Optimal Control Computation for Nonlinear Systems with State-dependent Stopping Criteria [⋆]

Qun Lin [a], Ryan Loxton [a], Kok Lay Teo [a], Yong Hong Wu [a]

[a] *Department of Mathematics and Statistics, Curtin University, Australia*

Abstract

In this paper, we consider a challenging optimal control problem in which the terminal time is determined by a stopping criterion. This stopping criterion is defined by a smooth surface in the state space; when the state trajectory hits this surface, the governing dynamic system stops. By restricting the controls to piecewise constant functions, we derive a finite-dimensional approximation of the optimal control problem. We then develop an efficient computational method, based on nonlinear programming, for solving the approximate problem. We conclude the paper with four numerical examples.

Key words: Nonlinear optimal control; Control parameterization; Nonlinear programming; Time-scaling transformation.

1 Introduction

The aim of optimal control is to manipulate a given control system in an optimal manner. One of the most famous problems in optimal control is the so-called *time-optimal control problem*, which involves designing a control law to steer a system from an initial state to a target state in minimum time. There are many effective computational methods for solving time-optimal control problems; these include the time-optimal switching algorithm by Kaya & Noakes (2003), and the control parameterization enhancing technique by Lee, Teo, Rehbock & Jennings (1997). Computational methods for solving more general optimal control problems are also available; see, for example, Gerdts & Kunkel (2008), Kaya & Martínez (2007), Hager (2000), Luus (2000), and von Stryk (1993).

In most optimal control problems, the time at which the control system stops—the so-called *terminal time*—is either fixed and known (and possibly infinite) or a free decision variable. In this paper, we consider a different type of optimal control problem in which the terminal time is neither fixed nor free; instead, it is defined as the first time at which the system trajectory reaches a certain *stopping surface*. When defined in this way, the terminal time is actually an implicit function of the control, as changing the control changes the system trajectory, which in turn changes the time at which the trajectory hits the stopping surface.

Optimal control problems of this type arise in aeronautical applications. Teo, Jepps, Moore & Hayes (1987) considered one such problem, where the aim is to maximize the range of a gliding projectile. The dynamic system in this problem consists of ordinary differential equations describing the glider's motion. These equations are only valid when the glider's altitude is positive—the glider crashes as soon as it hits the ground, and the time at which this occurs depends on the glider's control strategy. Thus, the glider's terminal time is not constant, but is instead determined by a stopping criterion.

Computational methods for solving such problems are discussed in Teo et al. (1987) and Teo, Goh & Lim (1989). These computational methods are based on a discretization scheme whereby the control is approximated by a piecewise constant function. The heights of this piecewise constant function are taken as decision variables to be chosen optimally, whereas the times at which it changes from one height to another—the so-called *switching times*—are pre-specified.

Recently in Lin, Loxton, Teo & Wu (2011), a new computational method that supersedes the old methods in Teo et al. (1987,1989) was developed. This new method

[⋆] This paper was not presented at any IFAC meeting. Corresponding author Q. Lin. Tel. +61-8-92664962. Fax +61-8-92663197.

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

is also based on a piecewise constant approximation of the control, but it allows *both* the control heights and the control switching times to be decision variables. This new approximation scheme is far more accurate than the old schemes in Teo et al. (1987,1989), which only allow the control heights to be chosen optimally. Unfortunately, a major disadvantage of the approximation scheme in Lin et al. (2011) is that it leads to an approximate nonlinear programming problem that is very difficult to solve numerically. The purpose of this paper is to present a new method for solving this approximate problem. To this end, we will introduce a novel procedure for transforming the approximate problem into a new problem that is easier to solve. We will then develop rigorous theory linking the two problems, before showing how to construct a solution of the approximate problem from a solution of the new problem. We will also develop an algorithm for solving the new problem using standard nonlinear programming techniques. By using this new approach, one can avoid the drawbacks of the existing methods in Teo et al. (1987,1989) and Lin et al. (2011).

2 Problem Statement

Consider the following nonlinear dynamic system:

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)), \quad t \geq 0, \tag{1}$$

and

$$\mathbf{x}(0) = \mathbf{x}^0, \tag{2}$$

where $\mathbf{x}(t) \in \mathbb{R}^n$ is the *state* at time t, $\mathbf{u}(t) \in \mathbb{R}^r$ is the *control* at time t, $\mathbf{x}^0 \in \mathbb{R}^n$ is a given initial state, and $\mathbf{f} : \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}^n$ is a given function.

Define

$$W := \{\, \mathbf{w} \in \mathbb{R}^r : \ a_i \leq w_i \leq b_i, \ i = 1, \ldots, r \,\},$$

where a_i and b_i, $i = 1, \ldots, r$, are given real numbers such that $a_i < b_i$. Any measurable function $\mathbf{u} : [0, \infty) \to \mathbb{R}^r$ such that $\mathbf{u}(t) \in W$ for almost all $t \geq 0$ is called an *admissible control*. Let $\mathcal{U}$ denote the class of all such admissible controls.

We assume that the following conditions are satisfied.

Assumption 2.1. The function $\mathbf{f}$ is continuously differentiable.

Assumption 2.2. There exists a real number $L_1 > 0$ such that

$$\|\mathbf{f}(\mathbf{v}, \mathbf{w})\| \leq L_1(1 + \|\mathbf{v}\| + \|\mathbf{w}\|), \quad (\mathbf{v}, \mathbf{w}) \in \mathbb{R}^n \times \mathbb{R}^r,$$

where $\|\cdot\|$ denotes the Euclidean norm.

