

# Minimizing control variation in nonlinear optimal control<sup>★</sup>

Ryan Loxton<sup>a</sup>, Qun Lin<sup>a</sup>, Kok Lay Teo<sup>a</sup>

<sup>a</sup>*Department of Mathematics and Statistics, Curtin University, Australia*

---

## Abstract

In any real system, changing the control signal from one value to another will usually cause wear and tear on the system's actuators. Thus, when designing a control law, it is important to consider not just predicted system performance, but also the cost associated with changing the control action. This latter cost is almost always ignored in the optimal control literature. In this paper, we consider a class of optimal control problems in which the variation of the control signal is explicitly penalized in the cost function. We develop an effective computational method, based on the control parameterization approach and a novel transformation procedure, for solving this class of optimal control problems. We then apply our method to three example problems in fisheries, train control, and chemical engineering.

*Key words:* Optimal control computation, Constrained optimal control, Total variation, Nonlinear optimization

---

## 1 Introduction

The optimal control literature is replete with standard problems in which the system cost function is expressed solely in terms of the final state reached by the system and/or an integral term involving the state and control values at each time. The cost of changing the control signal is rarely considered. Thus, two different control laws that result in the same cost value are deemed to perform equally—even if one of them is constant while the other fluctuates wildly. Obviously, in engineering applications, the constant control law would be preferred.

Results pertaining to optimal control problems with costs on control signal changes are scarce in the literature. Blatt (1976) derived necessary optimality conditions for an optimal control problem in which the control signal can assume two possible values and there is a cost associated with changing from one value to another. Subsequently, Stewart (1992) developed a computational algorithm for solving Blatt's problem numerically, and Matula (1987) derived further optimality conditions for solving similar (but more general) problems in which the total variation of the control signal is incorporated as a penalty in the cost function.

---

<sup>★</sup> This paper was not presented at any IFAC meeting. Corresponding author Q. Lin. Tel. +61-8-92664962. Fax +61-8-92663197.

*Email addresses:* r.loxton@curtin.edu.au (Ryan Loxton), q.lin@curtin.edu.au (Qun Lin), k.l.teo@curtin.edu.au (Kok Lay Teo).

State constraints are not considered in any of the early references mentioned above. This is obviously a serious limitation, as most optimal control problems arising in real-world applications involve state constraints (see, for example, the spacecraft control problems in Açikmeşe and Blackmore (2011); the underwater vehicle problems in Büskens and Maurer (2000); and the zinc sulphate purification problems in Wang, Gui, Teo, Loxton, and Yang (2009)). In recent years, several algorithms for solving optimal control problems with state constraints have been proposed. These include the *non-smooth Newton's method* (Gerdtts (2008); Gerdtts and Kunkel (2008)), the *constraint transcription method* (Loxton, Teo, Rehbock, and Yiu (2009); Wang et al. (2009)), the *exact penalty method* (Jiang, Lin, Yu, Teo, and Duan (2012); Yu, Li, Loxton, and Teo (2013)), the *lossless convexification method* (Açikmeşe and Blackmore (2010); Açikmeşe and Blackmore (2011)), and a new global optimization approach based on interval analysis (Zhao and Stadtherr (2011)). None of these methods, however, consider the cost of changing the control signal.

One method that does take this cost into account is the control parameterization method introduced by Teo and Jennings (1991). This method, which is implemented in the MISER optimal control software (Jennings, Fisher, Teo, and Goh (2004)), is applicable to a general class of optimal control problems in which the cost function includes the total variation of the control signal and the state variables are subject to inequality path constraints.

Out of the few optimal control methods that explicitly

consider the cost of control changes, the control parameterization method by Teo and Jennings (1991) is applicable to the widest class of problems. However, one disadvantage of this method is that it approximate the control by a piecewise-constant function with *fixed* switching times. More modern discretization schemes are now available in which *both* the control values and the control switching times are decision variables to be chosen optimally—see, for example, Loxton, Teo, and Rehbock (2008); Loxton et al. (2009); Lin, Loxton, Teo, and Wu (2011); and Yu et al. (2013). These new discretization methods are known to produce results of much higher accuracy.

Another major limitation of Teo and Jennings’s method is that it is only guaranteed to converge when the path constraints are *pure state constraints* (i.e. constraints that do not depend on the control explicitly). The convergence results in Teo and Jennings (1991) are actually invalid if the path constraints include the control explicitly, which is the case in container crane control problems (Sakawa and Shindo (1982)) and soft-landing control problems for autonomous spacecraft (Açikmeşe and Blackmore (2011)).

The computational method in Teo and Jennings (1991) involves three levels of approximation: first the time horizon is partitioned into a set of subintervals and the control is approximated by a piecewise-constant function consistent with this partition; then the non-smooth total variation term in the cost function is approximated by a smooth function; then finally the path constraints are approximated by a set of conventional constraints. There are four adjustable parameters—one governing the number of subintervals, one governing the approximation of the total variation term, and two governing the approximation of the path constraints. These parameters must be updated appropriately to ensure convergence. Since each level of approximation reduces accuracy, having three levels is clearly not ideal.

The aim of this paper is to introduce a superior approach that only involves two levels of approximation. This new approach is applicable to general optimal control problems with the following characteristics:

- (i) The cost function contains a total variation term measuring changes in the control signal;
- (ii) The state and control variables are subject to path constraints;
- (iii) The state variables are subject to terminal equality constraints.

In our recent work (see Loxton, Lin, Rehbock, and Teo (2012)), we have introduced a discretization strategy for approximating problems of this type by a *non-smooth* optimization problem. We have already discussed the convergence properties of this discretization strategy, but we have yet to investigate the practically significant

issue of actually computing an optimal control law. The purpose of this current paper is to address this issue.

## 2 Problem formulation

Consider the following nonlinear control system:

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)), & t \in [0, T], & (1) \\ \mathbf{x}(0) &= \mathbf{x}^0, & & (2)\end{aligned}$$

where  $\mathbf{x}(t) \in \mathbb{R}^n$  is the system’s state at time  $t$ ,  $\mathbf{u}(t) \in \mathbb{R}^r$  is the control signal at time  $t$ ,  $\mathbf{x}^0 \in \mathbb{R}^n$  is the initial state,  $T > 0$  is a given terminal time, and  $\mathbf{f} : \mathbb{R}^n \times \mathbb{R}^r \rightarrow \mathbb{R}^n$  is a given continuously differentiable function.

Let  $u_i : [0, T] \rightarrow \mathbb{R}$  denote the  $i$ th component of the control signal  $\mathbf{u} : [0, T] \rightarrow \mathbb{R}^r$ . Then the *total variation* of  $u_i$  (see Royden and Fitzpatrick (2010)) is defined by

$$\bigvee_0^T u_i = \sup \sum_{j=1}^m |u_i(t_j) - u_i(t_{j-1})|,$$

where the supremum is taken over all finite partitions  $\{t_j\}_{j=0}^m \subset [0, T]$  satisfying

$$0 = t_0 < t_1 < \dots < t_{m-1} < t_m = T.$$

The total variation of the vector-valued control signal  $\mathbf{u} : [0, T] \rightarrow \mathbb{R}^r$  is defined by

$$\bigvee_0^T \mathbf{u} = \sum_{i=1}^r \bigvee_0^T u_i. \quad (3)$$

If the total variation of  $\mathbf{u} : [0, T] \rightarrow \mathbb{R}^r$  is finite, then we say that  $\mathbf{u}$  is of *bounded variation*. Let  $\mathcal{U}$  denote the class of all such functions of bounded variation mapping  $[0, T]$  into  $\mathbb{R}^r$ . Any  $\mathbf{u} \in \mathcal{U}$  is called an *admissible control* for system (1)-(2).

The total variation defined by (3) measures the extent to which the control signal changes during the time horizon—the more change, the larger the total variation. Note that if the control signal is constant, then the total variation is zero.

It is easy to show that for each admissible control  $\mathbf{u} \in \mathcal{U}$ , there exists a corresponding real number  $K := K(\mathbf{u}) > 0$  such that

$$|\mathbf{u}(t)| \leq K, \quad t \in [0, T], \quad (4)$$

where  $|\cdot|$  denotes the Euclidean norm. Thus, bounded variation implies boundedness.

We assume that  $\mathbf{f} : \mathbb{R}^n \times \mathbb{R}^r \rightarrow \mathbb{R}^n$  in system (1)-(2) satisfies the following linear growth condition: there exists

a real number  $L > 0$  such that

$$|\mathbf{f}(\boldsymbol{\xi}, \boldsymbol{\omega})| \leq L(1 + |\boldsymbol{\xi}| + |\boldsymbol{\omega}|), \quad (\boldsymbol{\xi}, \boldsymbol{\omega}) \in \mathbb{R}^n \times \mathbb{R}^r. \quad (5)$$

This assumption is standard in control theory (see, for example, Ahmed (1988, 2006) and Lin et al. (2011)). Inequalities (4) and (5) ensure that there exists a unique solution of (1)-(2) corresponding to each  $\mathbf{u} \in \mathcal{U}$  (see Theorem 3.3.3 of Ahmed (2006)). Let  $\mathbf{x}(\cdot|\mathbf{u})$  denote this solution.

Now, suppose that system (1)-(2) is subject to the following *terminal state constraints*:

$$\Phi_j(\mathbf{x}(T|\mathbf{u})) = 0, \quad j = 1, \dots, q_1, \quad (6)$$

where each  $\Phi_j : \mathbb{R}^n \rightarrow \mathbb{R}$  is a given continuously differentiable function.

Suppose also that system (1)-(2) is subject to *continuous inequality constraints* of the following form:

$$h_j(\mathbf{x}(t|\mathbf{u}), \mathbf{u}(t)) \leq 0, \quad t \in [0, T], \quad j = 1, \dots, q_2, \quad (7)$$

where each  $h_j : \mathbb{R}^n \times \mathbb{R}^r \rightarrow \mathbb{R}$  is a given continuously differentiable function. Note that control bounds can be easily incorporated into (7).

Let  $\mathcal{F}$  denote the set of all  $\mathbf{u} \in \mathcal{U}$  satisfying constraints (6) and (7). Controls in  $\mathcal{F}$  are called *feasible controls*. Our goal is to determine a feasible control that results in minimal system cost. In the classical optimal control literature, system cost is measured by a function of the final state reached by the system (a so-called *Mayer term*), and/or an integral term involving the state and control variables at each time (a so-called *Lagrange term*). The cost of changing the control signal is almost always neglected. Thus, there is nothing to discourage volatile control strategies that would be difficult—and potentially dangerous—to implement in practice.

In this paper, we propose a different approach in which the system cost function depends on both the final state reached by the system and the total variation of the control signal. Thus, large fluctuations in the control signal (which will likely result in excessive wear and tear on the system) are discouraged.

Our optimal control problem is defined formally below.

**Problem 1.** Choose a feasible control  $\mathbf{u} \in \mathcal{F}$  to minimize the cost functional

$$J_1(\mathbf{u}) = \Psi(\mathbf{x}(T|\mathbf{u})) + \alpha \bigvee_0^T \mathbf{u}, \quad (8)$$

where  $\alpha \geq 0$  is a given weight and  $\Psi : \mathbb{R}^n \rightarrow \mathbb{R}$  is a given continuously differentiable function.

Note that any cost function containing a Lagrange integral term can be easily transformed into the form of (8) by introducing a dummy state variable. See Chapter 6 of Ahmed (1988) for details.

### 3 Numerical discretization

Problem 1 presents two major difficulties:

- (i) The continuous inequality constraints (7) must be satisfied at an *infinite* number of time points (every point in  $[0, T]$ ).
- (ii) There is no closed-form analytical formula for the total variation term in (8).

Regarding the first difficulty, Hartl, Sethi, and Vickson (1995) survey the various theoretical results available for dealing with continuous inequality constraints in optimal control. Several effective computational methods have also been developed; see Gerdt and Kunkel (2008), Loxton et al. (2009), Wang et al. (2009), Açikmeşe and Blackmore (2011), Zhao and Stadtherr (2011), and Yu et al. (2013) for details. However, none of these computational methods are applicable to Problem 1 because of the non-standard total variation term in (8).

In this section, we will introduce a piecewise-constant approximation scheme for the control function in Problem 1. Under this approximation scheme, the total variation term is dramatically simplified and can be computed easily via a simple formula. We direct the reader to Loxton et al. (2012) for a thorough discussion of the convergence properties of this piecewise-constant approximation scheme. In this current paper, we are concerned with the practical issue of actually computing a control policy, rather than convergence issues.

Let  $p \geq 2$  be a given integer. Furthermore, let  $\mathcal{T}$  denote the set of all  $\boldsymbol{\tau} \in \mathbb{R}^{p-1}$  such that

$$0 = \tau_0 \leq \tau_1 \leq \tau_2 \leq \dots \leq \tau_{p-1} \leq \tau_p = T. \quad (9)$$

Elements of  $\mathcal{T}$  are called *switching time vectors*.

