathbf{f}_i$ through $\boldsymbol{\Phi}_i$ and associated constraints, the direct multiple shooting approach helps avoiding issues related to nonsmoothness [8].

One can readily extend (P2): cost function $\tilde{m}(\boldsymbol{\xi}, \mathbf{d})$ can have switching states and intervals as arguments, not initial and final state only; simple constraints on switching states can be included in the nonsmooth term $\tilde{s}(\boldsymbol{\xi}, \mathbf{d})$; coupled boundary and switching conditions on switching states can be considered as $\tilde{\mathbf{b}}(\boldsymbol{\xi}) = \mathbf{0}$.

3.2 Proximal Operator

The proximal mapping of s_D can be evaluated component-wise, thanks to the separable structure of both the switching cost s and the feasible set D. Denoting $s_i := \sigma_i \|\cdot\|_0 + \chi_{D_i}$, it holds $z_i := \left[\mathrm{prox}_{\gamma s_D}(\mathbf{d})\right]_i = \mathrm{prox}_{\gamma s_i}(d_i)$, $i \in [1; N]$. Since characteristic function χ_{D_i} guarantees the projection onto D_i, for $i \in [1; N]$, it holds

$$z_i = \operatorname*{arg\,min}_{u_i \in D_i} \left\{ 2\gamma\sigma_i \|u_i\|_0 + (u_i - d_i)^2 \right\}.$$

If either $0 \notin D_i$ or $\sigma_i = 0$, then the first term attains a constant value in D_i, hence the problem turns into a projection, yielding $z_i = \mathrm{proj}_{D_i}(d_i) =: p_i$. On the other hand, if $0 \in D_i$ and $\sigma_i > 0$, the switching cost plays a role. Comparing the costs associated to $u_i \neq 0$ and $u_i = 0$, namely their ratio $r_i := \left[2\gamma\sigma_i + \mathrm{dist}^2_{D_i}(d_i)\right] / d_i^2$, a minimizer can be found: $z_i = p_i$ if $r_i < 1$, $z_i = \{0\} \cup p_i$ if $r_i = 1$, and $z_i = 0$ otherwise.

4 Constrained Composite Optimization

In Section 3, we have discussed a reformulation for the problem of interest, obtaining (P2). In a more general setting, the problem is to

$$\begin{aligned} \text{find} \quad & \mathbf{x} \in \mathbb{R}^n \\ \text{minimizing} \quad & f(\mathbf{x}) + g(\mathbf{x}) \\ \text{such that} \quad & \mathbf{c}(\mathbf{x}) = \mathbf{0} \end{aligned} \tag{P3}$$

with $f : \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{c} : \mathbb{R}^n \to \mathbb{R}^m$ sufficiently smooth, and $g : \mathbb{R}^n \to \mathbb{R}$ with an easily computable proximal mapping. Dimensions in (P3) are related to (P2) by $n = n_x(N+1)+N$ and $m = n_x N + n_b + 1$. In this context, inequality constraints can be considered by adding equality constraints with slack variables and a characteristic function in the nonsmooth cost term. As mentioned in Section 1, penalty methods [16] or interior point methods [6] have been used to deal with constrained composite optimization, with either softened or convex constraints. The augmented Lagrangian approach is considered here because (i) it is based on a sequence of unconstrained or simply constrained subproblems, (ii) it is often superior to pure penalty methods, (iii) it can handle nonconvex constraints, and (iv) it enjoys good warm-starting capabilities; see [5, 4]. In fact, this approach for solving (P3) is supported by Theorem 1 below, which extends results from smooth optimization. In particular, we adopt the primal-dual augmented Lagrangian function [13, 10], with safeguard [11]. Topics such as higher-order updates and infeasibility detection are beyond the scope and not considered here.

4.1 Necessary Criticality Conditions

Note 1. *Consider problem* $\min_{\mathbf{x}}\{f(\mathbf{x}) + g(\mathbf{x})\}$. *A point* $\mathbf{x}^\star$ *is called* optimal *if* $\mathbf{x}^\star \in \arg\min_{\mathbf{x}}\{f(\mathbf{x}) + g(\mathbf{x})\}$, critical *if* $\mathbf{x}^\star \in \operatorname{prox}_{\gamma g}(\mathbf{x}^\star - \gamma\nabla f(\mathbf{x}^\star))$ *for some* $\gamma \in (0, 1/L_f)$, L_f *being the Lipschitz constant of* ∇f, *and* stationary *if* $\mathbf{0} \in \nabla f(\mathbf{x}^\star) + \hat{\partial} g(\mathbf{x}^\star)$ *[18]. Consider constraints* $\mathbf{c}(\mathbf{x}) = \mathbf{0}$ *and the feasible set* $\mathcal{C} := \{\mathbf{x} | \mathbf{c}(\mathbf{x}) = \mathbf{0}\}$. *A point* $\mathbf{x}^\star$ *is called* feasible *if* $\mathbf{x}^\star \in \mathcal{C}$. *Consider* (P3). *A point* $\mathbf{x}^\star$ *is a* global solution *if* $\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathcal{C}}\{f(\mathbf{x}) + g(\mathbf{x})\}$, *and a* (local) solution *if it is feasible and there exists a closed ball* $\mathcal{B} := \{\mathbf{x} | \|\mathbf{x} - \mathbf{x}^\star\| \leq \epsilon\}$, $\epsilon > 0$, *such that, for all* $\mathbf{x} \in \mathcal{B} \cap \mathcal{C}$, *it holds* $(f+g)(\mathbf{x}) \geq (f+g)(\mathbf{x}^\star)$.