Let $\mathbf{x}(\cdot|\mathbf{u})$ denote the solution of (1)-(2) corresponding to the admissible control $\mathbf{u} \in \mathcal{U}$. Assumptions 2.1 and 2.2 ensure that $\mathbf{x}(\cdot|\mathbf{u})$ exists and is unique (see Theorem 3.1.6 of Ahmed (1988)).

We introduce the following *stopping surface* for control system (1)-(2):

$$\Omega := \{\, \mathbf{v} \in \mathbb{R}^n : \ \Phi(\mathbf{v}) = 0 \,\},$$

where $\Phi : \mathbb{R}^n \to \mathbb{R}$ is a given continuously differentiable function. System (1)-(2) stops once its state trajectory hits this surface. Thus, the *terminal time* for (1)-(2) is defined as follows:

$$T(\mathbf{u}) := \inf\{\, t > 0 : \ \mathbf{x}(t|\mathbf{u}) \in \Omega \,\}.$$

We assume that $T(\mathbf{u})$ is finite for each admissible control. This assumption is stated precisely below.

Assumption 2.3. There exists real numbers $T_{\min} > 0$ and $T_{\max} > 0$ such that

$$T_{\min} \leq T(\mathbf{u}) \leq T_{\max}, \quad \mathbf{u} \in \mathcal{U}.$$

System (1)-(2) starts in state $\mathbf{x}^0$ at time $t = 0$ and evolves according to (1) until its state trajectory reaches the stopping surface at time $t = T(\mathbf{u})$. The system then terminates; its final state is $\mathbf{x}(T(\mathbf{u})|\mathbf{u})$. Since Φ is continuous, one can easily show that

$$\mathbf{x}(T(\mathbf{u})|\mathbf{u}) \in \Omega, \quad \mathbf{u} \in \mathcal{U}. \tag{3}$$

We will use this inclusion later in the paper.

Now, we define a *cost function* J as follows:

$$J(\mathbf{u}) := \Psi(\mathbf{x}(t|\mathbf{u}))\big|_{t=T(\mathbf{u})}, \quad \mathbf{u} \in \mathcal{U}, \tag{4}$$

where $\Psi : \mathbb{R}^n \to \mathbb{R}$ is a given continuously differentiable function. Our goal is to choose an admissible control with minimum cost. We state this formally as the following optimal control problem.

Problem 1. Find an admissible control $\mathbf{u}^* \in \mathcal{U}$ such that

$$J(\mathbf{u}^*) = \inf_{\mathbf{u} \in \mathcal{U}} J(\mathbf{u}).$$

Note that the terminal time in Problem 1 is a function of the control. Thus, the control influences both the state trajectory and the time horizon over which the state trajectory evolves. This is quite different from standard optimal control problems in which the terminal time is fixed and known. Conventional optimal control methods, which assume that the terminal time is fixed, cannot be applied to Problem 1. The purpose of this paper is to develop a new method for solving Problem 1.

3 Problem Approximation

To proceed, we will use the control parameterization technique to approximate Problem 1 by a finite-dimensional optimization problem. Control parameterization has already been successfully applied to standard optimal control problems in which the terminal time is either fixed or free—see, for example, Chyba, Haberkorn, Singh, Smith & Choi (2009), Luus (2000), and Teo, Goh & Wong (1991).

Let $p \geq 2$ be a fixed integer and define

$$\mathcal{Z} := \prod_{k=1}^{p} W.$$

Thus, $\mathcal{Z}$ is the set of all p-tuples $(\boldsymbol{\zeta}^1, \ldots, \boldsymbol{\zeta}^p)$ such that $\boldsymbol{\zeta}^k \in W$, $k = 1, \ldots, p$.

Let $\mathcal{T}$ denote the set of all $\boldsymbol{\tau} = [\tau_1, \ldots, \tau_{p-1}]^\top \in \mathbb{R}^{p-1}$ such that

$$\tau_k \geq 0, \quad k = 1, \ldots, p-1,$$

and

$$\tau_{k-1} \leq \tau_k, \quad k = 2, \ldots, p-1.$$

For each $\boldsymbol{\tau} \in \mathcal{T}$, define corresponding intervals $\mathcal{I}_k(\boldsymbol{\tau})$, $k = 1, \ldots, p$ as follows:

$$\mathcal{I}_k(\boldsymbol{\tau}) := \begin{cases} [0, \tau_1), & \text{if } k = 1, \\ [\tau_{k-1}, \tau_k), & \text{if } k = 2, \ldots, p-1, \\ [\tau_{p-1}, \infty), & \text{if } k = p. \end{cases}$$

Clearly,

$$\mathcal{I}_i(\boldsymbol{\tau}) \cap \mathcal{I}_j(\boldsymbol{\tau}) = \emptyset, \quad i \neq j,$$

and

$$\bigcup_{k=1}^{p} \mathcal{I}_k(\boldsymbol{\tau}) = [0, \infty).$$

Thus, $\{\, \mathcal{I}_k(\boldsymbol{\tau}),\, k = 1, \ldots, p \,\}$ is a partition of $[0, \infty)$.