For given vectors  $\boldsymbol{\sigma}^k \in \mathbb{R}^r$ ,  $k = 1, \dots, p$ , define

$$(\boldsymbol{\sigma}^1, \dots, \boldsymbol{\sigma}^p) = [(\boldsymbol{\sigma}^1)^\top, \dots, (\boldsymbol{\sigma}^p)^\top]^\top \in \mathbb{R}^{pr}. \quad (10)$$

Let  $\mathcal{S}$  denote the set of all such  $\boldsymbol{\sigma} = (\boldsymbol{\sigma}^1, \dots, \boldsymbol{\sigma}^p) \in \mathbb{R}^{pr}$ . Elements of  $\mathcal{S}$  are called *control value vectors*.

We approximate the control  $\mathbf{u}$  by the following piecewise-constant function:

$$\mathbf{u}^p(t|\boldsymbol{\tau}, \boldsymbol{\sigma}) = \sum_{k=1}^p \boldsymbol{\sigma}^k \chi_{[\tau_{k-1}, \tau_k)}(t), \quad t \in [0, T], \quad (11)$$

where  $\boldsymbol{\tau} \in \mathcal{T}$  is a given switching time vector,  $\boldsymbol{\sigma} \in \mathcal{S}$  is a given control value vector, and  $\chi_{[\tau_{k-1}, \tau_k]} : \mathbb{R} \rightarrow \mathbb{R}$  is the characteristic function defined by

$$\chi_{[\tau_{k-1}, \tau_k]}(t) = \begin{cases} 1, & \text{if } t \in [\tau_{k-1}, \tau_k), \\ 0, & \text{otherwise.} \end{cases}$$

Throughout this paper, we use the convention that  $[\tau_{k-1}, \tau_k) = \emptyset$  if  $\tau_{k-1} = \tau_k$ .

Notice that equation (11) does not specify a value for  $\mathbf{u}^p$  at  $t = T$ . By convention, we define

$$\mathbf{u}^p(T|\boldsymbol{\tau}, \boldsymbol{\sigma}) = \mathbf{u}^p(T^-|\boldsymbol{\tau}, \boldsymbol{\sigma}) = \lim_{t \rightarrow T^-} \mathbf{u}^p(t|\boldsymbol{\tau}, \boldsymbol{\sigma}).$$

Let  $\sigma_i^k$  denote the  $i$ th component of  $\boldsymbol{\sigma}^k$ . Then it can be shown that

$$\bigvee_0^T \mathbf{u}^p(\cdot|\boldsymbol{\tau}, \boldsymbol{\sigma}) \leq \sum_{i=1}^r \sum_{k=1}^{p-1} |\sigma_i^{k+1} - \sigma_i^k|. \quad (12)$$

Thus,  $\mathbf{u}^p$  is of bounded variation—i.e.  $\mathbf{u}^p$  is an admissible control for Problem 1.

At first glance, it may appear that inequality (12) should be an equality. However, this is only true if the knot points  $\tau_k$ ,  $k = 0, \dots, p$ , are *distinct*. If two or more of the knot points coincide, then the inequality in (12) may be strict. For example, suppose that

$$\begin{aligned} r = 1, \quad p = 3, \quad T = 5, \quad \tau_0 = 0, \quad \tau_1 = 2, \\ \tau_2 = 2, \quad \tau_3 = 5, \quad \sigma^1 = 0, \quad \sigma^2 = 2, \quad \sigma^3 = 1. \end{aligned}$$

Then, noting that  $\tau_1$  and  $\tau_2$  coincide at  $t = 2$ , we see that

$$\bigvee_0^5 \mathbf{u}^p = 1 < 3 = |\sigma^2 - \sigma^1| + |\sigma^3 - \sigma^2|.$$

Thus, inequality (12) is strict in this case. A formal proof of inequality (12) is given in Loxton et al. (2012).

Substituting the piecewise-constant control (11) into the dynamic system (1)-(2) yields

$$\dot{\mathbf{x}}(t) = \sum_{k=1}^p \mathbf{f}(\mathbf{x}(t), \boldsymbol{\sigma}^k) \chi_{[\tau_{k-1}, \tau_k)}(t), \quad t \in [0, T], \quad (13)$$

$$\mathbf{x}(0) = \mathbf{x}^0. \quad (14)$$

Let  $\mathbf{x}^p(\cdot|\boldsymbol{\tau}, \boldsymbol{\sigma})$  denote the solution of (13)-(14) corresponding to  $\boldsymbol{\tau} \in \mathcal{T}$  and  $\boldsymbol{\sigma} \in \mathcal{S}$ . Then clearly,

$$\mathbf{x}^p(t|\boldsymbol{\tau}, \boldsymbol{\sigma}) = \mathbf{x}(t|\mathbf{u}^p(\cdot|\boldsymbol{\tau}, \boldsymbol{\sigma})), \quad t \in [0, T]. \quad (15)$$

Constraints (6) and (7) become

$$\Phi_j(\mathbf{x}^p(T|\boldsymbol{\tau}, \boldsymbol{\sigma})) = 0, \quad j = 1, \dots, q_1, \quad (16)$$

and

$$\sum_{k=1}^p h_j(\mathbf{x}^p(t|\boldsymbol{\tau}, \boldsymbol{\sigma}), \boldsymbol{\sigma}^k) \chi_{[\tau_{k-1}, \tau_k)}(t) \leq 0, \quad (17)$$

$$t \in [0, T], \quad j = 1, \dots, q_2.$$

Let  $\Gamma$  denote the set of all pairs  $(\boldsymbol{\tau}, \boldsymbol{\sigma}) \in \mathcal{T} \times \mathcal{S}$  such that (16) and (17) are satisfied. Then

$$(\boldsymbol{\tau}, \boldsymbol{\sigma}) \in \Gamma \iff \mathbf{u}^p(\cdot|\boldsymbol{\tau}, \boldsymbol{\sigma}) \in \mathcal{F}.$$

Define

$$J_2(\boldsymbol{\tau}, \boldsymbol{\sigma}) = \Psi(\mathbf{x}^p(T|\boldsymbol{\tau}, \boldsymbol{\sigma})) + \alpha \sum_{i=1}^r \sum_{k=1}^{p-1} |\sigma_i^{k+1} - \sigma_i^k|. \quad (18)$$

Then it follows from (12) and (15) that

$$J_1(\mathbf{u}^p) = \Psi(\mathbf{x}(T|\mathbf{u}^p)) + \alpha \bigvee_0^T \mathbf{u}^p \leq J_2(\boldsymbol{\tau}, \boldsymbol{\sigma}). \quad (19)$$

By following the arguments in Loxton et al. (2012), it can be shown that (19) holds with equality whenever  $(\boldsymbol{\tau}, \boldsymbol{\sigma})$  is a minimizer of  $J_2$ . On this basis, Problem 1 can be approximated by the following finite-dimensional optimization problem.

**Problem 2.** Choose a pair  $(\boldsymbol{\tau}, \boldsymbol{\sigma}) \in \Gamma$  to minimize the cost function  $J_2$ .

Notice that the total variation term in  $J_1$  has been transformed into a non-smooth absolute value term in  $J_2$ . This new term can be computed easily.

Let  $(\boldsymbol{\tau}^*, \boldsymbol{\sigma}^*) \in \Gamma$  be an optimal solution of Problem 2, where

$$\boldsymbol{\tau}^* = [\tau_1^*, \dots, \tau_{p-1}^*]^\top \in \mathcal{T}$$

and

$$\boldsymbol{\sigma}^* = (\boldsymbol{\sigma}^{1,*}, \dots, \boldsymbol{\sigma}^{p,*}) \in \mathcal{S}.$$

Then the corresponding piecewise-constant control for Problem 1 is defined as follows:

$$\mathbf{u}^{p,*}(t) = \begin{cases} \sum_{k=1}^p \boldsymbol{\sigma}^{k,*} \chi_{[\tau_{k-1}^*, \tau_k^*)}(t), & \text{if } t \in [0, T), \\ \mathbf{u}^{p,*}(T^-), & \text{if } t = T, \end{cases} \quad (20)$$

where  $\tau_0^* = 0$  and  $\tau_p^* = T$ . It can be shown that  $\mathbf{u}^{p,*}$  minimizes  $J_1$  over the space of feasible piecewise-constant controls (Loxton et al. (2012)). Thus, a suboptimal control for Problem 1 can be obtained by solving Problem 2.

Unfortunately, standard gradient-based optimization algorithms cannot be applied to Problem 2 because the cost function (18) is non-smooth. Furthermore, it is well known that variable switching times cause major computational difficulties for gradient-based optimization algorithms (Xu and Antsaklis (2004); Loxton et al. (2008); Loxton et al. (2009)). In the next section, we will present a novel transformation procedure for converting Problem 2 into a smooth optimization problem with fixed switching times. This new problem can be solved readily using existing optimization techniques.

## 4 Problem transformation

### 4.1 A new optimization problem

Let  $\mathcal{O}$  denote the set of all  $\boldsymbol{\theta} \in \mathbb{R}^p$  such that

$$\theta_k \geq 0, \quad k = 1, \dots, p,$$

and

$$\sum_{k=1}^p \theta_k = T.$$

Define

$$\boldsymbol{\zeta} = (\boldsymbol{\gamma}, \mathbf{v}^1, \dots, \mathbf{v}^{p-1}, \mathbf{w}^1, \dots, \mathbf{w}^{p-1}) \in \mathbb{R}^{(2p-1)r}, \quad (21)$$

where  $\boldsymbol{\gamma} \in \mathbb{R}^r$ ,  $\mathbf{v}^k \in \mathbb{R}^r$ ,  $\mathbf{w}^k \in \mathbb{R}^r$ , and the round bracket notation has the same meaning as in (10). Let  $\mathcal{Z}$  denote the set of all  $\boldsymbol{\zeta} \in \mathbb{R}^{(2p-1)r}$  in the form of (21) such that  $\mathbf{v}^k$  and  $\mathbf{w}^k$  satisfy the following non-negativity constraints:

$$v_i^k \geq 0, \quad w_i^k \geq 0, \quad i = 1, \dots, r, \quad k = 1, \dots, p-1.$$

Now, define functions  $\boldsymbol{\psi}^k : \mathcal{Z} \rightarrow \mathbb{R}^r$ ,  $k = 1, \dots, p$ , as follows:

$$\boldsymbol{\psi}^k(\boldsymbol{\zeta}) = \boldsymbol{\gamma} + \sum_{l=k}^{p-1} (\mathbf{v}^l - \mathbf{w}^l), \quad k = 1, \dots, p.$$

We consider the following dynamic system on the new time horizon  $[0, p]$ :

$$\dot{\mathbf{y}}(s) = \sum_{k=1}^p \theta_k \mathbf{f}(\mathbf{y}(s), \boldsymbol{\psi}^k(\boldsymbol{\zeta})) \chi_{[k-1, k)}(s), \quad (22)$$

$$\mathbf{y}(0) = \mathbf{x}^0, \quad (23)$$

where  $(\boldsymbol{\theta}, \boldsymbol{\zeta}) \in \mathcal{O} \times \mathcal{Z}$  is a given pair,  $\mathbf{x}^0 \in \mathbb{R}^n$  and  $\mathbf{f} : \mathbb{R}^n \times \mathbb{R}^r \rightarrow \mathbb{R}^n$  are as defined in Section 2, and  $\chi_{[k-1, k)} : \mathbb{R} \rightarrow \mathbb{R}$  is the characteristic function defined in Section 3.

Let  $\mathbf{y}(\cdot | \boldsymbol{\theta}, \boldsymbol{\zeta})$  denote the solution of (22)-(23) corresponding to  $(\boldsymbol{\theta}, \boldsymbol{\zeta}) \in \mathcal{O} \times \mathcal{Z}$ . Furthermore, let  $\Xi$  denote the set

of all  $(\boldsymbol{\theta}, \boldsymbol{\zeta}) \in \mathcal{O} \times \mathcal{Z}$  satisfying the following constraints:

$$\Phi_j(\mathbf{y}(p | \boldsymbol{\theta}, \boldsymbol{\zeta})) = 0, \quad j = 1, \dots, q_1,$$

and

$$\sum_{k=1}^p \theta_k h_j(\mathbf{y}(s | \boldsymbol{\theta}, \boldsymbol{\zeta}), \boldsymbol{\psi}^k(\boldsymbol{\zeta})) \chi_{[k-1, k)}(s) \leq 0, \\ s \in [0, p], \quad j = 1, \dots, q_2.$$

We now define a new optimization problem as follows.