Notice the similarity between L-stationarity [2] and criticality [18], and the relationship with optimality and stationarity concepts. Let us consider the following assumptions:

(A_1). functions $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{c} : \mathbb{R}^n \to \mathbb{R}^m$ are differentiable with Lipschitz continuous gradient;

(A_2). function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper and closed;

(A_3). function $f + g$ is bounded below in $\mathbb{R}^n$;

(A_4). there exists a (local) solution $\mathbf{x}^\star$ to (P3);

(A_5). gradients of the equality constraints are linearly independent at $\mathbf{x}^\star$ (LICQ);

(A_6). there exists a set $\mathcal{G}$ compact and a scalar $\epsilon > 0$ such that $\forall \mathbf{x} \in \{\mathbf{x} | \|\mathbf{x} - \mathbf{x}^\star\| \leq \epsilon\}$ it holds $\hat{\partial} g(\mathbf{x}) \subseteq \mathcal{G}$.

Some assumptions could be relaxed [18], but these are considered here for simplicity. Necessary conditions for optimality closely follow results from smooth optimization, as well as the proof outline [4].

Theorem 1 (Necessary conditions)**.** *Consider* (P3). *Let* $\mathbf{x}^\star$ *denote a (local) solution and Assumptions* (A_1)*–*(A_6) *be satisfied. Define function* $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ *as the mapping* $(\mathbf{x}, \mathbf{y}) \mapsto f(\mathbf{x}) - \mathbf{y}^\top \mathbf{c}(\mathbf{x})$. *Then, (i) there exists* $\mathbf{y}^\star \in \mathbb{R}^m$ *such that the following (criticality) condition holds*

$$\mathbf{x}^\star \in \operatorname{prox}_{\gamma g}\left(\mathbf{x}^\star - \gamma \nabla_x \mathcal{L}(\mathbf{x}^\star, \mathbf{y}^\star)\right) \tag{3}$$

for some $\gamma \in (0, 1/L_{\mathcal{L}})$, $L_{\mathcal{L}}$ *being the Lipschitz constant of* $\mathcal{L}(\cdot, \mathbf{y}^\star)$. *Furthermore, (ii) the following (stationarity) condition holds*

$$\mathbf{0} \in \nabla f(\mathbf{x}^\star) + \hat{\partial} g(\mathbf{x}^\star) - \nabla \mathbf{c}(\mathbf{x}^\star)\mathbf{y}^\star. \tag{4}$$

Proof. Introduce for $k = 1, 2, \ldots$ the cost function $\phi^k : \mathbb{R}^n \to \mathbb{R}$ defined by

$$\phi^k(\mathbf{x}) := (f + g)(\mathbf{x}) + \frac{k}{2}\|\mathbf{c}(\mathbf{x})\|^2 + \frac{\alpha}{2}\|\mathbf{x} - \mathbf{x}^\star\|^2,$$

where $\alpha > 0$. Since $\mathbf{x}^\star$ is a solution, there exists a closed, nonempty ball $\mathcal{B}$ such that $(f+g)(\mathbf{x}^\star) \leq (f+g)(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{B} \cap \mathcal{C}$; see Note 1. Consider $\mathbf{x}^k \in \arg\min_{\mathbf{x} \in \mathcal{B}} \phi^k(\mathbf{x})$ and the sequence $\{\mathbf{x}^k\}$. Since $\phi^k(\mathbf{x}^k) \leq \phi^k(\mathbf{x}^\star) = (f + g)(\mathbf{x}^\star)$ for all k, taking the limit $k \to \infty$ gives $\mathbf{c}(\mathbf{x}^\star) = \mathbf{0}$, due to ($A_3$), ($A_4$) and continuity of norms, and $\{\mathbf{x}^k\} \to \mathbf{x}^\star$; see [4]. Then, for k sufficiently large, $\mathbf{x}^k$ is interior to $\mathcal{B}$ and hence an unconstrained minimizer of ϕ^k. Then, it is also a critical point [18, Prop. 3.5], namely it satisfies

$$\mathbf{x}^k \in \operatorname{prox}_{\gamma g}\left(\mathbf{x}^k - \gamma \nabla \ell^k(\mathbf{x}^k)\right) \tag{5}$$

for any $\gamma \in (0, 1/L_\ell)$, being $\ell^k := \phi^k - g$ the smooth part of ϕ^k and L_ℓ the Lipschitz constant of $\nabla \ell^k$. Furthermore, $\mathbf{x}^k$ is also a stationary point [18, Prop. 3.5], i.e., it satisfies

$$\mathbf{0} \in \hat{\partial} \phi^k(\mathbf{x}^k) \tag{6}$$

thanks to (A_2). Under (A_5), for k sufficiently large, matrix $\boldsymbol{J}(\mathbf{x}^k) := \nabla \mathbf{c}(\mathbf{x}^k)^\top \nabla \mathbf{c}(\mathbf{x}^k)$ is invertible, since $\{\mathbf{x}^k\} \to \mathbf{x}^\star$; see [4]. Expanding the term $\hat{\partial} \phi^k(\mathbf{x}^k)$ using properties of subgradients, left-multiplying by $\nabla \mathbf{c}(\mathbf{x}^k)^\top$ and solving, an inclusion for $k\mathbf{c}(\mathbf{x}^k)$ is obtained. Hence, for all $k = 1, 2, \ldots$, there exists a $\mathbf{p}^k \in \hat{\partial} g(\mathbf{x}^k)$ such that the equality holds. The limit $k \to \infty$ gives

$$\{k\mathbf{c}(\mathbf{x}^k)\} \to -\boldsymbol{J}(\mathbf{x}^\star)^{-1} \nabla \mathbf{c}(\mathbf{x}^\star)^\top \left[\nabla f(\mathbf{x}^\star) + \mathbf{p}^\star\right] =: -\mathbf{y}^\star$$

with $\mathbf{p}^\star$ an accumulation point of $\{\mathbf{p}^k\}$, thanks to (A_6). Thus, $\{\nabla \ell^k(\mathbf{x}^k)\} \to \nabla f(\mathbf{x}^\star) - \nabla \mathbf{c}(\mathbf{x}^\star)^\top \mathbf{y}^\star$ for $k \to \infty$, which, taking the limit in (5)–(6) and substituting, gives (3)–(4). □

In general, if function g is nonconvex, the proximal mapping is set-valued, and it might be difficult to verify condition (3) in practice. The following result helps in this direction.