Now, for each $(\boldsymbol{\tau}, \boldsymbol{\zeta}) \in \mathcal{T} \times \mathcal{Z}$, define a corresponding function $\mathbf{u}^p(\cdot|\boldsymbol{\tau}, \boldsymbol{\zeta}) : [0, \infty) \to \mathbb{R}^r$ as follows:

$$\mathbf{u}^p(t|\boldsymbol{\tau}, \boldsymbol{\zeta}) := \sum_{k=1}^{p} \boldsymbol{\zeta}^k \chi_{\mathcal{I}_k(\boldsymbol{\tau})}(t), \quad t \geq 0, \tag{5}$$

where $\chi_{\mathcal{I}} : \mathbb{R} \to \mathbb{R}$ is the indicator function defined by

$$\chi_{\mathcal{I}}(t) := \begin{cases} 1, & \text{if } t \in \mathcal{I}, \\ 0, & \text{otherwise.} \end{cases}$$

We immediately see that

$$\mathbf{u}^p(t|\boldsymbol{\tau}, \boldsymbol{\zeta}) = \boldsymbol{\zeta}^k, \quad t \in \mathcal{I}_k(\boldsymbol{\tau}), \quad k = 1, \ldots, p.$$

This shows that $\mathbf{u}^p(\cdot|\boldsymbol{\tau}, \boldsymbol{\zeta})$ is piecewise constant and $\mathbf{u}^p(t|\boldsymbol{\tau}, \boldsymbol{\zeta}) \in W$ for all $t \geq 0$. Hence, $\mathbf{u}^p(\cdot|\boldsymbol{\tau}, \boldsymbol{\zeta})$ is an admissible control for Problem 1.

Let $\mathbf{x}^p(\cdot|\boldsymbol{\tau}, \boldsymbol{\zeta})$ denote the unique solution of (1)-(2) corresponding to the admissible control $\mathbf{u}^p(\cdot|\boldsymbol{\tau}, \boldsymbol{\zeta})$. More precisely,

$$\mathbf{x}^p(\cdot|\boldsymbol{\tau}, \boldsymbol{\zeta}) := \mathbf{x}(\cdot|\mathbf{u}^p(\cdot|\boldsymbol{\tau}, \boldsymbol{\zeta})).$$

Similarly, for each $(\boldsymbol{\tau}, \boldsymbol{\zeta}) \in \mathcal{T} \times \mathcal{Z}$, let

$$T^p(\boldsymbol{\tau}, \boldsymbol{\zeta}) := T(\mathbf{u}^p(\cdot|\boldsymbol{\tau}, \boldsymbol{\zeta})), \quad J^p(\boldsymbol{\tau}, \boldsymbol{\zeta}) := J(\mathbf{u}^p(\cdot|\boldsymbol{\tau}, \boldsymbol{\zeta})).$$

It follows from inclusion (3) that

$$\begin{aligned} &\mathbf{x}^p(t|\boldsymbol{\tau}, \boldsymbol{\zeta})\big|_{t=T^p(\boldsymbol{\tau}, \boldsymbol{\zeta})} \\ &\qquad = \mathbf{x}(t|\mathbf{u}^p(\cdot|\boldsymbol{\tau}, \boldsymbol{\zeta}))\big|_{t=T(\mathbf{u}^p(\cdot|\boldsymbol{\tau}, \boldsymbol{\zeta}))} \in \Omega. \end{aligned} \tag{6}$$

We have now expressed Problem 1's state, terminal time, and cost function in terms of the decision variables $\boldsymbol{\tau} \in \mathcal{T}$ and $\boldsymbol{\zeta} \in \mathcal{Z}$. This leads to the following approximate optimization problem.

Problem 2. Find a pair $(\boldsymbol{\tau}^*, \boldsymbol{\zeta}^*) \in \mathcal{T} \times \mathcal{Z}$ such that

$$J^p\left(\boldsymbol{\tau}^*, \boldsymbol{\zeta}^*\right) = \inf_{(\boldsymbol{\tau}, \boldsymbol{\zeta}) \in \mathcal{T} \times \mathcal{Z}} J^p(\boldsymbol{\tau}, \boldsymbol{\zeta}).$$

If $(\boldsymbol{\tau}^*, \boldsymbol{\zeta}^*)$ is an optimal solution for Problem 2, then $\mathbf{u}^p(\cdot|\boldsymbol{\tau}^*, \boldsymbol{\zeta}^*)$ is a corresponding suboptimal control for Problem 1. Does this suboptimal control converge to an optimal control as the number of subintervals p increases? In Lin et al. (2011), we showed that if $\mathbf{u}^*$ is an optimal control for Problem 1, and if $(\boldsymbol{\tau}^{p,*}, \boldsymbol{\zeta}^{p,*})$ is an optimal solution for Problem 2, then

$$\lim_{p \to \infty} J(\mathbf{u}^p(\cdot|\boldsymbol{\tau}^{p,*}, \boldsymbol{\zeta}^{p,*})) = J(\mathbf{u}^*).$$

Furthermore, if $\mathbf{u}^p(\cdot|\boldsymbol{\tau}^{p,*}, \boldsymbol{\zeta}^{p,*})$ converges to a function $\bar{\mathbf{u}} : [0, \infty) \to \mathbb{R}^r$ almost everywhere as $p \to \infty$, then $\bar{\mathbf{u}}$ is an optimal control for Problem 1. These two results suggest that $\mathbf{u}^p(\cdot|\boldsymbol{\tau}^{p,*}, \boldsymbol{\zeta}^{p,*})$ is a good approximation of the optimal control when p is large. Hence, by solving Problem 2 for large p, we can generate a high-quality suboptimal control for Problem 1. However, Problem 2 is difficult to solve using standard optimization techniques because it involves a nonlinear dynamic system with variable switching times (see Loxton, Teo & Rehbock (2008)). In the next two sections, we will develop an effective computational method for solving Problem 2.

4 A New Finite-Dimensional Problem

The decision variables in Problem 2 are the *control switching times* τ_k, $k = 1, \ldots, p-1$ and the *control* *heights* $\boldsymbol{\zeta}^k$, $k = 1, \ldots, p$. In Lin et al. (2011), we developed an algorithm for computing the partial derivatives of J^p with respect to these decision variables. We then showed that this algorithm can be used in conjunction with a gradient-based optimization method—for example, a conjugate gradient method—to solve Problem 2.