**Problem 3.** Choose  $(\boldsymbol{\theta}, \boldsymbol{\zeta}) \in \Xi$  to minimize the cost function

$$J_3(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \Psi(\mathbf{y}(p | \boldsymbol{\theta}, \boldsymbol{\zeta})) + \alpha \sum_{i=1}^r \sum_{k=1}^{p-1} (v_i^k + w_i^k).$$

We will show in the next subsection that Problem 3 is equivalent to Problem 2. This means that a solution of Problem 3 can be used to generate a solution of Problem 2, and vice versa. Note that Problem 3 has a *smooth cost function* and *fixed switching times* at the integers  $s = 1, \dots, p-1$ . Thus, from a computational point of view, Problem 3 is much easier than Problem 2.

### 4.2 Main results

Our aim in this subsection is to prove that Problems 2 and 3 are equivalent.

First, for each  $\boldsymbol{\theta} \in \mathcal{O}$ , define a corresponding *time-scaling function* as follows:

$$\mu(s | \boldsymbol{\theta}) = \begin{cases} \sum_{l=1}^{\lfloor s \rfloor} \theta_l + \theta_{\lfloor s \rfloor + 1} (s - \lfloor s \rfloor), & \text{if } s \in [0, p), \\ T, & \text{if } s = p. \end{cases}$$

Clearly,

$$\mu(k | \boldsymbol{\theta}) = \sum_{l=1}^k \theta_l, \quad k = 0, \dots, p.$$

For each  $\boldsymbol{\theta} \in \mathcal{O}$ , define

$$\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}) = [\mu(1 | \boldsymbol{\theta}), \dots, \mu(p-1 | \boldsymbol{\theta})]^\top \in \mathbb{R}^{p-1}.$$

Since each component of  $\boldsymbol{\theta} \in \mathcal{O}$  is non-negative,

$$0 \leq \mu(k | \boldsymbol{\theta}) \leq \sum_{l=1}^p \theta_l = T, \quad k = 1, \dots, p-1, \quad (24)$$

and

$$\mu(k-1|\boldsymbol{\theta}) \leq \mu(k|\boldsymbol{\theta}), \quad k = 2, \dots, p-1. \quad (25)$$

It follows from (24) and (25) that  $\tilde{\tau}(\boldsymbol{\theta})$  satisfies (9). Thus,  $\tilde{\tau}(\boldsymbol{\theta}) \in \mathcal{T}$  is a valid switching time vector for Problem 2.

Now, for each  $\boldsymbol{\zeta} \in \mathcal{Z}$  defined by (21), let

$$\tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta}) = (\boldsymbol{\psi}^1(\boldsymbol{\zeta}), \dots, \boldsymbol{\psi}^p(\boldsymbol{\zeta})) \in \mathbb{R}^{pr}.$$

We immediately see that  $\tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta}) \in \mathcal{S}$ . Thus, each pair in  $\mathcal{O} \times \mathcal{Z}$  generates a corresponding pair in  $\mathcal{T} \times \mathcal{S}$  through the relation  $(\boldsymbol{\theta}, \boldsymbol{\zeta}) \mapsto (\tilde{\tau}(\boldsymbol{\theta}), \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta}))$ . Solving the dynamic system (13)-(14) with  $\boldsymbol{\tau} = \tilde{\tau}(\boldsymbol{\theta})$  and  $\boldsymbol{\sigma} = \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta})$  yields the state trajectory  $\mathbf{x}^p(\cdot|\tilde{\tau}(\boldsymbol{\theta}), \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta}))$ . Our next result reveals the relationship between  $\mathbf{x}^p(\cdot|\tilde{\tau}(\boldsymbol{\theta}), \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta}))$  and the solution of the new system (22)-(23).

**Theorem 4.1.** For each  $(\boldsymbol{\theta}, \boldsymbol{\zeta}) \in \mathcal{O} \times \mathcal{Z}$ ,

$$\mathbf{y}(s|\boldsymbol{\theta}, \boldsymbol{\zeta}) = \mathbf{x}^p(t|\tilde{\tau}(\boldsymbol{\theta}), \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta})) \Big|_{t=\mu(s|\boldsymbol{\theta})}. \quad (26)$$

*Proof.* For notational simplicity, let

$$\tilde{\mathbf{x}}(s) = \mathbf{x}^p(t|\tilde{\tau}(\boldsymbol{\theta}), \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta})) \Big|_{t=\mu(s|\boldsymbol{\theta})}.$$

We will prove (26) by showing that  $\tilde{\mathbf{x}}$  is the unique solution of (22)-(23).

Since  $\mu(0|\boldsymbol{\theta}) = 0$  and  $\mathbf{x}^p(0|\tilde{\tau}(\boldsymbol{\theta}), \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta})) = \mathbf{x}^0$ ,

$$\begin{aligned} \tilde{\mathbf{x}}(0) &= \mathbf{x}^p(t|\tilde{\tau}(\boldsymbol{\theta}), \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta})) \Big|_{t=\mu(0|\boldsymbol{\theta})} \\ &= \mathbf{x}^p(0|\tilde{\tau}(\boldsymbol{\theta}), \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta})) \\ &= \mathbf{x}^0. \end{aligned}$$

Hence,  $\tilde{\mathbf{x}}$  satisfies the initial condition (23).

Let  $\mathcal{E}'$  denote the set of all  $k \in \{1, \dots, p\}$  such that  $\theta_k = 0$ , and let  $\mathcal{E}'' = \{1, \dots, p\} \setminus \mathcal{E}'$ . If  $k \in \mathcal{E}'$ , then  $\mu(s|\boldsymbol{\theta}) = \mu(k-1|\boldsymbol{\theta})$  for all  $s \in (k-1, k)$ . Thus,

$$\begin{aligned} \tilde{\mathbf{x}}(s) &= \mathbf{x}^p(t|\tilde{\tau}(\boldsymbol{\theta}), \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta})) \Big|_{t=\mu(s|\boldsymbol{\theta})} \\ &= \mathbf{x}^p(t|\tilde{\tau}(\boldsymbol{\theta}), \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta})) \Big|_{t=\mu(k-1|\boldsymbol{\theta})} \\ &= \tilde{\mathbf{x}}(k-1). \end{aligned}$$

This implies that for each  $k \in \mathcal{E}'$ ,

$$\dot{\tilde{\mathbf{x}}}(s) = \mathbf{0} = \theta_k \mathbf{f}(\tilde{\mathbf{x}}(s), \boldsymbol{\psi}^k(\boldsymbol{\zeta})), \quad s \in (k-1, k). \quad (27)$$

Now, if  $k \in \mathcal{E}''$ , then  $\mu(k-1|\boldsymbol{\theta}) < \mu(s|\boldsymbol{\theta}) < \mu(k|\boldsymbol{\theta})$  and  $\dot{\mu}(s|\boldsymbol{\theta}) = \theta_k$  for all  $s \in (k-1, k)$ . Thus, by differentiating

$\tilde{\mathbf{x}}$  using the chain rule and then applying equation (13) with  $\boldsymbol{\tau} = \tilde{\tau}(\boldsymbol{\theta})$  and  $\boldsymbol{\sigma} = \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta})$ , we obtain

$$\dot{\tilde{\mathbf{x}}}(s) = \theta_k \mathbf{f}(\tilde{\mathbf{x}}(s), \boldsymbol{\psi}^k(\boldsymbol{\zeta})), \quad s \in (k-1, k). \quad (28)$$

Equations (27) and (28) imply that  $\tilde{\mathbf{x}}$  satisfies the following differential equation almost everywhere on  $[0, p]$ :

$$\dot{\tilde{\mathbf{x}}}(s) = \sum_{k=1}^p \theta_k \mathbf{f}(\tilde{\mathbf{x}}(s), \boldsymbol{\psi}^k(\boldsymbol{\zeta})) \chi_{[k-1, k)}(s), \quad s \in [0, p].$$

It follows that  $\tilde{\mathbf{x}}$  is the unique solution of (22)-(23), as required.  $\square$

Theorem 4.1 shows how the time-scaling function  $\mu(\cdot|\boldsymbol{\theta})$  links the dynamic system in Problem 2 with the dynamic system in Problem 3. The next result links the feasible regions of these two problems.

**Theorem 4.2.** Let  $(\boldsymbol{\theta}, \boldsymbol{\zeta}) \in \mathcal{O} \times \mathcal{Z}$  be a given pair. Then  $(\boldsymbol{\theta}, \boldsymbol{\zeta}) \in \Xi$  if and only if  $(\tilde{\tau}(\boldsymbol{\theta}), \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta})) \in \Gamma$ . That is,  $(\boldsymbol{\theta}, \boldsymbol{\zeta})$  is feasible for Problem 3 if and only if  $(\tilde{\tau}(\boldsymbol{\theta}), \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta}))$  is feasible for Problem 2.

*Proof.* For notational simplicity, we write  $\mathbf{x}^p(\cdot)$  instead of  $\mathbf{x}^p(\cdot|\tilde{\tau}(\boldsymbol{\theta}), \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta}))$ , and  $\mathbf{y}(\cdot)$  instead of  $\mathbf{y}(\cdot|\boldsymbol{\theta}, \boldsymbol{\zeta})$ . Since  $(\boldsymbol{\theta}, \boldsymbol{\zeta}) \in \mathcal{O} \times \mathcal{Z}$  is fixed, this notation will not cause confusion.

Suppose that  $(\boldsymbol{\theta}, \boldsymbol{\zeta})$  is feasible for Problem 3. Then

$$\Phi_j(\mathbf{y}(p)) = 0, \quad j = 1, \dots, q_1, \quad (29)$$

and

$$\begin{aligned} \theta_k h_j(\mathbf{y}(s), \boldsymbol{\psi}^k(\boldsymbol{\zeta})) &\leq 0, \quad s \in [k-1, k), \\ k &= 1, \dots, p, \quad j = 1, \dots, q_2. \end{aligned} \quad (30)$$

Recall that  $\mu(p|\boldsymbol{\theta}) = T$ . Thus, it follows from (29) and Theorem 4.1 that for each  $j = 1, \dots, q_1$ ,

$$\Phi_j(\mathbf{x}^p(T)) = \Phi_j(\mathbf{x}^p(t)) \Big|_{t=\mu(p|\boldsymbol{\theta})} = \Phi_j(\mathbf{y}(p)) = 0.$$

This shows that  $(\tilde{\tau}(\boldsymbol{\theta}), \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta}))$  satisfies the terminal constraints (16) in Problem 2.

Now, let  $t' \in [0, T]$ . If  $t' = T$ , then the continuous inequality constraints (17) are automatically satisfied. Thus, we assume that  $t' \in [0, T)$ . Then we can find an integer  $\varsigma$  such that  $\theta_\varsigma > 0$  and  $\mu(\varsigma-1|\boldsymbol{\theta}) \leq t' < \mu(\varsigma|\boldsymbol{\theta})$ . Thus, by considering (30) for  $k = \varsigma$  and dividing both sides by  $\theta_\varsigma$ , we obtain, for each  $j = 1, \dots, q_2$ ,

$$h_j(\mathbf{y}(s), \boldsymbol{\psi}^\varsigma(\boldsymbol{\zeta})) \leq 0, \quad s \in [\varsigma-1, \varsigma). \quad (31)$$

The time-scaling function  $\mu(\cdot|\boldsymbol{\theta})$  is clearly a non-decreasing surjection from  $[0, p]$  to  $[0, T]$ . Hence, there exists a  $s' \in [\zeta - 1, \zeta)$  such that  $\mu(s'|\boldsymbol{\theta}) = t'$ . By Theorem 4.1 and inequality (31), for each  $j = 1, \dots, q_2$ ,

$$\begin{aligned} h_j(\mathbf{x}^p(t'), \boldsymbol{\psi}^\zeta(\boldsymbol{\zeta})) &= h_j(\mathbf{x}^p(t), \boldsymbol{\psi}^\zeta(\boldsymbol{\zeta})) \Big|_{t=\mu(s'|\boldsymbol{\theta})} \\ &= h_j(\mathbf{y}(s'), \boldsymbol{\psi}^\zeta(\boldsymbol{\zeta})) \leq 0. \end{aligned}$$

Thus, for each  $j = 1, \dots, q_2$ ,

$$\begin{aligned} \sum_{k=1}^p h_j(\mathbf{x}^p(t'), \boldsymbol{\psi}^k(\boldsymbol{\zeta})) \chi_{[\tilde{\tau}_{k-1}(\boldsymbol{\theta}), \tilde{\tau}_k(\boldsymbol{\theta})]}(t') \\ = h_j(\mathbf{x}^p(t'), \boldsymbol{\psi}^\zeta(\boldsymbol{\zeta})) \leq 0, \end{aligned}$$

where  $\tilde{\tau}_k(\boldsymbol{\theta}) = \mu(k|\boldsymbol{\theta})$ . Since  $t' \in [0, T)$  was chosen arbitrarily, this shows that  $(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}), \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta}))$  satisfies the continuous inequality constraints (17) in Problem 2. It follows that  $(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}), \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta}))$  is feasible for Problem 2. The reverse implication is proved in a similar manner.  $\square$

We now characterize the optimal solution of Problem 3.