Lemma 1. *Consider the mappings in Section 2. Then*

$$0 = \operatorname{dist}_{\operatorname{prox}_{\gamma g}(\mathbf{x}^\star - \gamma \nabla_x \mathcal{L}(\mathbf{x}^\star, \mathbf{y}^\star))} (\mathbf{x}^\star) \tag{7}$$

is equivalent to condition (3), for any γ.

Proof. Due to definition in Section 2 and continuity of norms, (7) implies that $\mathbf{x}^\star$ is an element of $\operatorname{prox}_{\gamma g}(\mathbf{x}^\star - \gamma \nabla_x \mathcal{L}(\mathbf{x}^\star, \mathbf{y}^\star))$, which proves (7) $\Rightarrow$ (3). Conversely, the fact that $\mathbf{x}^\star$ is the unique global minimizer of the distance to $\mathbf{x}^\star$ itself, due to properties of norms, together with condition (3), implies (7), thus proving (3) $\Rightarrow$ (7). The two implications give the equivalence. □

Based on definitions in Note 1, necessary conditions in Theorem 1, and Lemma 1, we introduce the concept of approximate solution. Given positive tolerances on feasibility and criticality, denoted $\eta^\star$ and $\epsilon^\star$, respectively, we refer to a point $(\mathbf{x}_k, \mathbf{y}_k)$ as an *approximate (local) solution* to (P3) if it satisfies

$$\left\| \operatorname{dist}_{\operatorname{prox}_{\gamma g}(\mathbf{x}_k - \gamma \nabla_x \mathcal{L}(\mathbf{x}_k, \mathbf{y}_k))} (\mathbf{x}_k) \right\| \leq \gamma \epsilon^\star \tag{8a}$$

$$\|\mathbf{c}(\mathbf{x}_k)\| \leq \eta^\star \tag{8b}$$

for some $\gamma \in (0, 1/L_{\mathcal{L}})$, $L_{\mathcal{L}}$ being the Lipschitz constant of $\nabla_x \mathcal{L}(\cdot, \mathbf{y}_k)$.

5 Primal-Dual Augmented Lagrangian Proximal Method

Necessary conditions provided by Theorem 1 resemble results from smooth optimization, in particular the Lagrange function $\mathcal{L}$ and the existence of Lagrange multipliers $\mathbf{y}^\star$. Based on this, in the spirit of augmented Lagrangian methods [5, 4], we aim at solving (P3) through a sequence of subproblems which

$$\begin{aligned} \text{find} \quad & \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m \\ \text{minimizing} \quad & F(\mathbf{x}, \mathbf{y}, \mathbf{y}_k^e, \mu_k) + G(\mathbf{x}, \mathbf{y}, \beta_k) \end{aligned} \tag{P4}$$

where $\{\mathbf{y}_k^e\}$ is a sequence of estimates of the Lagrange multipliers, $\{\mu_k\}$ is a sequence of positive penalty parameters (a positive definite, diagonal matrix could be adopted too [5]) and $\{\beta_k\}$ is a sequence of positive bound parameters. Function F, defined by

$$F(\mathbf{x}, \mathbf{y}, \mathbf{y}^e, \mu) := f(\mathbf{x}) + \frac{1}{2\mu} \|\mathbf{c}(\mathbf{x}) - \mu \mathbf{y}^e\|^2 + \frac{1}{2\mu} \|\mathbf{c}(\mathbf{x}) + \mu(\mathbf{y} - \mathbf{y}^e)\|^2, \tag{9}$$

recalls the primal-dual augmented Lagrange function [13]. Function G consists of the nonsmooth term g and the characteristic function $\chi_{\lfloor \beta \rceil}$, denoting $\lfloor \beta \rceil$ the closed hyper-interval $[-\beta, \beta]^m$, namely

$$G(\mathbf{x}, \mathbf{y}, \beta) := g(\mathbf{x}) + \chi_{\lfloor \beta \rceil}(\mathbf{y}). \tag{10}$$