Although this approach yields satisfactory results, it has two major shortcomings:

(i) Integrating the dynamic system (1)-(2) numerically is very difficult when the control switching times are variable, as they are in Problem 2.
(ii) Computing the partial derivatives of J^p involves integrating two auxiliary dynamic systems, and these auxiliary systems are not well-defined if two or more switching times coincide (i.e. if $\tau_{k-1} = \tau_k$ for some k). The algorithm in Lin et al. (2011) will stall if this occurs.

The purpose of this paper is to develop a new method that does not suffer from these drawbacks. We will not tackle Problem 2 directly; instead, we will introduce a new optimization problem and show that a solution of this new problem can be used to generate a solution of Problem 2. As we will see, solving the new problem is much easier than solving Problem 2.

Our approach is inspired by the so-called *time-scaling transformation* first introduced by Lee et al. (1997) to solve time-optimal control problems. In a time-optimal control problem, the optimal control is usually a bang-bang control oscillating between its maximum and minimum values. The times at which the control switches between these values must be chosen optimally. Thus, the control switching times in a time-optimal control problem are decision variables, just like in Problem 2. The time-scaling transformation can be used to map these switching times to fixed points in a new time horizon, which results in a new problem that is easier to solve. The standard time-scaling transformation works by introducing a new time variable s, and then relating s to t through the following differential equation:

$$\frac{dt(s)}{ds} = v(s), \quad s \in [0, p],$$

together with the boundary conditions

$$t(0) = 0, \quad t(p) = T,$$

where $p - 1$ is the number of switches, $v : [0, p] \to [0, \infty)$ is a piecewise constant function, and T is the terminal time. The time-scaling transformation is a useful tool for circumventing the difficulties caused by variable switching times. However, the standard time-scaling transformation is only applicable to optimal control problems in which the terminal time T is a fixed constant or a free decision variable. As such, it cannot be applied to Problem 2, whose terminal time $T^p(\boldsymbol{\tau}, \boldsymbol{\zeta})$ is an implicit function of the decision variables $\boldsymbol{\tau}$ and $\boldsymbol{\zeta}$.

4.1 Problem Statement

As in Section 3, let $p \geq 2$ be a fixed integer. Define

$$\Theta := \{\, \boldsymbol{\theta} \in \mathbb{R}^{p-1} : \ \theta_1 > 0; \ \theta_k \geq 0, \ k = 2, \ldots, p-1 \,\}$$

and

$$\tilde{\mathcal{I}}_k := \begin{cases} (k-1, k), & \text{if } k = 1, \ldots, p-1, \\ (p-1, \infty), & \text{if } k = p. \end{cases}$$

Consider the following *switched system* evolving in the state space $\mathbb{R}^n$:

$$\dot{\mathbf{y}}(s) = \theta_k \mathbf{f}(\mathbf{y}(s), \boldsymbol{\zeta}^k), \quad s \in \tilde{\mathcal{I}}_k, \quad k = 1, \ldots, p, \tag{7}$$

and

$$\mathbf{y}(k) = \mathbf{y}(k^+) = \begin{cases} \mathbf{x}^0, & \text{if } k = 0, \\ \mathbf{y}(k^-), & \text{if } k = 1, \ldots, p-1, \end{cases} \tag{8}$$

where $\theta_p := 1$, $(\boldsymbol{\theta}, \boldsymbol{\zeta}) \in \Theta \times \mathcal{Z}$ is a given pair, and the positive and negative superscripts denote limits from the right and left, respectively. By Theorem 3.1.6 of Ahmed (1988), there exists a unique solution of (7)-(8) corresponding to each $(\boldsymbol{\theta}, \boldsymbol{\zeta}) \in \Theta \times \mathcal{Z}$. Let $\mathbf{y}(\cdot|\boldsymbol{\theta}, \boldsymbol{\zeta})$ denote this solution.

For each $(\boldsymbol{\theta}, \boldsymbol{\zeta}) \in \Theta \times \mathcal{Z}$, define

$$S^p(\boldsymbol{\theta}, \boldsymbol{\zeta}) := \inf\{\, s > 0 : \ \mathbf{y}(s|\boldsymbol{\theta}, \boldsymbol{\zeta}) \in \Omega \,\}, \tag{9}$$

where Ω is the stopping surface defined in Section 2. If $S^p(\boldsymbol{\theta}, \boldsymbol{\zeta})$ is finite, then clearly

$$\mathbf{y}(s|\boldsymbol{\theta}, \boldsymbol{\zeta})\big|_{s = S^p(\boldsymbol{\theta}, \boldsymbol{\zeta})} \in \Omega. \tag{10}$$

We will show in the next subsection that $S^p(\boldsymbol{\theta}, \boldsymbol{\zeta})$ defined by (9) is indeed finite.

Now, define a new cost function $\tilde{J}^p : \Theta \times \mathcal{Z} \to \mathbb{R}$ as follows:

$$\tilde{J}^p(\boldsymbol{\theta}, \boldsymbol{\zeta}) = \Psi(\mathbf{y}(s|\boldsymbol{\theta}, \boldsymbol{\zeta}))\big|_{s = S^p(\boldsymbol{\theta}, \boldsymbol{\zeta})}.$$

We consider the following optimization problem.