**Theorem 4.3.** Let  $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*) \in \Xi$  be an optimal solution of Problem 3, where

$$\boldsymbol{\zeta}^* = (\boldsymbol{\gamma}^*, \mathbf{v}^{1,*}, \dots, \mathbf{v}^{p-1,*}, \mathbf{w}^{1,*}, \dots, \mathbf{w}^{p-1,*}). \quad (32)$$

Then

$$v_i^{k,*} w_i^{k,*} = 0, \quad i = 1, \dots, r, \quad k = 1, \dots, p-1. \quad (33)$$

*Proof.* Suppose that (33) is violated for some  $i$  and  $k$ . Let  $\mathcal{J}_1$  denote the set of index pairs  $(i, k)$  such that  $v_i^{k,*} w_i^{k,*} > 0$ , and let  $\mathcal{J}_2$  denote the set of index pairs  $(i, k)$  such that  $v_i^{k,*} w_i^{k,*} = 0$ . Since  $v_i^{k,*}$  and  $w_i^{k,*}$  are non-negative,  $\mathcal{J}_1 \cup \mathcal{J}_2 = \{1, \dots, r\} \times \{1, \dots, p-1\}$ .

Define

$$\hat{v}_i^k = \begin{cases} \max\{v_i^{k,*} - w_i^{k,*}, 0\}, & \text{if } (i, k) \in \mathcal{J}_1, \\ v_i^{k,*}, & \text{if } (i, k) \in \mathcal{J}_2, \end{cases}$$

and

$$\hat{w}_i^k = \begin{cases} \max\{w_i^{k,*} - v_i^{k,*}, 0\}, & \text{if } (i, k) \in \mathcal{J}_1, \\ w_i^{k,*}, & \text{if } (i, k) \in \mathcal{J}_2. \end{cases}$$

Furthermore, define

$$\hat{\boldsymbol{\zeta}} = (\boldsymbol{\gamma}^*, \hat{\mathbf{v}}^1, \dots, \hat{\mathbf{v}}^{p-1}, \hat{\mathbf{w}}^1, \dots, \hat{\mathbf{w}}^{p-1}),$$

where  $\hat{\mathbf{v}}^k = [\hat{v}_1^k, \dots, \hat{v}_r^k]^\top$  and  $\hat{\mathbf{w}}^k = [\hat{w}_1^k, \dots, \hat{w}_r^k]^\top$ . We immediately see that  $\hat{\boldsymbol{\zeta}} \in \mathcal{Z}$ . Furthermore,

$$\hat{v}_i^k - \hat{w}_i^k = v_i^{k,*} - w_i^{k,*}, \quad i = 1, \dots, r, \quad k = 1, \dots, p-1.$$

Hence, for each  $k = 1, \dots, p$ ,

$$\begin{aligned} \boldsymbol{\psi}^k(\hat{\boldsymbol{\zeta}}) &= \boldsymbol{\gamma}^* + \sum_{l=k}^{p-1} (\hat{\mathbf{v}}^l - \hat{\mathbf{w}}^l) \\ &= \boldsymbol{\gamma}^* + \sum_{l=k}^{p-1} (\mathbf{v}^{l,*} - \mathbf{w}^{l,*}) = \boldsymbol{\psi}^k(\boldsymbol{\zeta}^*). \end{aligned}$$

This implies that  $\mathbf{y}(s|\boldsymbol{\theta}^*, \hat{\boldsymbol{\zeta}}) = \mathbf{y}(s|\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*)$  for all  $s \in [0, p]$ . Thus, since  $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*)$  is feasible for Problem 3,  $(\boldsymbol{\theta}^*, \hat{\boldsymbol{\zeta}})$  is also feasible for Problem 3. Furthermore,

$$\begin{aligned} J_3(\boldsymbol{\theta}^*, \hat{\boldsymbol{\zeta}}) &= \Psi(\mathbf{y}(p|\boldsymbol{\theta}^*, \hat{\boldsymbol{\zeta}})) + \alpha \sum_{i=1}^r \sum_{k=1}^{p-1} (\hat{v}_i^k + \hat{w}_i^k) \\ &= \Psi(\mathbf{y}(p|\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*)) + \alpha \sum_{i=1}^r \sum_{k=1}^{p-1} (v_i^{k,*} + w_i^{k,*}). \quad (34) \end{aligned}$$

Now, recall that  $v_i^{k,*}$  and  $w_i^{k,*}$  are both non-negative. Hence, if  $(i, k) \in \mathcal{J}_1$ , then  $v_i^{k,*} > 0$  and  $w_i^{k,*} > 0$ . This implies that

$$\begin{aligned} \hat{v}_i^k + \hat{w}_i^k &= \max\{v_i^{k,*} - w_i^{k,*}, 0\} + \max\{w_i^{k,*} - v_i^{k,*}, 0\} \\ &= |v_i^{k,*} - w_i^{k,*}| < v_i^{k,*} + w_i^{k,*}. \end{aligned}$$

Consequently, we have the following implication:

$$(i, k) \in \mathcal{J}_1 \implies \hat{v}_i^k + \hat{w}_i^k < v_i^{k,*} + w_i^{k,*}.$$

Thus, from (34) and our assumption that  $\mathcal{J}_1 \neq \emptyset$ ,

$$\begin{aligned} J_3(\boldsymbol{\theta}^*, \hat{\boldsymbol{\zeta}}) &< \Psi(\mathbf{y}(p|\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*)) + \alpha \sum_{i=1}^r \sum_{k=1}^{p-1} (v_i^{k,*} + w_i^{k,*}) \\ &= J_3(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*). \end{aligned}$$

But since  $(\boldsymbol{\theta}^*, \hat{\boldsymbol{\zeta}})$  is feasible for Problem 3, this contradicts the optimality of  $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*)$ . Thus, our assumption that  $\mathcal{J}_1 \neq \emptyset$  is false. It follows that equation (33) must hold for all  $i$  and  $k$ .  $\square$

We are now ready to show that a solution of Problem 3 can be used to generate a solution of Problem 2.

**Theorem 4.4.** Let  $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*) \in \Xi$  be an optimal solution of Problem 3, where  $\boldsymbol{\zeta}^*$  is as defined in equation (32). Then  $(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}^*), \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta}^*))$  is an optimal solution of Problem 2.

*Proof.* By Theorem 4.1,

$$\begin{aligned} \mathbf{x}^p(T|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}^*), \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta}^*)) &= \mathbf{x}^p(t|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}^*), \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta}^*)) \Big|_{t=\mu(p|\boldsymbol{\theta}^*)} \\ &= \mathbf{y}(p|\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*). \end{aligned} \quad (35)$$

Theorem 4.3 implies that for each index pair  $(i, k)$ , either  $v_i^{k,*} = 0$  or  $w_i^{k,*} = 0$ . If  $v_i^{k,*} = 0$ , then since  $w_i^{k,*} \geq 0$ ,

$$|w_i^{k,*} - v_i^{k,*}| = |w_i^{k,*}| = w_i^{k,*} = v_i^{k,*} + w_i^{k,*}.$$

Similarly, if  $w_i^{k,*} = 0$ , then

$$|w_i^{k,*} - v_i^{k,*}| = |-v_i^{k,*}| = v_i^{k,*} = v_i^{k,*} + w_i^{k,*}.$$

Thus, for each  $i = 1, \dots, r$  and  $k = 1, \dots, p-1$ ,

$$|\psi_i^{k+1}(\boldsymbol{\zeta}^*) - \psi_i^k(\boldsymbol{\zeta}^*)| = |w_i^{k,*} - v_i^{k,*}| = v_i^{k,*} + w_i^{k,*}, \quad (36)$$

where  $\psi_i^k(\boldsymbol{\zeta}^*)$  denotes the  $i$ th component of  $\boldsymbol{\psi}^k(\boldsymbol{\zeta}^*)$ . From (35) and (36),

$$\begin{aligned} J_2(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}^*), \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta}^*)) &= \Psi(\mathbf{x}^p(T|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}^*), \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta}^*))) \\ &\quad + \alpha \sum_{i=1}^r \sum_{k=1}^{p-1} |\psi_i^{k+1}(\boldsymbol{\zeta}^*) - \psi_i^k(\boldsymbol{\zeta}^*)| \\ &= \Psi(\mathbf{y}(p|\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*)) + \alpha \sum_{i=1}^r \sum_{k=1}^{p-1} (v_i^{k,*} + w_i^{k,*}). \end{aligned} \quad (37)$$

Now, let  $(\bar{\boldsymbol{\tau}}, \bar{\boldsymbol{\sigma}}) \in \Gamma$  be an arbitrary feasible pair for Problem 2, where  $\bar{\boldsymbol{\tau}} = [\bar{\tau}_1, \dots, \bar{\tau}_{p-1}]^\top$  and

$$\bar{\boldsymbol{\sigma}} = (\bar{\sigma}^1, \dots, \bar{\sigma}^p).$$

Define  $\bar{\boldsymbol{\theta}} = [\bar{\theta}_1, \dots, \bar{\theta}_p]^\top \in \mathbb{R}^p$  as follows:

$$\bar{\theta}_k = \bar{\tau}_k - \bar{\tau}_{k-1}, \quad k = 1, \dots, p,$$

where  $\bar{\tau}_0 = 0$  and  $\bar{\tau}_p = T$ . Then clearly,  $\bar{\boldsymbol{\theta}} \in \mathcal{O}$  and  $\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}) = \bar{\boldsymbol{\tau}}$ . For each  $k$ , define  $\bar{\mathbf{v}}^k = [\bar{v}_1^k, \dots, \bar{v}_r^k]^\top$  and  $\bar{\mathbf{w}}^k = [\bar{w}_1^k, \dots, \bar{w}_r^k]^\top$  as follows:

$$\begin{aligned} \bar{v}_i^k &= \max\{\bar{\sigma}_i^k - \bar{\sigma}_i^{k+1}, 0\}, \quad i = 1, \dots, r, \\ \bar{w}_i^k &= \max\{\bar{\sigma}_i^{k+1} - \bar{\sigma}_i^k, 0\}, \quad i = 1, \dots, r. \end{aligned}$$

Furthermore, define

$$\bar{\boldsymbol{\zeta}} = (\bar{\boldsymbol{\sigma}}^p, \bar{\mathbf{v}}^1, \dots, \bar{\mathbf{v}}^{p-1}, \bar{\mathbf{w}}^1, \dots, \bar{\mathbf{w}}^{p-1}) \in \mathbb{R}^{(2p-1)r}.$$

Then  $\bar{\boldsymbol{\zeta}} \in \mathcal{Z}$ . For each  $i = 1, \dots, r$  and  $k = 1, \dots, p-1$ ,

$$\begin{aligned} \bar{v}_i^k - \bar{w}_i^k &= \max\{\bar{\sigma}_i^k - \bar{\sigma}_i^{k+1}, 0\} - \max\{\bar{\sigma}_i^{k+1} - \bar{\sigma}_i^k, 0\} \\ &= \bar{\sigma}_i^k - \bar{\sigma}_i^{k+1} \end{aligned} \quad (38)$$

and

$$\begin{aligned} \bar{v}_i^k + \bar{w}_i^k &= \max\{\bar{\sigma}_i^k - \bar{\sigma}_i^{k+1}, 0\} + \max\{\bar{\sigma}_i^{k+1} - \bar{\sigma}_i^k, 0\} \\ &= |\bar{\sigma}_i^k - \bar{\sigma}_i^{k+1}| = |\bar{\sigma}_i^{k+1} - \bar{\sigma}_i^k|. \end{aligned} \quad (39)$$

Using (38), we obtain, for  $k = 1, \dots, p$ ,

$$\begin{aligned} \boldsymbol{\psi}^k(\bar{\boldsymbol{\zeta}}) &= \bar{\boldsymbol{\sigma}}^p + \sum_{l=k}^{p-1} (\bar{\mathbf{v}}^l - \bar{\mathbf{w}}^l) \\ &= \bar{\boldsymbol{\sigma}}^p + \sum_{l=k}^{p-1} (\bar{\boldsymbol{\sigma}}^l - \bar{\boldsymbol{\sigma}}^{l+1}) = \bar{\boldsymbol{\sigma}}^k. \end{aligned}$$

This shows that  $\tilde{\boldsymbol{\sigma}}(\bar{\boldsymbol{\zeta}}) = \bar{\boldsymbol{\sigma}}$ .