In fact, through $\chi_{\lfloor\beta\rceil}$ in (10), explicit bounds are imposed on the dual variables $\mathbf{y}$, thus, regularizing the associated subproblems, as discussed in [10, Section 4]. Also, requiring bounded estimates $\mathbf{y}^e$ for (9) leads to a safeguarded method [11]. Between iterations, such estimates can be obtained with a first-order multiplier update. The primal and primal-dual updates [13], denoted $\hat{\mathbf{y}}^p$ and $\hat{\mathbf{y}}$, respectively, read

$$\hat{\mathbf{y}}_k^p := \mathbf{y}_k^e - \mathbf{c}(\mathbf{x}_k^\star)/\mu_k, \tag{11a}$$
$$\hat{\mathbf{y}}_k := 2\hat{\mathbf{y}}_k^p - \mathbf{y}_k^\star, \tag{11b}$$

given solution $(\mathbf{x}_k^\star, \mathbf{y}_k^\star)$ to (P4), multiplier estimate $\mathbf{y}_k^e$, and penalty parameter μ_k. Since (P4) is unconstrained, and $\chi_{\lfloor\beta\rceil}$ is a characteristic function, the following necessary conditions hold, for some $\gamma > 0$ [12, 18]:

$$\mathbf{x}_k^\star \in \operatorname{prox}_{\gamma g}\left(\mathbf{x}_k^\star - \gamma\nabla_x F(\mathbf{x}_k^\star, \mathbf{y}_k^\star, \mathbf{y}_k^e, \mu_k)\right), \tag{12a}$$
$$\mathbf{y}_k^\star \in \operatorname{proj}_{\lfloor\beta_k\rceil}\left(\mathbf{y}_k^\star - \gamma\nabla_y F(\mathbf{x}_k^\star, \mathbf{y}_k^\star, \mathbf{y}_k^e, \mu_k)\right), \tag{12b}$$

and, similarly to (8), an approximate solution to (P4) is characterized by

$$\left\| \begin{pmatrix} \operatorname{dist}_{\operatorname{prox}_{\gamma g}(\mathbf{x}_k^\star - \gamma\nabla_x F(\mathbf{x}_k^\star, \mathbf{y}_k^\star))}(\mathbf{x}_k^\star) \\ \operatorname{dist}_{\operatorname{proj}_{\lfloor\beta_k\rceil}(\mathbf{y}_k^\star - \gamma\nabla_y F(\mathbf{x}_k^\star, \mathbf{y}_k^\star))}(\mathbf{y}_k^\star) \end{pmatrix} \right\| \leq \gamma\epsilon_k \tag{13}$$

given an optimality tolerance $\epsilon_k > 0$.

Note 2. *The method we are proposing does not differ substantially from well-established methods in the augmented Lagrangian framework. Indeed, looking at proximal mappings as generalized projections,* (P4) *seems a generalization of bound-constrained subproblem in BCL [5] and pdBCL [13] methods. Furthermore, by modifying functions F and G in* (9)–(10) *or estimates $\mathbf{y}^e$, classical (Hestenes-Powell's) augmented Lagrangian and quadratic penalty approaches can be recovered, see [10, Table 1]. Nonetheless,* (P4) *is significantly dissimilar to classical subproblems, in that it is an unconstrained composite problem, which is amenable to proximal methods, such as FISTA [3], PNOPT [12] and PANOC [16], among others.*

5.1 Algorithm

We illustrate Algorithm 1, also referred to as `pdALX`, based on the primal-dual Augmented Lagrangian proXimal method considered above. This resembles a proximal version of the primal-dual bound-constrained Lagrangian (pdBCL) method [13, 10], based upon the bound-constrained Lagrangian (BCL) method [5]. Algorithm 1 consists of outer and inner iterations: the latter for solving the subproblems via any suitable proximal method, the former for reducing the constraint violation while estimating the Lagrange multipliers. The approximate solution in Step 6 is defined according to (13) and is obtained via any suitable proximal method, adopting $(\mathbf{x}_k, \mathbf{y}_k^e)$ as initial guess. Many parameters are used in Algorithm 1, which are defined as follows [5, 13]: initial penalty $\mu_0 \in (0, 1)$ and bound $\beta_0 \in (0, \infty)$ parameter, initial feasibility $\eta_0 \in [\eta_\star, 1/2)$ and optimality $\epsilon_0 \in [\epsilon^\star, 1/2)$

Algorithm 1 `pdALX`

Input: $\mathbf{x}_0$, $\mathbf{y}_0$

Output: $\mathbf{x}^\star$, $\mathbf{y}^\star$

set μ_0, β_0, η_0, ϵ_0, $\eta^\star$, $\epsilon^\star$
set ρ_y, $\overline{\alpha}_\mu$, $\overline{\alpha}_\beta$, $\overline{\alpha}_\eta$, $\overline{\alpha}_\epsilon$, $\underline{\alpha}_\mu$, $\underline{\alpha}_\beta$, $\underline{\alpha}_\eta$, $\underline{\alpha}_\epsilon$
$k \leftarrow 0$
while $(\mathbf{x}_k, \mathbf{y}_k)$ does not satisfy (8) **do**
  $\mathbf{y}_k^e \leftarrow \mathrm{proj}_{\lfloor \beta_k \rceil}(\mathbf{y}_k)$
  $(\mathbf{x}_k^\star, \mathbf{y}_k^\star) \leftarrow$ approximate solution to (P4)
  $\hat{\mathbf{y}}_k \leftarrow 2\,[\mathbf{y}_k^e - \mathbf{c}(\mathbf{x}_k^\star)/\mu_k] - \mathbf{y}_k^\star$
  if $\|\mathbf{c}(\mathbf{x}_k^\star)\| \leq \eta_k$ **then**
    if $\|\mathbf{y}_k^\star\|_\infty \geq \rho_y \beta_k$ **then**
      $\mu_{k+1} \leftarrow \overline{\alpha}_\mu \mu_k$, $\beta_{k+1} \leftarrow \overline{\alpha}_\beta \beta_k$
    else
      $\mu_{k+1} \leftarrow \mu_k$, $\beta_{k+1} \leftarrow \beta_k$
    end if
    $\eta_{k+1} \leftarrow \eta_k \mu_{k+1}^{\overline{\alpha}_\eta}$, $\epsilon_{k+1} \leftarrow \epsilon_k \mu_{k+1}^{\overline{\alpha}_\epsilon}$
  else
    $\mu_{k+1} \leftarrow \underline{\alpha}_\mu \mu_k$, $\beta_{k+1} \leftarrow \underline{\alpha}_\beta \beta_k$
    $\eta_{k+1} \leftarrow \eta_0 \mu_{k+1}^{\underline{\alpha}_\eta}$, $\epsilon_{k+1} \leftarrow \epsilon_0 \mu_{k+1}^{\underline{\alpha}_\epsilon}$
  end if
  $(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) \leftarrow (\mathbf{x}_k^\star, \hat{\mathbf{y}}_k)$
  $k \leftarrow k + 1$
end while
return $(\mathbf{x}^\star, \mathbf{y}^\star) \leftarrow (\mathbf{x}_k, \mathbf{y}_k)$