Problem 3. Find a pair $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*) \in \Theta \times \mathcal{Z}$ such that

$$\tilde{J}^p(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*) = \inf_{(\boldsymbol{\theta}, \boldsymbol{\zeta}) \in \Theta \times \mathcal{Z}} \tilde{J}^p(\boldsymbol{\theta}, \boldsymbol{\zeta}).$$

We will show in the next subsection that Problem 3 is equivalent to Problem 2. This means that a solution of Problem 2 can be generated from a solution of Problem 3, and vice versa. The main virtue of Problem 3 is that its governing dynamic system (7)-(8) has fixed switching times at $s = 1, \dots, p-1$. This makes Problem 3 much easier to solve than Problem 2.

4.2 Relationship Between Problems 2 and 3

We will now establish several results linking Problem 2 with Problem 3. Our discussion culminates with a proof of the *Equivalence Theorem*—a fundamental result showing that a solution of Problem 2 can be easily obtained from a solution of Problem 3, and vice versa.

First, for each $\boldsymbol{\theta} \in \Theta$, define a *time-scaling function* $\mu(\cdot|\boldsymbol{\theta}) : [0, \infty) \to \mathbb{R}$ as follows:

$$\mu(s|\boldsymbol{\theta}) := \begin{cases} \displaystyle\sum_{j=1}^{\lfloor s \rfloor} \theta_j + \theta_{\lfloor s \rfloor + 1}(s - \lfloor s \rfloor), & \text{if } s \in [0, p-1), \\ \displaystyle\sum_{j=1}^{p-1} \theta_j + \theta_p(s - p + 1), & \text{if } s \in [p-1, \infty), \end{cases}$$

where $\lfloor \cdot \rfloor$ denotes the floor function and $\theta_p := 1$ (recall that $\Theta \subset \mathbb{R}^{p-1}$).

It's easy to see that $\mu(\cdot|\boldsymbol{\theta})$ is a piecewise linear function. Some other important properties of $\mu(\cdot|\boldsymbol{\theta})$ are stated in the following lemma.

Lemma 4.1. For each $\boldsymbol{\theta} \in \Theta$, the corresponding time-scaling function $\mu(\cdot|\boldsymbol{\theta})$ has the following properties:

(a) $\mu(0|\boldsymbol{\theta}) = 0$ and $\mu(k|\boldsymbol{\theta}) = \theta_1 + \cdots + \theta_k$, $k = 1, \dots, p$.
(b) $\mu(\cdot|\boldsymbol{\theta})$ is non-negative and $\mu(s|\boldsymbol{\theta}) > 0$ for all $s > 0$.
(c) $\mu(\cdot|\boldsymbol{\theta})$ is non-decreasing.
(d) $\mu(\cdot|\boldsymbol{\theta})$ is continuous.
(e) $\dot{\mu}(s|\boldsymbol{\theta}) = \theta_k$, $s \in \tilde{\mathcal{I}}_k$, $k = 1, \dots, p$.

Proof. Follows easily from the definition of $\mu(\cdot|\boldsymbol{\theta})$. □

Now, define a vector-valued function $\tilde{\boldsymbol{\tau}} : \Theta \to \mathbb{R}^{p-1}$ as follows:

$$\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}) := \left[\mu(1|\boldsymbol{\theta}), \dots, \mu(p-1|\boldsymbol{\theta})\right]^\top, \quad \boldsymbol{\theta} \in \Theta.$$

It follows from Lemma 4.1(b,c) that

$$\mu(k|\boldsymbol{\theta}) \geq 0, \quad k = 1, \dots, p-1,$$

and

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

Thus, $\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}) \in \mathcal{T}$. This implies that the state trajectory $\mathbf{x}^p(\cdot|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}), \boldsymbol{\zeta})$ is well-defined. Since $\mu(\cdot|\boldsymbol{\theta})$ is non-negative, we may substitute $t = \mu(s|\boldsymbol{\theta})$ into $\mathbf{x}^p(\cdot|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}), \boldsymbol{\zeta})$. Our first major result shows that applying this substitution yields $\mathbf{y}(\cdot|\boldsymbol{\theta}, \boldsymbol{\zeta})$, the solution of the switched system (7)-(8).

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

$$\mathbf{y}(s|\boldsymbol{\theta}, \boldsymbol{\zeta}) = \mathbf{x}^p(t|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}), \boldsymbol{\zeta})\big|_{t=\mu(s|\boldsymbol{\theta})}, \quad s \geq 0.$$

Proof. Let $(\boldsymbol{\theta}, \boldsymbol{\zeta}) \in \Theta \times \mathcal{Z}$ be arbitrary but fixed. Define

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

Since both $\mu(\cdot|\boldsymbol{\theta})$ and $\mathbf{x}^p$ are continuous, $\tilde{\mathbf{x}}^p$ is also continuous. Thus,

$$\tilde{\mathbf{x}}^p(k) = \tilde{\mathbf{x}}^p(k^+) = \tilde{\mathbf{x}}^p(k^-), \quad k = 1, \dots, p-1. \qquad (11)$$

Furthermore, by Lemma 4.1(a) and equation (2),

$$\tilde{\mathbf{x}}^p(0) = \tilde{\mathbf{x}}^p(0^+) = \mathbf{x}^0. \qquad (12)$$

Let $k \in \{1, \dots, p\}$ be a fixed integer. Since $\boldsymbol{\theta} \in \Theta$ and $\theta_p = 1$, either $\theta_k = 0$ or $\theta_k > 0$. If $\theta_k = 0$, then

$$\mu(s|\boldsymbol{\theta}) = \mu(k-1|\boldsymbol{\theta}), \quad s \in \tilde{\mathcal{I}}_k,$$

and thus

$$\tilde{\mathbf{x}}^p(s) = \tilde{\mathbf{x}}^p(k-1), \quad s \in \tilde{\mathcal{I}}_k.$$