Since  $(\bar{\boldsymbol{\tau}}, \bar{\boldsymbol{\sigma}}) = (\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}), \tilde{\boldsymbol{\sigma}}(\bar{\boldsymbol{\zeta}}))$  is feasible for Problem 2, it follows from Theorem 4.2 that  $(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\zeta}})$  is feasible for Problem 3. Also, from Theorem 4.1,  $\mathbf{y}(p|\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\zeta}}) = \mathbf{x}^p(T|\bar{\boldsymbol{\tau}}, \bar{\boldsymbol{\sigma}})$ . Thus, (39) implies

$$\begin{aligned} J_3(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\zeta}}) &= \Psi(\mathbf{y}(p|\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\zeta}})) + \alpha \sum_{i=1}^r \sum_{k=1}^{p-1} (\bar{v}_i^k + \bar{w}_i^k) \\ &= \Psi(\mathbf{x}^p(T|\bar{\boldsymbol{\tau}}, \bar{\boldsymbol{\sigma}})) + \alpha \sum_{i=1}^r \sum_{k=1}^{p-1} |\bar{\sigma}_i^{k+1} - \bar{\sigma}_i^k| \\ &= J_2(\bar{\boldsymbol{\tau}}, \bar{\boldsymbol{\sigma}}). \end{aligned} \quad (40)$$

By combining (40) and (37), and recalling that  $(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\zeta}})$  is feasible for Problem 3, we obtain

$$J_2(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}^*), \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta}^*)) = J_3(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*) \leq J_3(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\zeta}}) = J_2(\bar{\boldsymbol{\tau}}, \bar{\boldsymbol{\sigma}}).$$

Since  $(\bar{\boldsymbol{\tau}}, \bar{\boldsymbol{\sigma}})$  was chosen arbitrarily, this shows that  $(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}^*), \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta}^*))$  is optimal for Problem 2.  $\square$

We now prove the converse of Theorem 4.4.

**Theorem 4.5.** Let  $(\boldsymbol{\tau}^*, \boldsymbol{\sigma}^*) \in \Gamma$  be an optimal solution of Problem 2, where

$$\boldsymbol{\tau}^* = [\tau_1^*, \dots, \tau_{p-1}^*]^\top, \quad \boldsymbol{\sigma}^* = (\boldsymbol{\sigma}^{1,*}, \dots, \boldsymbol{\sigma}^{p,*}).$$

Define

$$\begin{aligned} \theta_k^* &= \tau_k^* - \tau_{k-1}^*, \\ v_i^{k,*} &= \max\{\sigma_i^{k,*} - \sigma_i^{k+1,*}, 0\}, \\ w_i^{k,*} &= \max\{\sigma_i^{k+1,*} - \sigma_i^{k,*}, 0\}, \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{\theta}^* &= [\theta_1^*, \dots, \theta_p^*]^\top, \\ \boldsymbol{\zeta}^* &= (\boldsymbol{\sigma}^{p,*}, \mathbf{v}^{1,*}, \dots, \mathbf{v}^{p-1,*}, \mathbf{w}^{1,*}, \dots, \mathbf{w}^{p-1,*}). \end{aligned}$$

Then  $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*)$  is an optimal solution of Problem 3.



*Proof.* It is clear that  $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*) \in \mathcal{O} \times \mathcal{Z}$ . By following similar arguments to those used in the proof of Theorem 4.4, it is possible to show that  $\boldsymbol{\tau}^* = \tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}^*)$  and  $\boldsymbol{\sigma}^* = \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta}^*)$ . Thus, it follows from Theorem 4.2 that  $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*)$  is feasible for Problem 3.

Now, by Theorem 4.1,

$$\begin{aligned} \Psi(\mathbf{x}^p(T|\boldsymbol{\tau}^*, \boldsymbol{\sigma}^*)) &= \Psi(\mathbf{x}^p(t|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}^*), \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta}^*))) \Big|_{t=\mu(p|\boldsymbol{\theta}^*)} \\ &= \Psi(\mathbf{y}(p|\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*)). \end{aligned} \quad (41)$$

It is easy to see from the definitions of  $v_i^{k,*}$  and  $w_i^{k,*}$  that either  $v_i^{k,*} = 0$  or  $w_i^{k,*} = 0$ . Thus, for each  $i = 1, \dots, r$  and  $k = 1, \dots, p-1$ ,

$$\begin{aligned} |\sigma_i^{k+1,*} - \sigma_i^{k,*}| &= |\psi_i^{k+1}(\boldsymbol{\zeta}^*) - \psi_i^k(\boldsymbol{\zeta}^*)| \\ &= |v_i^{k,*} - w_i^{k,*}| = v_i^{k,*} + w_i^{k,*}. \end{aligned} \quad (42)$$

Combining (41) and (42) gives

$$\begin{aligned} J_2(\boldsymbol{\tau}^*, \boldsymbol{\sigma}^*) &= \Psi(\mathbf{x}^p(T|\boldsymbol{\tau}^*, \boldsymbol{\sigma}^*)) + \alpha \sum_{i=1}^r \sum_{k=1}^{p-1} |\sigma_i^{k+1,*} - \sigma_i^{k,*}| \\ &= \Psi(\mathbf{y}(p|\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*)) + \alpha \sum_{i=1}^r \sum_{k=1}^{p-1} (v_i^{k,*} + w_i^{k,*}) \\ &= J_3(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*). \end{aligned} \quad (43)$$

Now, suppose that  $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*)$  is not an optimal solution for Problem 3. Then there exists a pair  $(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\zeta}}) \in \Xi$  such that  $J_3(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\zeta}}) < J_3(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*)$ , where

$$\bar{\boldsymbol{\zeta}} = (\bar{\gamma}, \bar{v}^1, \dots, \bar{v}^{p-1}, \bar{w}^1, \dots, \bar{w}^{p-1}) \in \mathcal{Z}.$$

It follows from Theorem 4.2 that  $(\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}), \tilde{\boldsymbol{\sigma}}(\bar{\boldsymbol{\zeta}}))$  is feasible for Problem 2. Furthermore, from Theorem 4.1,

$$\begin{aligned} \Psi(\mathbf{y}(p|\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\zeta}})) &= \Psi(\mathbf{x}^p(t|\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}), \tilde{\boldsymbol{\sigma}}(\bar{\boldsymbol{\zeta}}))) \Big|_{t=\mu(p|\bar{\boldsymbol{\theta}})} \\ &= \Psi(\mathbf{x}^p(T|\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}), \tilde{\boldsymbol{\sigma}}(\bar{\boldsymbol{\zeta}}))). \end{aligned} \quad (44)$$

Since  $\bar{v}_i^k$  and  $\bar{w}_i^k$  are non-negative,

$$|\psi_i^{k+1}(\bar{\boldsymbol{\zeta}}) - \psi_i^k(\bar{\boldsymbol{\zeta}})| = |\bar{v}_i^k - \bar{w}_i^k| \leq \bar{v}_i^k + \bar{w}_i^k. \quad (45)$$

By using equations (44) and (45), we obtain inequality  $J_2(\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}), \tilde{\boldsymbol{\sigma}}(\bar{\boldsymbol{\zeta}})) \leq J_3(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\zeta}})$ . Hence, from (43),

$$J_2(\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}), \tilde{\boldsymbol{\sigma}}(\bar{\boldsymbol{\zeta}})) \leq J_3(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\zeta}}) < J_3(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*) = J_2(\boldsymbol{\tau}^*, \boldsymbol{\sigma}^*).$$

But this contradicts the optimality of  $(\boldsymbol{\tau}^*, \boldsymbol{\sigma}^*)$ . Thus,  $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*)$  must be an optimal solution of Problem 3. This completes the proof.  $\square$

The equivalence of Problems 2 and 3 now follows immediately from Theorems 4.4 and 4.5.

**Corollary 4.1.** Problems 2 and 3 are equivalent.

According to Theorem 4.4, Problem 2 can be solved *indirectly* by first solving Problem 3, and then using the solution of Problem 3 to “backtrack” and generate a corresponding solution of Problem 2. Problem 3, a smooth optimization problem with fixed switching times, is clearly much easier to solve than Problem 2, which has a non-smooth cost function and variable switching times. In the next section, we will show that Problem 3 can be solved efficiently using a recently-developed exact penalty method (Yu et al. (2013)).

We emphasize that the transformation procedure described in this section yields an equivalent problem, not an approximation. Thus, our approach has obvious advantages over the existing approach proposed by Teo and Jennings (1991), which involves approximating the total variation term by a smooth function, and then solving a sequence of approximate problems. Using our approach, only one problem (Problem 3) needs to be solved.

Finally, we note that the idea of transforming the variable switching times in Problem 2 into fixed switching times in Problem 3 is motivated by the so-called *time-scaling transformation*, a popular tool for solving switched system optimal control problems (see Xu and Antsaklis (2004); Loxton et al. (2009); Yu et al. (2013)).

## 5 A computational algorithm for Problem 3

Problem 3 is a smooth dynamic optimization problem subject to continuous inequality constraints. Each continuous inequality constraint can be viewed as an infinite number of interior-point constraints—one for each point in  $[0, p]$ . Unfortunately, standard optimization algorithms can only handle a finite number of constraints, and thus such algorithms are not directly applicable to Problem 3. In this section, we will develop a computational algorithm for solving Problem 3 based on the exact penalty method introduced by Yu et al. (2013). The main idea is to approximate Problem 3 by a *penalty problem* in which constraint violations are penalized in the cost function. This penalty problem can be solved readily using standard optimization techniques such as sequential quadratic programming or interior-point methods (Nocedal and Wright (2006); Luenberger and Ye (2008)).

### 5.1 The penalty problem

Let  $\mathcal{V}$  denote the set of all  $\boldsymbol{\theta} \in \mathbb{R}^p$  such that

$$0 \leq \theta_k \leq T, \quad k = 1, \dots, p.$$

Note that  $\mathcal{O} \subset \mathcal{V}$ . Define a *constraint violation function* on  $\mathcal{V} \times \mathcal{Z}$  as follows:

$$\Delta(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \left\{ \sum_{k=1}^p \theta_k - T \right\}^2 + \sum_{j=1}^{q_1} \Phi_j(\mathbf{y}(p|\boldsymbol{\theta}, \boldsymbol{\zeta}))^2 + \sum_{j=1}^{q_2} \sum_{k=1}^p \int_{k-1}^k \max\{g_j^k(s|\boldsymbol{\theta}, \boldsymbol{\zeta}), 0\}^2 ds,$$

where

$$g_j^k(s|\boldsymbol{\theta}, \boldsymbol{\zeta}) = \theta_k h_j(\mathbf{y}(s|\boldsymbol{\theta}, \boldsymbol{\zeta}), \boldsymbol{\psi}^k(\boldsymbol{\zeta})), \\ s \in [k-1, k], \quad j = 1, \dots, q_2.$$

Then for each  $(\boldsymbol{\theta}, \boldsymbol{\zeta}) \in \mathcal{O} \times \mathcal{Z}$ , we have  $\Delta(\boldsymbol{\theta}, \boldsymbol{\zeta}) = 0$  if and only if  $(\boldsymbol{\theta}, \boldsymbol{\zeta})$  is feasible for Problem 3.

Now, define a *penalty function*  $G_\rho$  as follows:

$$G_\rho(\boldsymbol{\theta}, \boldsymbol{\zeta}, \epsilon) = J_3(\boldsymbol{\theta}, \boldsymbol{\zeta}) + \epsilon^{-\delta_1} \Delta(\boldsymbol{\theta}, \boldsymbol{\zeta}) + \rho \epsilon^{\delta_2}, \quad (46)$$

where  $\epsilon > 0$  is a new decision variable,  $\rho > 0$  is the penalty parameter, and  $\delta_1$  and  $\delta_2$  are fixed parameters satisfying  $1 \leq \delta_2 \leq \delta_1$ .

The second term in equation (46) is designed to penalize constraint violations, while the third term penalizes non-zero values of  $\epsilon$ . When  $\epsilon$  is small, the coefficient  $\epsilon^{-\delta_1}$  in the second term is large, thus causing constraint violations to be penalized very severely. Consequently, minimizing the penalty function for large values of  $\rho$  will likely lead to feasible points of Problem 3. With this in mind, we define the following penalty problem.

**Problem 4.** Choose  $(\boldsymbol{\theta}, \boldsymbol{\zeta}, \epsilon) \in \mathcal{V} \times \mathcal{Z} \times (0, \bar{\epsilon}]$  to minimize the penalty function  $G_\rho$ , where  $\bar{\epsilon} > 0$  is a given constant.

The following convergence results are based on the results in Yu et al. (2013).

**Theorem 5.1.** Let  $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*, \epsilon^*)$  be a local solution of Problem 4. Then  $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*)$  is a local solution of Problem 3 if and only if  $\epsilon^* = 0$ .