tolerance; feasibility $\eta^\star \in [0, 1/2)$ and optimality $\epsilon^\star \in [0, 1/2)$ tolerance; margin $\rho_y \in [0, 1]$, penalty $\overline{\alpha}_\mu, \underline{\alpha}_\mu \in (0, 1)$, and bound $\overline{\alpha}_\beta, \underline{\alpha}_\beta > 1$ factors; feasibility $\overline{\alpha}_\eta \in (0, \min(\overline{\alpha}_\epsilon, 1))$, $\underline{\alpha}_\eta \in (0, \min(\underline{\alpha}_\epsilon, 1))$, and optimality $\overline{\alpha}_\epsilon, \underline{\alpha}_\epsilon > 0$ tolerance factors. The proof of global convergence in Section 5.2 requires $\overline{\alpha}_\mu \overline{\alpha}_\beta, \underline{\alpha}_\mu \underline{\alpha}_\beta < 1$.

Another algorithm, here referred to as `ALX`, is obtained by neglecting the last term in (9)–(10) and discarding decision variables $\mathbf{y}$. The resulting method is similar to the classical augmented Lagrangian approach: it solves smaller subproblems, it cannot take advantage of the dual regularization [13] but it can be safeguarded [11]. These two algorithms, namely `ALX` and `pdALX`, which relate to each other as BCL [5] to pdBCL [13], are compared in Section 6.

5.2 Convergence analysis

This section gives results about sequences generated by `pdALX` and shows its global convergence, under standing Assumptions (A_1)–(A_6) and supposing that

(A_7). a compact set $\mathcal{B}_x \subset \mathbb{R}^n$ contains the sequence $\{\mathbf{x}_k^\star\}$;

(A_8). denoting K a subsequence of integers such that $\lim_{k\in K} \mathbf{x}_k^\star = \mathbf{x}^\star$, a compact set $\mathcal{B}_y \subset \mathbb{R}^m$ contains the subsequence $\{\mathbf{y}_k^\star\}_K$.

Lemma 2. *Consider $\hat{\mathbf{y}}_k^p$ and $\hat{\mathbf{y}}_k$ defined in (11). Then, executing Algorithm 1, for k sufficiently large, the following holds: $\hat{\mathbf{y}}_k = \hat{\mathbf{y}}_k^p = \mathbf{y}_k^\star$.*

Proof. Consider necessary condition (12b). The fact that, by construction, β_k is eventually sufficiently large and $\gamma \in (0, \infty)$ gives $\mathbf{0} = \nabla_y F(\mathbf{x}_k^\star, \mathbf{y}_k^\star, \mathbf{y}_k^e, \mu_k)$. Using definitions (9) and (11) yields $\mathbf{0} = \mathbf{c}(\mathbf{x}^\star) + \mu_k(\mathbf{y}_k^\star - \mathbf{y}_k^e)$, and then $\mathbf{y}_k^\star = \mathbf{y}_k^e - \mathbf{c}(\mathbf{x}_k^\star)/\mu_k = \hat{\mathbf{y}}_k^p = \hat{\mathbf{y}}_k$. □

Lemma 3. *Consider Algorithm 1 and let Assumptions (A_1)–(A_8) hold. Let $\{\mu_k\}$, $\{\beta_k\}$, and $\{\epsilon_k\}$ be given sequences of positive numbers such that $\{\epsilon_k\} \to 0$, and let $\{\mathbf{y}_k^e\}$ be any sequence of vectors in $\mathbb{R}^m$. Let $\{(\mathbf{x}_k^\star, \mathbf{y}_k^\star)\}$ be a sequence of points satisfying (13). Denoting $\mathbf{x}^\star$ a limit point of $\{\mathbf{x}_k^\star\}$, let K be a subsequence of the integers such that $\lim_{k\in K} \mathbf{x}_k^\star = \mathbf{x}^\star$ and set $\mathbf{y}^\star := \lim_{k\in K} \hat{\mathbf{y}}_k$, considering $\hat{\mathbf{y}}_k$ defined in (11b). If $\mathbf{c}(\mathbf{x}^\star) = \mathbf{0}$, then $(\mathbf{x}^\star, \mathbf{y}^\star)$ satisfies (3), ii.e., it is a critical point for (P3).*