This implies that

$$\dot{\tilde{\mathbf{x}}}^p(s) = \mathbf{0} = \theta_k \mathbf{f}(\tilde{\mathbf{x}}^p(s), \boldsymbol{\zeta}^k), \quad s \in \tilde{\mathcal{I}}_k. \qquad (13)$$

On the other hand, if $\theta_k > 0$ then $\mu(\cdot|\boldsymbol{\theta})$ is strictly increasing on $\tilde{\mathcal{I}}_k$ and

$$\mu(s|\boldsymbol{\theta}) \in \mathcal{I}_k(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta})), \quad s \in \tilde{\mathcal{I}}_k.$$

Thus, differentiating $\tilde{\mathbf{x}}^p$ using the chain rule and then applying Lemma 4.1(e) gives

$$\dot{\tilde{\mathbf{x}}}^p(s) = \theta_k \mathbf{f}(\tilde{\mathbf{x}}^p(s), \boldsymbol{\zeta}^k), \quad s \in \tilde{\mathcal{I}}_k. \qquad (14)$$

Equations (11)-(14) show that $\tilde{\mathbf{x}}^p$ is the unique solution of (7)-(8). This yields $\tilde{\mathbf{x}}^p = \mathbf{y}(\cdot|\boldsymbol{\theta}, \boldsymbol{\zeta})$, as required. □

We now show that $S^p(\boldsymbol{\theta}, \boldsymbol{\zeta})$ defined by equation (9) is finite.

Theorem 4.2. For each $(\boldsymbol{\theta}, \boldsymbol{\zeta}) \in \Theta \times \mathcal{Z}$,

$$0 < S^p(\boldsymbol{\theta}, \boldsymbol{\zeta}) \in \mathbb{R}.$$

Proof. We first prove that $S^p(\boldsymbol{\theta},\boldsymbol{\zeta}) \in \mathbb{R}$. Note that $\mu(\cdot|\boldsymbol{\theta})$ is a surjection from $[0,\infty)$ to $[0,\infty)$. Thus, there exists a point $s' \in [0,\infty)$ such that

$$\mu(s'|\boldsymbol{\theta}) = T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}).$$

If $s' = 0$, then

$$T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}) = \mu(s'|\boldsymbol{\theta}) = \mu(0|\boldsymbol{\theta}) = 0,$$

which contradicts Assumption 2.3. Thus, $s' > 0$.

By Theorem 4.1 and inclusion (6),

$$\begin{aligned}\mathbf{y}(s'|\boldsymbol{\theta},\boldsymbol{\zeta}) &= \mathbf{x}^p(t|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta})\big|_{t=\mu(s'|\boldsymbol{\theta})} \\ &= \mathbf{x}^p(t|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta})\big|_{t=T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta})} \in \Omega.\end{aligned}$$

This shows that the set $\{\, s > 0 : \ \mathbf{y}(s|\boldsymbol{\theta},\boldsymbol{\zeta}) \in \Omega \,\}$ on the right-hand side of equation (9) contains at least one element, so $S^p(\boldsymbol{\theta},\boldsymbol{\zeta})$ is well-defined.

Now, it's clear that $S^p(\boldsymbol{\theta},\boldsymbol{\zeta})$ is non-negative. Suppose that $S^p(\boldsymbol{\theta},\boldsymbol{\zeta}) = 0$. Then there exists a sequence $\{s_i\}_{i=1}^{\infty} \subset (0,\infty)$ such that $s_i \to 0+$ as $i \to \infty$ and

$$\mathbf{y}(s_i|\boldsymbol{\theta},\boldsymbol{\zeta}) \in \Omega, \quad i \geq 1.$$

Lemma 4.1(b) implies that

$$\mu(s_i|\boldsymbol{\theta}) > 0, \quad i \geq 1.$$

Furthermore, since $s_i \to 0+$ as $i \to \infty$, $\mu(s_i|\boldsymbol{\theta}) \to 0+$ as $i \to \infty$. Now, by Theorem 4.1,

$$\mathbf{x}^p(t|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta})\big|_{t=\mu(s_i|\boldsymbol{\theta})} = \mathbf{y}(s_i|\boldsymbol{\theta},\boldsymbol{\zeta}) \in \Omega, \quad i \geq 1.$$

Thus,

$$T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}) = \inf\{\, t > 0 : \ \mathbf{x}^p(t|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}) \in \Omega \,\} = 0.$$

But this contradicts Assumption 2.3. Hence, we must have $S^p(\boldsymbol{\theta},\boldsymbol{\zeta}) > 0$. □

Recall from Theorem 4.1 that the time-scaling function links the solution of the switched system (7)-(8) with the solution of the original system (1)-(2). Our next result shows that the time-scaling function maps $s = S^p(\boldsymbol{\theta},\boldsymbol{\zeta})$ in the new time horizon to $t = T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta})$ in the old time horizon.

Theorem 4.3. For each $(\boldsymbol{\theta},\boldsymbol{\zeta}) \in \Theta \times \mathcal{Z}$,

$$T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}) = \mu(s|\boldsymbol{\theta})\big|_{s=S^p(\boldsymbol{\theta},\boldsymbol{\zeta})}.$$

Proof. Let $(\boldsymbol{\theta},\boldsymbol{\zeta}) \in \Theta \times \mathcal{Z}$. For simplicity, we will write μ instead of $\mu(\cdot|\boldsymbol{\theta})$ and S^p instead of $S^p(\boldsymbol{\theta},\boldsymbol{\zeta})$.