**Theorem 5.2.** Let  $\{\rho_k\}_{k=1}^\infty$  be an increasing sequence of penalty parameters such that  $\rho_k \rightarrow \infty$  as  $k \rightarrow \infty$ . Furthermore, let  $(\boldsymbol{\theta}^{k,*}, \boldsymbol{\zeta}^{k,*}, \epsilon^{k,*})$  be a solution of Problem 4 with  $\rho = \rho_k$ . Then any limit point of the sequence  $\{(\boldsymbol{\theta}^{k,*}, \boldsymbol{\zeta}^{k,*})\}_{k=1}^\infty$  is a solution of Problem 3.

Theorems 5.1 and 5.2 suggest that Problem 4 is a good approximation of Problem 3 when  $\rho$  is large. Since Problem 4 only involves bound constraints, it is clearly much easier to solve than Problem 3.

In the next subsection, we will derive formulae for the partial derivatives of  $G_\rho$  in Problem 4. These formulae

can be integrated with a gradient-based optimization method (e.g. sequential quadratic programming) to solve Problem 4 efficiently (Nocedal and Wright (2006); Luenberger and Ye (2008)). On this basis, we present the following algorithm for solving Problem 3:

1. Choose  $(\boldsymbol{\theta}^0, \boldsymbol{\zeta}^0) \in \mathcal{V} \times \mathcal{Z}$  (initial guess),  $\rho_0 > 0$  (initial penalty parameter),  $\epsilon_{\min} > 0$  (tolerance),  $\delta_1 \geq 1$  (fixed parameter), and  $\delta_2 \geq \delta_1$  (another fixed parameter).
2. Set  $\bar{\epsilon} \rightarrow \epsilon^0$  and  $\rho_0 \rightarrow \rho$ .
3. Starting with  $(\boldsymbol{\theta}^0, \boldsymbol{\zeta}^0, \epsilon^0)$  as the initial guess, use a gradient-based optimization algorithm to solve Problem 4. Let  $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*, \epsilon^*)$  denote the solution obtained.
4. If  $\epsilon^* < \epsilon_{\min}$ , then stop: take  $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*)$  as the solution of Problem 3. Otherwise, set  $10\rho \rightarrow \rho$  and go to Step 5.
5. Set  $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*, \epsilon^*) \rightarrow (\boldsymbol{\theta}^0, \boldsymbol{\zeta}^0, \epsilon^0)$  and return to Step 3.

This algorithm returns an optimal pair  $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*)$  for Problem 3. The corresponding suboptimal control for Problem 1 is  $\mathbf{u}^p(\cdot|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}^*), \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta}^*))$ , which is defined by equation (11) with  $\boldsymbol{\tau} = \tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}^*)$  and  $\boldsymbol{\sigma} = \tilde{\boldsymbol{\sigma}}(\boldsymbol{\zeta}^*)$ .

## 5.2 Gradient formulae

To solve Problem 4 using a gradient-based optimization method (as required in the above algorithm), we must be able to compute the partial derivatives of the penalty function  $G_\rho$ . These partial derivatives cannot be determined using elementary differentiation rules because the decision vectors  $\boldsymbol{\theta}$  and  $\boldsymbol{\zeta}$  influence the penalty function *implicitly* through the dynamic system (22)-(23). To derive the partial derivatives of  $G_\rho$ , we follow the approach described in Teo, Goh, and Wong (1991).

Define the *Hamiltonian function* as follows:

$$H_k(\epsilon, \theta_k, \boldsymbol{\zeta}, \mathbf{x}, \boldsymbol{\lambda}) = \epsilon^{-\delta_1} \sum_{j=1}^{q_2} \max\{\theta_k h_j(\mathbf{x}, \boldsymbol{\psi}^k(\boldsymbol{\zeta})), 0\}^2 + \theta_k \boldsymbol{\lambda}^\top \mathbf{f}(\mathbf{x}, \boldsymbol{\psi}^k(\boldsymbol{\zeta})),$$

where  $\boldsymbol{\lambda} \in \mathbb{R}^n$  is called the *costate vector*. Consider the following *costate system*:

$$\dot{\boldsymbol{\lambda}}(s) = - \left[ \frac{\partial H_k(\epsilon, \theta_k, \boldsymbol{\zeta}, \mathbf{y}(s), \boldsymbol{\lambda}(s))}{\partial \mathbf{x}} \right]^\top, \quad (47) \\ s \in [k-1, k], \quad k = 1, \dots, p,$$

and

$$\boldsymbol{\lambda}(p) = \left[ \frac{\partial \Psi(\mathbf{y}(p))}{\partial \mathbf{x}} \right]^\top + 2\epsilon^{-\delta_1} \sum_{j=1}^{q_1} \Phi_j(\mathbf{y}(p)) \left[ \frac{\partial \Phi_j(\mathbf{y}(p))}{\partial \mathbf{x}} \right]^\top, \quad (48)$$

where  $\mathbf{y}(s) = \mathbf{y}(s|\boldsymbol{\theta}, \boldsymbol{\zeta})$ . Let  $\boldsymbol{\lambda}(\cdot|\boldsymbol{\theta}, \boldsymbol{\zeta})$  denote the solution of (47)-(48). Then based on the results in Chapter 5 of Teo et al. (1991), we have the following formulae for the partial derivatives of  $G_\rho$ :

$$\frac{\partial G_\rho}{\partial \theta_k} = 2\epsilon^{-\delta_1} \left\{ \sum_{l=1}^p \theta_l - T \right\} + \int_{k-1}^k \frac{\partial H_k(\epsilon, \theta_k, \boldsymbol{\zeta}, \mathbf{y}(s), \boldsymbol{\lambda}(s))}{\partial \theta_k} ds, \quad (49)$$

$$\frac{\partial G_\rho}{\partial \gamma_i} = \sum_{l=1}^p \int_{l-1}^l \frac{\partial H_l(\epsilon, \theta_l, \boldsymbol{\zeta}, \mathbf{y}(s), \boldsymbol{\lambda}(s))}{\partial \gamma_i} ds, \quad (50)$$

$$\frac{\partial G_\rho}{\partial v_i^k} = \alpha + \sum_{l=1}^k \int_{l-1}^l \frac{\partial H_l(\epsilon, \theta_l, \boldsymbol{\zeta}, \mathbf{y}(s), \boldsymbol{\lambda}(s))}{\partial v_i^k} ds, \quad (51)$$

$$\frac{\partial G_\rho}{\partial w_i^k} = \alpha + \sum_{l=1}^k \int_{l-1}^l \frac{\partial H_l(\epsilon, \theta_l, \boldsymbol{\zeta}, \mathbf{y}(s), \boldsymbol{\lambda}(s))}{\partial w_i^k} ds, \quad (52)$$

$$\begin{aligned} \frac{\partial G_\rho}{\partial \epsilon} &= -\delta_1 \epsilon^{-\delta_1-1} \left\{ \sum_{l=1}^p \theta_l - T \right\}^2 + \delta_2 \rho \epsilon^{\delta_2-1} \\ &\quad - \delta_1 \epsilon^{-\delta_1-1} \sum_{j=1}^{q_1} \Phi_j(\mathbf{y}(p))^2 \\ &\quad + \sum_{l=1}^p \int_{l-1}^l \frac{\partial H_l(\epsilon, \theta_l, \boldsymbol{\zeta}, \mathbf{y}(s), \boldsymbol{\lambda}(s))}{\partial \epsilon} ds. \end{aligned} \quad (53)$$

Equations (49)-(53) can be combined with existing nonlinear programming software to solve Problem 4 numerically. To evaluate (49)-(53), it is necessary to first solve the state system (22)-(23) forward in time, and then solve the costate system (47)-(48) backwards in time. An alternative gradient computation scheme in which all integration proceeds in the forward direction can also be derived; see Vincent and Grantham (1981), Kaya and Noakes (2003), Loxton et al. (2008), and Lin et al. (2011) for details.

Gradient-based optimization methods such as sequential quadratic programming are designed to find local (rather than global) solutions. The quality of the local solution obtained depends on the initial guess used to start the optimization process. Thus, to ensure global or near-global optimality, the algorithm presented above for solving Problem 3 should be used with several different initial guesses. Alternatively, this algorithm can be combined with a global optimization method such as the filled function method (Wu, Bai, Lee, and Yang (2007)) or the particle swarm method (Chen, Zhang, Chung, Zhong, Wu, and Shi (2010)).

## 6 Numerical simulations

For numerical testing, we wrote a Fortran program to implement the computational approach described in the

previous sections. This program uses the optimization software NLPQLP (Schittkowski (2007)) to solve Problem 4 as a nonlinear optimization problem, with the cost function's gradient calculated using the formulae in equations (49)-(53). The state and costate systems are solved using the Runge-Kutta method of order 6. The parameters  $\delta_1$  and  $\delta_2$  in the exact penalty function (46) are set equal to  $\delta_1 = 2$  and  $\delta_2 = 1$ .

### 6.1 Optimal fishery harvesting

For our first example, we consider the fishery harvesting problem described in Teo and Jennings (1991) and Teo et al. (1991). This problem involves determining an optimal fishing policy that maximizes total revenue, while at the same time ensuring long-term sustainability of the fish population. The problem dynamics are given below:

$$\dot{x}(t) = a_0 \{ (1 - u(t))x(t) - x(t)^2 \}, \quad t \in [0, 1], \quad (54)$$

$$x(0) = x_0, \quad (55)$$

where  $x(t)$  denotes the fish population at time  $t$  (as a fraction of the carrying capacity of the environment),  $u(t)$  denotes the harvesting effort at time  $t$ ,  $x_0 > 0$  denotes the initial population level, and  $a_0$  is a constant.

The harvesting effort (the control function for this problem) is subject to the following bound constraint:

$$0 \leq u(t) \leq 1, \quad t \in [0, 1]. \quad (56)$$

In addition, the following state constraint is imposed to prevent overfishing:

$$x(t) \geq x_{\min}, \quad t \in [0, 1], \quad (57)$$

where  $x_{\min} > 0$  is a constant. Since  $x_{\min}$  is strictly positive, constraint (57) ensures that the fish population does not become extinct.

The total revenue obtained from harvesting is given by

$$R = \int_0^1 e^{-\varsigma_1 t} \{ b_1 (1 + b_2 (1 - e^{-\varsigma_2 t})) u(t) x(t) - c_1 u(t) - c_2 u(t)^2 \} dt, \quad (58)$$

where  $\varsigma_1, \varsigma_2, b_1, b_2, c_1$ , and  $c_2$  are constants.

We choose the following values for the model constants:

$$\begin{aligned} a_0 &= 0.5, \quad x_0 = 0.45, \quad x_{\min} = 0.4, \quad \varsigma_1 = 1, \quad \varsigma_2 = 5, \\ b_1 &= 1.4, \quad b_2 = 0.25, \quad c_1 = 0.2, \quad c_2 = 0.1. \end{aligned}$$

Our primary goal is to maximize total revenue. However, we are also interested in minimizing control changes, as it is obviously impractical to make frequent changes to

the fishing effort. Thus, we consider a cost function  $J$  that is the sum of negative revenue and total variation of the harvesting function:

$$J = -R + \alpha \bigvee_0^1 u, \quad (59)$$

where  $R$  is defined by (58). Our optimal control problem is defined as follows: *Choose the harvesting function  $u$  to minimize (59) subject to the dynamic system (54)-(55) and the constraints (56) and (57).*

We first solved this problem for  $\alpha = 0$  (i.e. no penalty on control changes). Using our program with  $p = 10$ ,  $\rho_0 = 1$ , and  $\epsilon_{\min} = 10^{-7}$ , we obtained an optimal revenue of  $R = 0.22076$ . For comparison, we also solved the problem using the optimal control software MISER 3.3 (Jennings et al. (2004)). MISER 3.3 returned an optimal revenue of  $R = 0.22051$ , slightly less than our result. This is expected, as MISER 3.3 uses a coarse discretization procedure in which the switching times are fixed constants, not decision variables. In contrast, our program optimizes the switching times—it can even combine two or more consecutive switching times into a single point if it is optimal to do so. MISER 3.3 does not provide this level of flexibility.

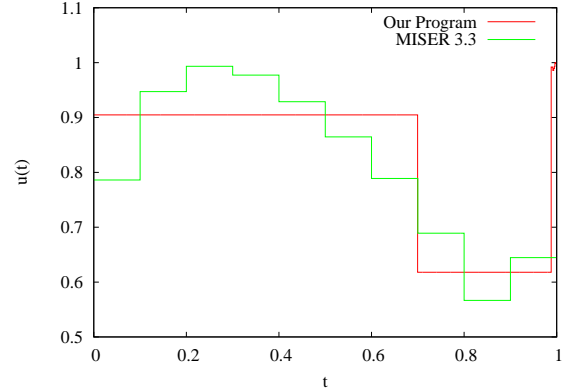
The optimal control produced by our program has a total variation of 0.68188, whereas the optimal control produced by MISER 3.3 has a total variation of 0.71169. These optimal controls, and the corresponding state trajectories, are shown in Figure 1.