Proof. Since $\{\epsilon_k\} \to 0$, $\gamma \in (0, \infty)$, and condition (13) is satisfied at Step 6, it holds

$$\lim_{k\in K} \left\| \begin{pmatrix} \mathrm{dist}_{\mathrm{prox}_{\gamma g}(\mathbf{x}_k^\star - \gamma \nabla_x F(\mathbf{x}_k^\star, \mathbf{y}_k^\star))}(\mathbf{x}_k^\star) \\ \mathrm{dist}_{\mathrm{proj}_{\lfloor \beta_k \rceil}(\mathbf{y}_k^\star - \gamma \nabla_y F(\mathbf{x}_k^\star, \mathbf{y}_k^\star))}(\mathbf{y}_k^\star) \end{pmatrix} \right\| \leq \lim_{k\in K} \gamma \epsilon_k = 0,$$

which implies (12), due to continuity of norms, for $k \to \infty$ in K. Using definition (9) and Lemma 2, (12a) yields

$$\mathbf{x}_k^\star \in \mathrm{prox}_{\gamma g}\left(\mathbf{x}_k^\star - \gamma\left(\nabla f(\mathbf{x}_k^\star) - \nabla \mathbf{c}(\mathbf{x}_k^\star)\mathbf{y}_k^\star\right)\right)$$

for k sufficiently large. Condition (3) is obtained by substituting function $\mathcal{L}$ as defined in Theorem 1, proving the result. □

Lemma 3 shows that convergence to critical points occurs provided the constraint violation is forced to zero. Some lemmas are stated to simplify the global convergence proof.

Lemma 4. *Algorithm 1 generates sequences $\{\beta_k\}$, $\{\mathbf{y}_k^e\}$, and $\{\mathbf{y}_k^\star\}$ such that $\|\mathbf{y}_k^e\|_\infty \leq \beta_k$, $\|\mathbf{y}_k^\star\|_\infty \leq \beta_k$ for all k.*

Proof. Projection $\mathrm{proj}_{\lfloor \beta_k \rceil}$ at Step 5 and characteristic function $\chi_{\lfloor \beta_k \rceil}$ in (10) guarantee the result. □

Lemma 5. *Suppose that $\{\mu_k\} \to 0$ as Algorithm 1 is executed. Then $\{\mu_k \beta_k\} \to 0$.*

Proof. By construction, it holds $\mu_k \beta_k = \overline{\alpha}^{\overline{n}_k} \underline{\alpha}^{\underline{n}_k} \mu_0 \beta_0$, with $\overline{\alpha} := \overline{\alpha}_\mu \overline{\alpha}_\beta < 1$ and $\underline{\alpha} := \underline{\alpha}_\mu \underline{\alpha}_\beta < 1$, for all k and for some non-decreasing sequences $\{\overline{n}_k\}$ and $\{\underline{n}_k\}$. By inspection, since $\{\mu_k\} \to 0$, it is $\{\overline{n}_k + \underline{n}_k\} \to \infty$. This yields $\mu_k \beta_k \leq [\max(\overline{\alpha}, \underline{\alpha})]^{(\overline{n}_k + \underline{n}_k)} \mu_0 \beta_0 \to 0$, since $\mu_0 \beta_0 < \infty$. □

Lemma 6. *Suppose that $\{\mu_k\} \to 0$ as Algorithm 1 is executed. Then $\{\mu_k \|\mathbf{y}_k^e\|\} \to 0$ and $\{\mu_k \|\mathbf{y}_k^\star\|\} \to 0$.*

Proof. See [13, Section 4.3]. □

Lemma 7. *Executing Algorithm 1, iterates satisfy* $\|\mathbf{c}(\mathbf{x}_k^\star)\| \leq \mu_k \|\mathbf{y}_k^e\| + \mu_k \|\mathbf{y}_k^\star\|/2 + \mu_k \|\hat{\mathbf{y}}_k\|/2$ *for all* k.

Proof. Use (11) and the triangle inequality. □

Theorem 2 (Global subsequential convergence)**.** *Let Assumptions* (A_1)*–*(A_8) *hold and* $\{(\mathbf{x}_k^\star, \mathbf{y}_k^\star)\}$ *be the sequence of points generated by Algorithm 1 with tolerances* $\eta^\star = 0$ *and* $\epsilon^\star = 0$*. Then Lemma 3 holds, and* $(\mathbf{x}^\star, \mathbf{y}^\star)$ *as defined in Lemma 3 is a feasible critical point for* (P3)*.*

Proof. Algorithm 1 generates sequences $\{\mu_k\}$, $\{\beta_k\}$, $\{\epsilon_k\}$, and $\{\mathbf{y}_k^e\}$ such that $\{\epsilon_k\} \to 0$, and points $(\mathbf{x}_k^\star, \mathbf{y}_k^\star)$ satisfying condition (13). Hence, Lemma 3 holds. Then, to show that $(\mathbf{x}^\star, \mathbf{y}^\star)$ is a feasible critical point for (P3), it remains to prove that $\mathbf{x}^\star$ is feasible, i.e., $\mathbf{c}(\mathbf{x}^\star) = \mathbf{0}$. There are two cases to consider: (a) $\{\mu_k\} \to 0$, and (b) $\{\mu_k\}$ is bounded away from zero.

(a. Lemma 2 and Lemma 6 give $\{\mu_k \|\hat{\mathbf{y}}_k\|\} \to 0$ as $\{\mu_k\} \to 0$. Then, Lemma 7 implies $\{\|\mathbf{c}(\mathbf{x}_k^\star)\|\} \to 0$, since the right-hand side goes to zero.

(b. Since $\{\mu_k\}$ is bounded away from zero, there exists an integer k_μ such that $\|\mathbf{c}(\mathbf{x}_k^\star)\| \leq \eta_k$ for all $k \geq k_\mu$. Since $\{\eta_k\} \to 0$, this implies $\{\|\mathbf{c}(\mathbf{x}_k^\star)\|\} \to 0$.

Properties of norms and continuity of $\mathbf{c}$ yield $\mathbf{c}(\mathbf{x}^\star) = \mathbf{0}$, concluding the proof. □

Although a detailed analysis of the proposed methods is beyond the scope of this paper, we conjecture most properties of classical or primal-dual augmented Lagrangian methods are retained, e.g., local convergence rates.