Recall from Theorem 4.2 that $S^p > 0$. Thus, by Lemma 4.1(b),

$$\mu(S^p) > 0. \tag{15}$$

Furthermore, by Theorem 4.1 and inclusion (10),

$$\mathbf{x}^p(t|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta})\big|_{t=\mu(S^p)} = \mathbf{y}(S^p|\boldsymbol{\theta},\boldsymbol{\zeta}) \in \Omega. \tag{16}$$

Together, (15) and (16) imply that

$$T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}) \leq \mu(S^p). \tag{17}$$

Now, recall that μ is a surjection from $[0,\infty)$ to $[0,\infty)$. Hence, as in the proof of Theorem 4.2, there exists a point $s' \in (0,\infty)$ such that $\mu(s') = T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta})$. By Theorem 4.1 and inclusion (6),

$$\begin{aligned}\mathbf{y}(s'|\boldsymbol{\theta},\boldsymbol{\zeta}) &= \mathbf{x}^p(t|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta})\big|_{t=\mu(s')} \\ &= \mathbf{x}^p(t|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta})\big|_{t=T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta})} \in \Omega.\end{aligned}$$

Thus,

$$S^p \leq s'.$$

Therefore, since μ is non-decreasing,

$$\mu(S^p) \leq \mu(s') = T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}). \tag{18}$$

Combining inequalities (17) and (18) completes the proof. □

Together, Theorems 4.1 and 4.3 imply the following important result.

Theorem 4.4. For each $(\boldsymbol{\theta},\boldsymbol{\zeta}) \in \Theta \times \mathcal{Z}$,

$$\tilde{J}^p(\boldsymbol{\theta},\boldsymbol{\zeta}) = J^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}).$$

Proof. Recall from Theorem 4.3 that

$$T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}) = \mu(s|\boldsymbol{\theta})\big|_{s=S^p(\boldsymbol{\theta},\boldsymbol{\zeta})}.$$

Hence, by Theorem 4.1,

$$\begin{aligned}\mathbf{y}(s|\boldsymbol{\theta},\boldsymbol{\zeta})\big|_{s=S^p(\boldsymbol{\theta},\boldsymbol{\zeta})} &= \mathbf{x}^p(t|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta})\big|_{t=\mu(S^p(\boldsymbol{\theta},\boldsymbol{\zeta})|\boldsymbol{\theta})} \\ &= \mathbf{x}^p(t|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta})\big|_{t=T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta})}.\end{aligned}$$

Thus,

$$\begin{aligned}\tilde{J}^p(\boldsymbol{\theta},\boldsymbol{\zeta}) &= \Psi(\mathbf{y}(s|\boldsymbol{\theta},\boldsymbol{\zeta}))\big|_{s=S^p(\boldsymbol{\theta},\boldsymbol{\zeta})} \\ &= \Psi(\mathbf{x}^p(t|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}))\big|_{t=T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta})} \\ &= J^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}),\end{aligned}$$

as required. □

We are now ready to prove our main result showing that Problems 2 and 3 are equivalent.

Theorem 4.5 (Equivalence Theorem)**.** Let $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*)$ be a given pair in $\Theta \times \mathcal{Z}$. Then $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*)$ is optimal for Problem 3 if and only if $(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}^*), \boldsymbol{\zeta}^*)$ is optimal for Problem 2.

Proof. First, suppose that $(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}^*), \boldsymbol{\zeta}^*)$ is an optimal solution for Problem 2. Let $(\boldsymbol{\theta}, \boldsymbol{\zeta}) \in \Theta \times \mathcal{Z}$ be arbitrary but fixed. Then $(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}), \boldsymbol{\zeta}) \in \mathcal{T} \times \mathcal{Z}$. By Theorem 4.4,

$$\tilde{J}^p(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*) = J^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}^*), \boldsymbol{\zeta}^*) \le J^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}), \boldsymbol{\zeta}) = \tilde{J}^p(\boldsymbol{\theta}, \boldsymbol{\zeta}).$$

Hence, $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*)$ is optimal for Problem 3.

Conversely, suppose that $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*) \in \Theta \times \mathcal{Z}$ is an optimal solution for Problem 3. Let $(\boldsymbol{\tau}, \boldsymbol{\zeta}) \in \mathcal{T} \times \mathcal{Z}$ be fixed. We consider two cases: (i) $\boldsymbol{\tau} = \mathbf{0}$; and (ii) $\boldsymbol{\tau} \neq \mathbf{0}$.

For case (i), define vectors $\bar{\boldsymbol{\theta}} = [\bar{\theta}_1, \ldots, \bar{\theta}_{p-1}]^\top \in \mathbb{R}^{p-1}$ and $\bar{\boldsymbol{\zeta}} = (\bar{\boldsymbol{\zeta}}^1, \ldots, \bar{\boldsymbol{\zeta}}^p) \in \mathbb{R}^{pr}$ as follows:

$$\bar{\theta}_k := 1, \quad k = 1, \ldots, p-1,$$

and

$$\bar{\boldsymbol{\zeta}}^k := \boldsymbol{\zeta}^p, \quad k = 1, \ldots, p.$$

Then clearly $(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\zeta}}) \in \Theta \times \mathcal{Z}$. Furthermore, since $\tau_k = 0$ for each $k = 1, \ldots, p-1$,

$$\mathbf{u}^p(t|\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}), \bar{\boldsymbol{\zeta}}) = \mathbf{u}^p(t|\boldsymbol{\tau}, \boldsymbol{\zeta}) = \boldsymbol{\zeta}^p, \quad t \ge 0.$$

Thus,

$$J^p(\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}), \bar{\boldsymbol{\zeta}}) = J^p(\boldsymbol{\tau}, \boldsymbol{\zeta}).$$