To illustrate the effect of  $\alpha$  on the optimal control policy, we now consider the fishery problem with an additional non-zero lower bound on the subinterval durations  $\theta_k$ . Thus, in the corresponding Problem 3 (the transformed problem), we impose the constraint  $\theta_k \geq 0.05$ .

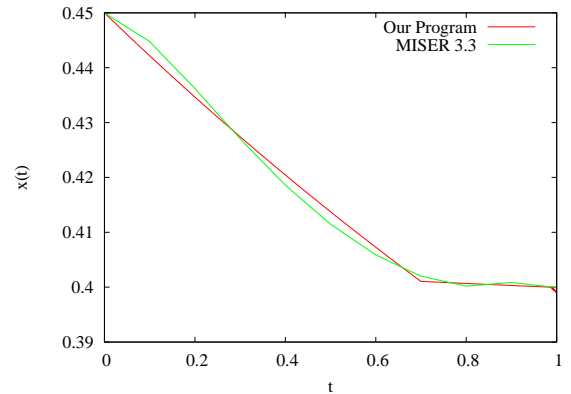
We used our program (with the same parameters as before) to solve the fishery problem for  $\alpha = 0$ ,  $\alpha = 0.001$ ,  $\alpha = 0.005$ , and  $\alpha = 0.01$ . Our numerical results are summarized in Table 1. The optimal fishing policies and corresponding state trajectories are shown in Figure 2. Figure 2(a) clearly shows how increasing  $\alpha$  “smooths” the optimal control. MISER 3.3, which implements the method by Jennings and Teo (1991), produces similar results: for  $\alpha = 0.01$ , MISER 3.3 gives an optimal revenue of  $R = 0.21912$ .

## 6.2 Optimal train control

Consider a train travelling along an uneven track. Let  $x_1$  denote the train’s position along the track and let  $x_2$  denote the train’s velocity. Then according to Vanderbei (2001), the train’s motion can be described by the



(a) Optimal harvesting policy.



(b) Fish population profile.

Fig. 1. Numerical results for Example 6.1 with  $\alpha = 0$ .

$\alpha$	Revenue	Total Variation
0	0.22055	0.60319
0.001	0.22029	0.28304
0.005	0.21982	0.10088
0.010	0.21918	0

Table 1  
Numerical results for Example 6.1 with a non-zero lower bound on the subinterval duration.

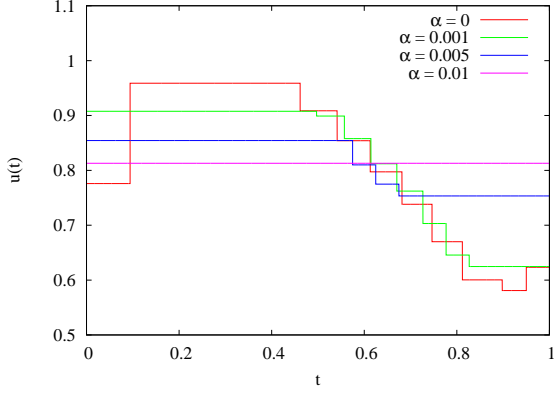
following differential equations:

$$\dot{x}_1(t) = x_2(t), \quad (60)$$

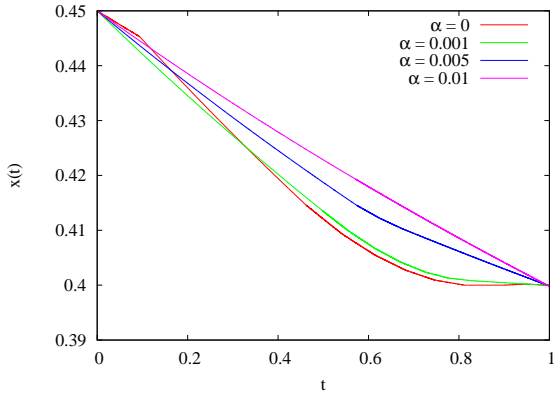
and

$$\dot{x}_2(t) = h(x_1(t)) - (a_1 + a_2 x_2(t) + a_3 x_2(t)^2) + u_1(t) - u_2(t), \quad (61)$$

where  $u_1(t)$  is the acceleration supplied by the engine at time  $t$ ,  $u_2(t)$  is the deceleration supplied by the brakes at time  $t$ ,  $h(x_1(t))$  models the acceleration/deceleration



(a) Optimal harvesting policy.



(b) Optimal fish population profile.

Fig. 2. Numerical results for Example 6.1 with a non-zero lower bound on the subinterval duration.

caused by elevation changes, and  $a_1$ ,  $a_2$ , and  $a_3$  are constants related to the friction force.

Let  $x_1^0$  denote the train's initial position and let  $x_2^0$  denote the train's initial velocity. Then

$$x_1(0) = x_1^0, \quad x_2(0) = x_2^0. \quad (62)$$

The control variables  $u_1$  and  $u_2$  are subject to the following bound constraints:

$$0 \leq u_1(t) \leq 10, \quad 0 \leq u_2(t) \leq 2, \quad t \in [0, T]. \quad (63)$$

At the final time  $t = T$ , the train is required to be in position  $x_1^f$  with velocity  $x_2^f$ . Thus, we impose the following terminal state constraints:

$$x_1(T) = x_1^f, \quad x_2(T) = x_2^f. \quad (64)$$

Since the train cannot travel backwards, we also impose the following continuous inequality constraint:

$$x_2(t) \geq 0, \quad t \in [0, T]. \quad (65)$$

We consider the following cost function:

$$J = \int_0^T u_1(t)x_2(t)dt + \alpha \bigvee_0^T u_1 + \alpha \bigvee_0^T u_2. \quad (66)$$

The first term in (66) measures fuel costs (as a function of the total work performed), while the last two terms penalize changes in the control action.

The optimal train control problem is defined as follows: *Choose  $u_1$  and  $u_2$  to minimize (66) subject to the dynamic system (60)-(62) and the constraints (63)-(65).*

As in Vanderbei (2001), we choose the  $h$  function to be

$$h(x_1(t)) = \frac{b_1 - b_2}{\pi} \tan^{-1} \frac{x_1(t) - c_1}{\epsilon} + \frac{b_2 - b_3}{\pi} \tan^{-1} \frac{x_1(t) - c_2}{\epsilon}.$$

This choice for  $h$  corresponds to the following sequence of elevation changes: an initial uphill climb, followed by a level track, then finishing with a downhill run.

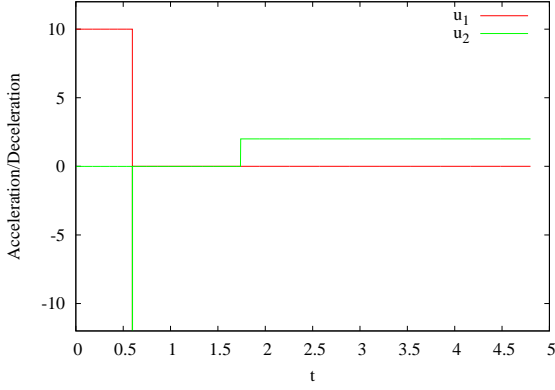
We suppose that the train's initial position is  $x_1^0 = 0$ , the final position is  $x_1^f = 6$ , the initial and final velocities are  $x_2^0 = x_2^f = 0$ , and the journey time is  $T = 4.8$ . For the other constants in the train model, we use the same values as Vanderbei (2001):

$$a_1 = 0.3, \quad a_2 = 0.14, \quad a_3 = 0.16, \quad \epsilon = 0.05, \\ b_1 = 2, \quad b_2 = 0, \quad b_3 = -2, \quad c_1 = 2, \quad c_2 = 4.$$

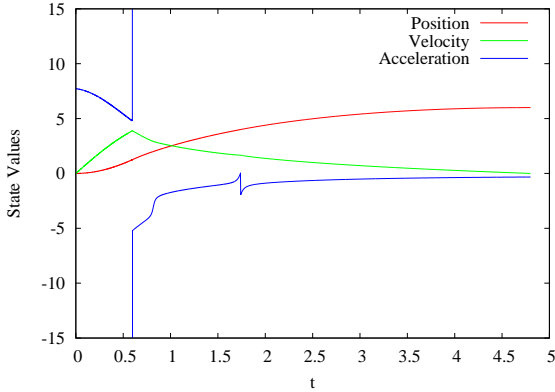
Using our program with parameters  $p = 6$ ,  $\rho_0 = 100$ , and  $\epsilon_{\min} = 10^{-7}$ , we solved the train control problem for  $\alpha = 0$  to obtain an optimal cost of  $J = 12.31826$ . The optimal control and optimal state trajectories are shown in Figure 3. Notice that the control and acceleration functions have large "spikes" near  $t = 0.5$ . This phenomenon, which was also observed in Vanderbei (2001), is caused by numerical error. We now show that, by imposing a small cost on control changes, these numerical errors can be eliminated.

Using our program (this time with  $p = 6$ ,  $\rho_0 = 10$ , and  $\epsilon_{\min} = 10^{-7}$ ) to solve the train control problem for  $\alpha = 0.001$ , we obtained an optimal fuel cost of 12.31195. The optimal controls and corresponding optimal state trajectories are shown in Figure 4. Note that the large spikes have disappeared. Thus, including a total variation term in the cost function can help to regularize the optimal control computation.

Note that the optimal fuel cost for  $\alpha = 0.001$  is actually slightly less than the optimal fuel cost for  $\alpha = 0$ . This indicates that our previous solution for  $\alpha = 0$  is a non-global local solution. Re-starting our program for  $\alpha = 0$



(a) Optimal control functions.



(b) Optimal state trajectories.

Fig. 3. Numerical results for Example 6.2 with  $\alpha = 0$ .

using the solution for  $\alpha = 0.001$  as the initial guess, we obtained an improved solution with an optimal fuel cost of 12.31182.

Our program achieved much better results than MISER for this example. For  $\alpha = 0$ , MISER gives an optimal cost of 13.04940. This is worse than our result, even with a penalty on control changes. For  $\alpha = 0.1$ , MISER gives an optimal cost of 13.05253.

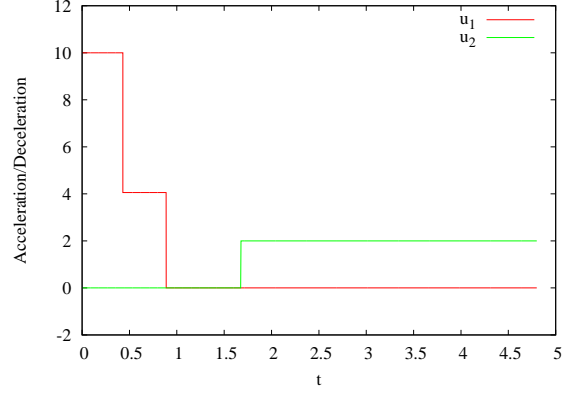
### 6.3 Optimal control of an exothermic chemical reaction

Our final example comes from Zhao and Stadtherr (2011). Consider the reaction  $A + B \rightarrow C$  occurring in an isothermal semibatch reactor operating at temperature  $T = 343.15$  Kelvin. The problem dynamics are

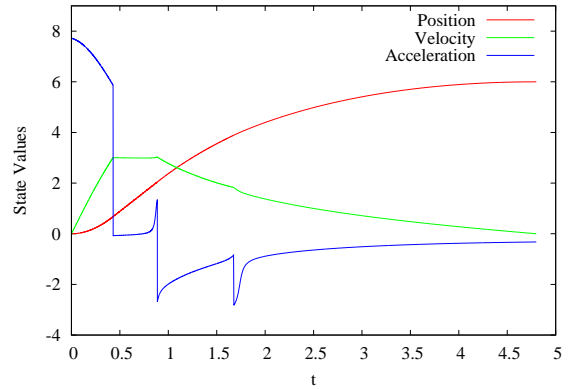
$$\dot{x}_1(t) = -kx_1(t)x_2(t) - \frac{u(t)x_1(t)}{x_3(t)}, \quad (67)$$

$$\dot{x}_2(t) = -kx_1(t)x_2(t) + \frac{u(t)(c_B - x_2(t))}{x_3(t)}, \quad (68)$$

$$\dot{x}_3(t) = u(t), \quad (69)$$



(a) Optimal control functions.



(b) Optimal state trajectories.

Fig. 4. Numerical results for Example 6.2 with  $\alpha = 0.001$ .

where  $x_1(t)$  denotes the concentration of  $A$  at time  $t$ ,  $x_2(t)$  denotes the concentration of  $B$  at time  $t$ ,  $x_3(t)$  denotes the volume at time  $t$ ,  $u(t)$  denotes the volumetric flow rate of the feed stream at time  $t$ ,  $c_B = 2$  is the concentration of  $B$  in the feed stream, and  $k = 0.0482$  is a reaction rate constant.