6 Numerical Results

In this section, we illustrate the proposed method through the optimal switching control problem of Lotka-Volterra dynamics [15, 17]. Including the tracking cost via differential state x_3, the Mayer cost is $m(\mathbf{x}(0), \mathbf{x}(T)) = x_3(T)$.

$$\begin{cases} \dot{x}_1 = x_1 - x_1 x_2 - x_1 v_1 \\ \dot{x}_2 = x_1 x_2 - x_2 - x_2 v_2 \\ \dot{x}_3 = (x_1 - 1)^2 + (x_2 - 1)^2 \end{cases} \tag{14}$$

Referring to (P1), dynamics $\mathbf{f}_i$ are based on (14) and the control sequence $\{\mathbf{v}^-, \mathbf{v}^+, \mathbf{v}^-, \dots\}$, with $\mathbf{v}^- = (0, 0)^\top$ and $\mathbf{v}^+ = (0.4, 0.2)^\top$, for $i \in [1; N]$, $N = 20$. We consider fixed final time $T = 12$, fixed initial state $\mathbf{x}(0) = (0.5, 0.7, 0)^\top$, terminal conditions $x_{[1;2]}(T) \in [0.95, 1.05]$, feasible set D consisting of $D_i = \{0\} \cup [\underline{d}, \infty)$ and switching cost $\sigma_i = \underline{\sigma}$, for $i \in [1; N]$. We are interested in three problem setups: I with $(\underline{d}, \underline{\sigma}) = (0, 0)$, II with $(\underline{d}, \underline{\sigma}) = (0.1, 0)$,

Table 1: Summary of numerical results

Setup	Algorithm	Iter.	Time [s]	Obj. val.	Constr. viol.
I	`ALX`	11	14.27	1.5268	$5.1 \cdot 10^{-8}$
500 i.i.	`pdALX`	16	32.51	1.5001	$2.2 \cdot 10^{-7}$
I	`ALX`	10	15.98	1.8004	$4.0 \cdot 10^{-7}$
750 i.i.	`pdALX`	10	19.11	1.4895	$5.1 \cdot 10^{-7}$
II	`ALX`	10	12.83	1.7285	$8.9 \cdot 10^{-7}$
500 i.i.	`pdALX`	16	32.05	1.7210	$2.5 \cdot 10^{-7}$
II	`ALX`	10	16.94	1.9765	$5.5 \cdot 10^{-7}$
750 i.i.	`pdALX`	10	17.92	1.7115	$5.1 \cdot 10^{-7}$
III	`ALX`	12	19.85	4.8849	$7.0 \cdot 10^{-7}$
500 i.i.	`pdALX`	16	31.90	4.9001	$2.2 \cdot 10^{-7}$
III	`ALX`	11	28.31	4.8791	$5.4 \cdot 10^{-7}$
750 i.i.	`pdALX`	11	20.36	4.6903	$3.7 \cdot 10^{-7}$

and III with $(\underline{d}, \underline{\sigma}) = (0, 0.2)$. Notice that the non-negative parameter $\underline{d}$ corresponds to the dwell-time.

We adopt a background time grid with $n = 200$ points to integrate dynamics and sensitivities, with the explicit Euler method; we set tolerances $\eta^\star = 10^{-6}$ and $\epsilon^\star = 10^{-9}$. Subproblems are solved using FISTA [3], with a maximum number of (inner) iterations and estimating the Lipschitz constant via backtracking.

Table 1 summarizes the solution process and results for different problem setups and solver settings. Starting from a simple initial guess, solutions are found with relatively few (outer) iterations and within the feasibility tolerance. The maximum number of inner iterations (i.i. in Table 1) affects both the solution process and results, due to the different subproblems' solution, especially for `pdALX`. Concerning `pdALX`, allowing for more inner iterations leads to improved objective value, less outer iterations and less computation time; see Table 1. The corresponding state trajectories and switching intervals are depicted in Fig. 1–2. Recall that these may be associated to local minima, which are likely introduced by the switching time formulation [15]. Furthermore, as standard proximal methods are based on necessary criticality conditions, obtained solutions are optimal in a weak sense, especially for problems with discrete choices [2, 7].

7 Conclusions

Sparse switching times optimization, via direct multiple shooting, yields constrained composite problems. An original method for such problems has been introduced, embedding

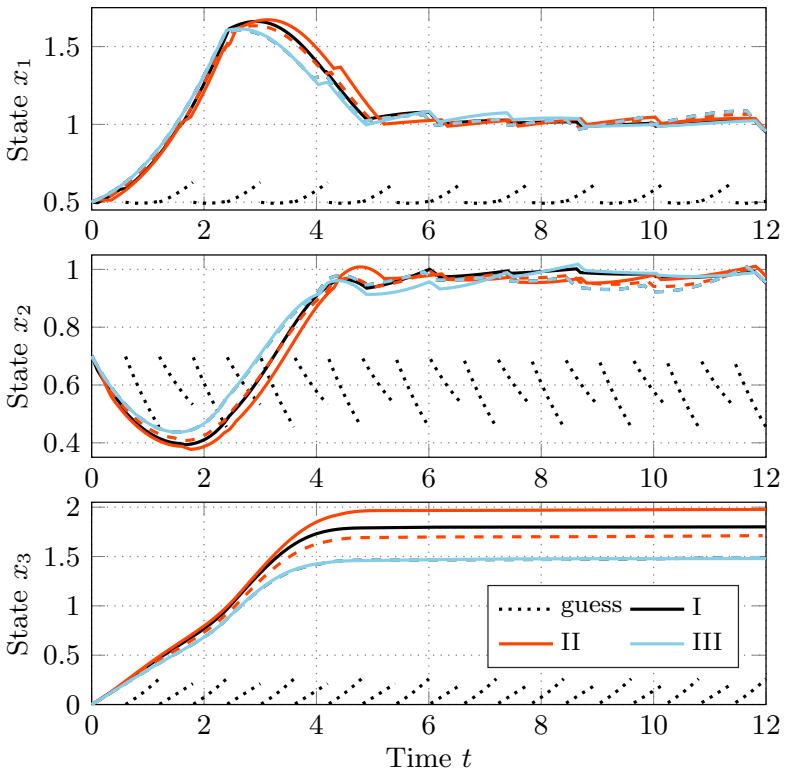

Figure 1: Differential states x_1, x_2, and x_3 versus time t, obtained for problems I, II, and III, and initial guess, adopting `ALX` (solid), `pdALX` (dashed), and maximum 750 inner iterations.

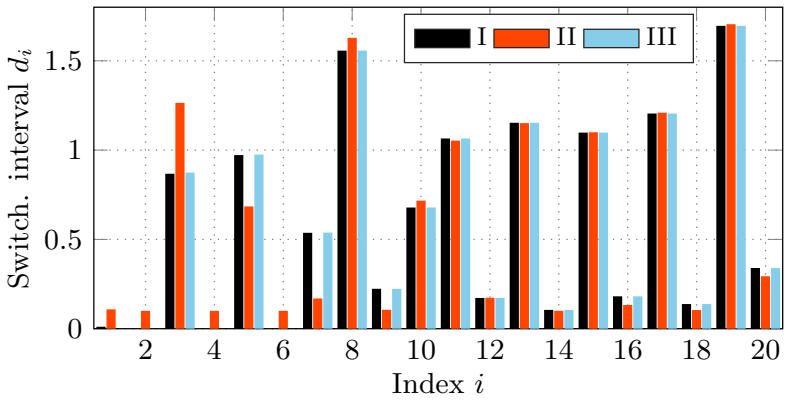

Figure 2: Switching intervals $\mathbf{d}$ obtained for problems I, II, and III, adopting `pdALX` and maximum 750 inner iterations.

proximal methods in the augmented Lagrangian framework. A numerical example indicated some challenges for future research: optimality concepts need to be sharpened and local minima to be escaped, structure exploitation and extensions to mixed-integer optimal control could be investigated.

Acknowledgment

The author thanks the anonymous reviewers, for their valuable comments, and Matthias Gerdts, for his acute remarks and ingenious questions.

References

[1] U. Ali and M. Egerstedt. Optimal control of switched dynamical systems under dwell time constraints. In *53rd IEEE Conference on Decision and Control (CDC)*, pages 4673–4678, 12 2014.

[2] Amir Beck and Nadav Hallak. Optimization problems involving group sparsity terms. *Mathematical Programming*, 4 2018.

[3] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[4] Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2016.

[5] Andrew R. Conn, Nicholas I. M. Gould, and Philippe L. Toint. A globally convergent augmented lagrangian algorithm for optimization with general constraints and simple bounds. *SIAM Journal on Numerical Analysis*, 28(2):545–572, 4 1991.

[6] Marie-Caroline Corbineau, Emilie Chouzenoux, and Jean-Christophe Pesquet. PIPA: A new proximal interior point algorithm for large-scale convex optimization. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1343–1347, Calgary, AB, 4 2018. IEEE.

[7] Alberto De Marchi. On the mixed-integer linear-quadratic optimal control with switching cost. *IEEE Control Systems Letters*, 3(4):990–995, 10 2019.

[8] Kathrin Flaßkamp, Todd Murphey, and Sina Ober-Blöbaum. Switching time optimization in discretized hybrid dynamical systems. In *51st IEEE Conference on Decision and Control (CDC)*, pages 707–712, 12 2012.

[9] Matthias Gerdts. *Optimal Control of ODEs and DAEs*. De Gruyter, 2011.

[10] Philip E. Gill and Daniel P. Robinson. A primal-dual augmented lagrangian. *Computational Optimization and Applications*, 51(1):1–25, 1 2012.

[11] Christian Kanzow and Daniel Steck. An example comparing the standard and safeguarded augmented lagrangian methods. *Operations Research Letters*, 45(6):598–603, 2017.

[12] Jason D. Lee, Yuekai Sun, and Michael A. Saunders. Proximal Newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014.

[13] Daniel P. Robinson. *Primal-Dual Methods for Nonlinear Optimization*. PhD thesis, University of California, San Diego, 9 2007.

[14] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317. Springer, Berlin, 2011.

[15] Sebastian Sager. *Numerical methods for mixed-integer optimal control problems*. PhD thesis, University of Heidelberg, 2005. Interdisciplinary Center for Scientific Computing.

[16] Lorenzo Stella, Andreas Themelis, Pantelis Sopasakis, and Panagiotis Patrinos. A simple and efficient algorithm for nonlinear model predictive control. In *56th IEEE Conference on Decision and Control (CDC)*, pages 1939–1944. IEEE, 2017.

[17] Bartolomeo Stellato, Sina Ober-Blöbaum, and Paul J. Goulart. Second-order switching time optimization for switched dynamical systems. *IEEE Transaction on Automatic Control*, 62(10):5407–5414, 2017.

[18] Andreas Themelis, Lorenzo Stella, and Panagiotis Patrinos. Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms. *SIAM Journal on Optimization*, 28(3):2274–2303, 2018.