The objective is to maximize the amount of product  $C$  at the final time  $t = 20$  hours. The initial conditions are

$$x_1(0) = 2, \quad x_2(0) = 0.5, \quad x_3(0) = 0.7. \quad (70)$$

The volumetric flow rate is subject to the following bounds:

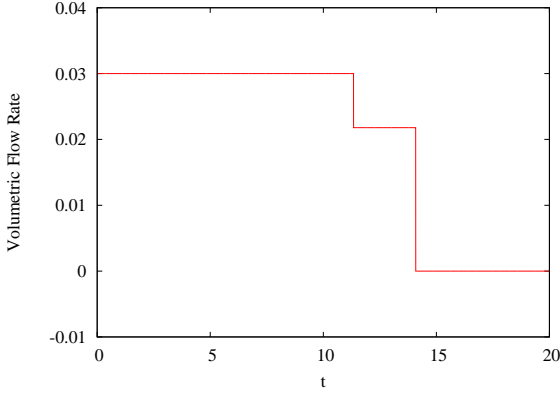
$$0 \leq u(t) \leq 0.03, \quad t \in [0, 20]. \quad (71)$$

In addition, there are two continuous inequality constraints: a maximum volume constraint and a cooling-failure safety constraint. These constraints are formulated as follows:

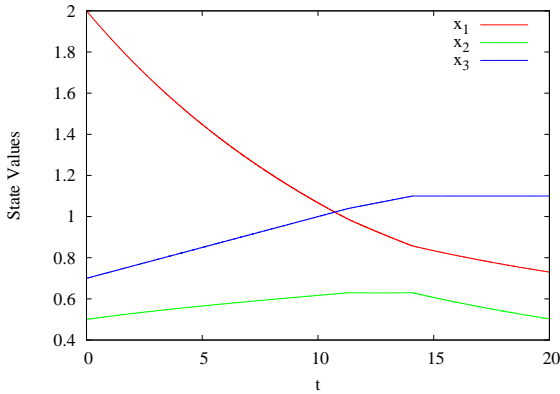
$$x_3(t) \leq V_{\max}, \quad t \in [0, 20], \quad (72)$$

and

$$T + mx_2(t) \leq T_{\max}, \quad t \in [0, 20], \quad (73)$$



(a) Optimal control function.



(b) Optimal state trajectories.

Fig. 5. Numerical results for Example 6.3 with  $\alpha = 0$ .

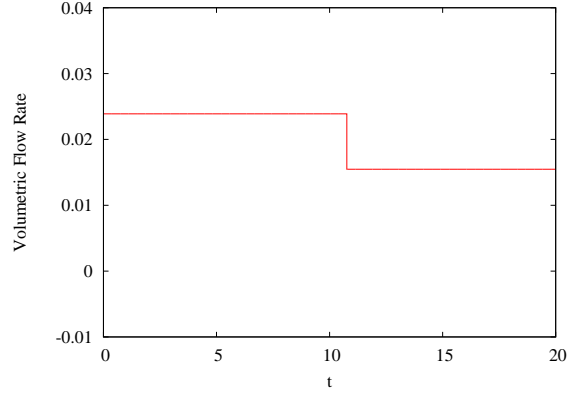
where  $V_{\max} = 1.1$ ,  $T = 343.15$ ,  $T_{\max} = 353.15$ , and  $m = 15.8730$ . Our aim is to minimize the following cost function:

$$J = x_1(20)x_3(20) - x_1(0)x_3(0) + \alpha \bigvee_0^{20} u. \quad (74)$$

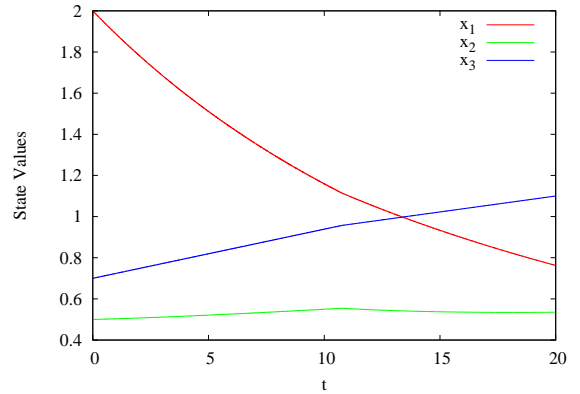
The optimal control problem is defined as: *Choose the volumetric flow rate  $u$  to minimize (74) subject to the dynamic system (67)-(70) and the constraints (71)-(73).*

We first considered this problem for  $\alpha = 0$ . Using our program with  $p = 7$ ,  $\rho_0 = 0.1$ , and  $\epsilon_{\min} = 10^{-8}$ , we obtained an optimal cost of  $-0.59666$ , which is slightly better than the results in Zhao and Stadtherr (2011). Table 5 in Zhao and Stadtherr (2011) reports computation times of up to 11 hours. Our program took less than one minute. The optimal control and corresponding state trajectories produced by our program are shown in Figure 5. Note that the optimal control only involves 2 switches, even though  $p - 1 = 6$  switches are available.

We now consider the case when  $\alpha = 2$ . Using our program with  $p = 7$ ,  $\rho_0 = 0.1$ , and  $\epsilon_{\min} = 10^{-8}$ , we ob-



(a) Optimal control function.



(b) Optimal state trajectories.

Fig. 6. Numerical results for Example 6.3 with  $\alpha = 2$ .

tained an optimal cost of  $-0.56116$  (minus the bounded variation term). The optimal control and optimal state trajectories are shown in Figure 6. Note that the new optimal control only involves 1 switch. Thus, imposing a cost on control changes results in a simpler control, at the expense of slightly increased cost.

## 7 Conclusion

This paper introduces a novel optimal control problem in which control changes are penalized in the cost function via a total variation term. In general, this total variation term cannot be computed analytically. We developed a numerical discretization procedure that involves approximating the control signal by a piecewise-constant function. After applying this discretization procedure, the total variation term can be evaluated easily. We then developed a transformation procedure for transforming the non-smooth approximate problem into an equivalent problem that can be solved readily using existing techniques. Numerical results indicate that this approach is effective for a range of practical problems.

## References

- [1] B. Açikmeşe and L. Blackmore, "Lossless convexification of a class of non-convex optimal control problems for linear systems," in *Proceedings of the 2010 American Control Conference*, Baltimore, USA, 2010.
- [2] B. Açikmeşe and L. Blackmore, "Lossless convexification of a class of optimal control problems with non-convex control constraints," *Automatica*, vol. 47, no. 2, pp. 341-347, 2011.
- [3] N. U. Ahmed, *Elements of Finite-Dimensional Systems and Control Theory*. Essex: Longman Scientific and Technical, 1988.
- [4] N. U. Ahmed, *Dynamic Systems and Control with Applications*. Singapore: World Scientific, 2006.
- [5] J. M. Blatt, "Optimal control with a cost of switching control," *Journal of the Australian Mathematical Society – Series B: Applied Mathematics*, vol. 19, no. 3, pp. 316-332, 1976.
- [6] C. Büskens and H. Maurer, "SQP-methods for solving optimal control problems with control and state constraints: Adjoint variables, sensitivity analysis, and real-time control," *Journal of Computational and Applied Mathematics*, vol. 120, no. 1-2, pp. 85-108, 2000.
- [7] W. N. Chen, J. Zhang, H. S. H. Chung, W. L. Zhong, W. G. Wu, and Y. H. Shi, "A novel set-based particle swarm optimization method for discrete optimization problems," *IEEE Transactions on Evolutionary Computation*, vol. 14, no. 2, pp. 278-300, 2010.
- [8] M. Gerdts, "Global convergence of a nonsmooth Newton method for control-state constrained optimal control problems," *SIAM Journal on Optimization*, vol. 19, no. 1, pp. 326-350, 2008.
- [9] M. Gerdts and M. Kunkel, "A nonsmooth Newton's method for discretized optimal control problems with state and control constraints," *Journal of Industrial and Management Optimization*, vol. 4, no. 2, pp. 247-270, 2008.
- [10] R. F. Hartl, S. P. Sethi, and R. G. Vickson, "A survey of the maximum principles for optimal control problems with state constraints," *SIAM Review*, vol. 37, no. 2, pp. 181-218, 1995.
- [11] L. S. Jennings, M. E. Fisher, K. L. Teo, and C. J. Goh, *MISER 3 Optimal Control Software: Theory and User Manual*, version 3. University of Western Australia, 2004.
- [12] C. Jiang, Q. Lin, C. Yu, K. L. Teo, and G. R. Duan, "An exact penalty method for free terminal time optimal control problem with continuous inequality constraints," *Journal of Optimization Theory and Applications*, vol. 154, no. 1, pp. 30-53, 2012.
- [13] C. Y. Kaya and J. L. Noakes, "Computational method for time-optimal switching control," *Journal of Optimization Theory and Applications*, vol. 117, no. 1, pp. 69-92, 2003.
- [14] Q. Lin, R. Loxton, K. L. Teo, and Y. H. Wu, "A new computational method for a class of free terminal time optimal control problems," *Pacific Journal of Optimization*, vol. 7, no. 1, pp. 63-81, 2011.
- [15] R. Loxton, Q. Lin, V. Rehbock, and K. L. Teo, "Control parameterization for optimal control problems with continuous inequality constraints: New convergence results," *Numerical Algebra, Control and Optimization*, vol. 2, no. 3, pp. 571-599, 2012.
- [16] R. Loxton, K. L. Teo, and V. Rehbock, "Optimal control problems with multiple characteristic time points in the objective and constraints," *Automatica*, vol. 44, no. 11, pp. 2923-2929, 2008.
- [17] R. Loxton, K. L. Teo, V. Rehbock, and K. F. C. Yiu, "Optimal control problems with a continuous inequality constraint on the state and the control," *Automatica*, vol. 45, no. 10, pp. 2250-2257, 2009.
- [18] D. G. Luenberger and Y. Ye, *Linear and Nonlinear Programming*, 3rd edition. New York: Springer, 2008.
- [19] J. Matula, "On an extremum problem," *Journal of the Australian Mathematical Society – Series B: Applied Mathematics*, vol. 28, no. 3, pp. 376-392, 1987.
- [20] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd edition. New York: Springer, 2006.
- [21] H. L. Royden and P. M. Fitzpatrick, *Real Analysis*, 4th edition. Boston: Prentice Hall, 2010.
- [22] Y. Sakawa and Y. Shindo, "Optimal control of container cranes," *Automatica*, vol. 18, no. 3, pp. 257-266, 1982.
- [23] K. Schittkowski, *NLPQLP: A Fortran Implementation of a Sequential Quadratic Programming Algorithm with Distributed and Non-Monotone Line Search*, version 2.24. University of Bayreuth, 2007.
- [24] D. E. Stewart, "A numerical algorithm for optimal control problems with switching costs," *Journal of the Australian Mathematical Society – Series B: Applied Mathematics*, vol. 34, no. 2, pp. 212-228, 1992.
- [25] K. L. Teo, C. J. Goh, and K. H. Wong, *A Unified Computational Approach to Optimal Control Problems*. Essex: Longman Scientific and Technical, 1991.
- [26] K. L. Teo and L. S. Jennings, "Optimal control with a cost on changing control," *Journal of Optimization Theory and Applications*, vol. 68, no. 2, pp. 335-357, 1991.
- [27] R. J. Vanderbei, "Case studies in trajectory optimization: Trains, planes, and other pastimes," *Optimization and Engineering*, vol. 2, no. 2, pp. 215-243, 2001.
- [28] T. L. Vincent and W. J. Grantham, *Optimality in Parametric Systems*. New York: John Wiley, 1981.
- [29] L. Y. Wang, W. H. Gui, K. L. Teo, R. Loxton, and C. H. Yang, "Time delayed optimal control problems with multiple characteristic time points: Computation and industrial applications," *Journal of Industrial and Management Optimization*, vol. 5, no. 4, pp. 705-718, 2009.
- [30] Z. Y. Wu, F. S. Bai, H. W. J. Lee, and Y. J. Yang, "A filled function method for constrained global optimization," *Journal of Global Optimization*, vol. 39, no. 4, pp. 495-507, 2007.
- [31] X. Xu and P. J. Antsaklis, "Optimal control of switched systems based on parameterization of the switching instants," *IEEE Transactions on Automatic Control*, vol. 49, no. 1, pp. 2-16, 2004.
- [32] C. Yu, B. Li, R. Loxton, and K. L. Teo, "Optimal discrete-valued control computation," *Journal of Global Optimization*, in press.
- [33] Y. Zhao and M. A. Stadtherr, "Rigorous global optimization for dynamic systems subject to inequality path constraints," *Industrial and Engineering Chemistry Research*, vol. 50, no. 22, pp. 12678-12693, 2011.