By Theorem 4.4,

$$\begin{aligned} J^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}^*), \boldsymbol{\zeta}^*) &= \tilde{J}^p(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*) \\ &\le \tilde{J}^p(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\zeta}}) \\ &= J^p(\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}), \bar{\boldsymbol{\zeta}}) \\ &= J^p(\boldsymbol{\tau}, \boldsymbol{\zeta}). \end{aligned} \tag{19}$$

Now, for case (ii), there exists a $k \in \{1, \ldots, p-1\}$ such that $\tau_k > 0$. Let

$$\varsigma := \min\{\, k : \ \tau_k > 0 \,\}.$$

Furthermore, define vectors $\bar{\boldsymbol{\theta}} = [\bar{\theta}_1, \ldots, \bar{\theta}_{p-1}]^\top \in \mathbb{R}^{p-1}$ and $\bar{\boldsymbol{\zeta}} = (\bar{\boldsymbol{\zeta}}^1, \ldots, \bar{\boldsymbol{\zeta}}^p) \in \mathbb{R}^{pr}$ as follows:

$$\bar{\theta}_k := \begin{cases} \tau_\varsigma, & \text{if } k = 1, \\ \tau_{k+\varsigma-1} - \tau_{k+\varsigma-2}, & \text{if } k = 2, \ldots, p-\varsigma, \\ 1, & \text{if } k = p-\varsigma+1, \ldots, p-1, \end{cases}$$

and

$$\bar{\boldsymbol{\zeta}}^k := \begin{cases} \boldsymbol{\zeta}^{k+\varsigma-1}, & \text{if } k = 1, \ldots, p-\varsigma, \\ \boldsymbol{\zeta}^p, & \text{if } k = p-\varsigma+1, \ldots, p. \end{cases}$$

It's easy to see that $(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\zeta}}) \in \Theta \times \mathcal{Z}$.

Now, suppose $t \in \mathcal{I}_k(\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}))$ for some $k = 1, \ldots, p-\varsigma$. Then $t \in \mathcal{I}_{k+\varsigma-1}(\boldsymbol{\tau})$. Thus,

$$\begin{aligned} \mathbf{u}^p(t|\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}), \bar{\boldsymbol{\zeta}}) = \mathbf{u}^p(t|\boldsymbol{\tau}, \boldsymbol{\zeta}) = \boldsymbol{\zeta}^{k+\varsigma-1}, \\ t \in \mathcal{I}_k(\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}})), \ \ k = 1, \ldots, p-\varsigma. \end{aligned} \tag{20}$$

On the other hand, if $t \in \mathcal{I}_k(\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}))$ for some integer $k = p-\varsigma+1, \ldots, p$, then $t \ge \tau_{p-1}$. Hence,

$$\begin{aligned} \mathbf{u}^p(t|\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}), \bar{\boldsymbol{\zeta}}) = \mathbf{u}^p(t|\boldsymbol{\tau}, \boldsymbol{\zeta}) = \boldsymbol{\zeta}^p, \quad t \in \mathcal{I}_k(\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}})), \\ k = p-\varsigma+1, \ldots, p. \end{aligned} \tag{21}$$

Equations (20) and (21) imply that

$$J^p(\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}), \bar{\boldsymbol{\zeta}}) = J^p(\boldsymbol{\tau}, \boldsymbol{\zeta}).$$

Therefore, by Theorem 4.4,

$$\begin{aligned} J^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}^*), \boldsymbol{\zeta}^*) &= \tilde{J}^p(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*) \\ &\le \tilde{J}^p(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\zeta}}) \\ &= J^p(\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}), \bar{\boldsymbol{\zeta}}) \\ &= J^p(\boldsymbol{\tau}, \boldsymbol{\zeta}). \end{aligned} \tag{22}$$

Since $(\boldsymbol{\tau}, \boldsymbol{\zeta}) \in \mathcal{T} \times \mathcal{Z}$ was chosen arbitrarily, (19) and (22) show that $(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}^*), \boldsymbol{\zeta}^*)$ is optimal for Problem 2. □

It follows from the Equivalence Theorem that if $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*)$ is optimal for Problem 3, then $(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}^*), \boldsymbol{\zeta}^*)$ is optimal for Problem 2. The corresponding suboptimal control for Problem 1 is $\mathbf{u}^p(\cdot|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}^*), \boldsymbol{\zeta}^*)$, which has switching times at $t = \mu(k|\boldsymbol{\theta}^*)$, $k = 1, \ldots, p-1$. By Lemma 4.1(a),

$$\mu(k|\boldsymbol{\theta}^*) - \mu(k-1|\boldsymbol{\theta}^*) = \theta_k^*, \quad k = 1, \ldots, p-1.$$

Thus, θ_k^* is the time duration between the suboptimal control's $(k-1)$th and kth switching times.

5 Gradient Computation for Problem 3

Problem 3 is a nonlinear optimization problem in which the decision variables θ_k, $k = 1, \ldots, p-1$ and $\boldsymbol{\zeta}^k$, $k = 1, \ldots, p$ need to be chosen to minimize the cost function $\tilde{J}^p$. In this section, we will develop an algorithm for computing the partial derivatives of $\tilde{J}^p$. This algorithm can be combined with a nonlinear programming method to solve Problem 3 efficiently. A solution of Problem 2 can then be generated via the Equivalence Theorem.

First, define

$$\delta_{k,j} := \begin{cases} 1, & \text{if } k = j, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$\hat{\delta}_{k,j} := \begin{cases} 1, & \text{if } k \leq j, \\ 0, & \text{otherwise.} \end{cases}$$

Next, for each $k = 1, \ldots, p-1$, consider the following auxiliary switched system:

$$\begin{aligned} \dot{\boldsymbol{\psi}}^k(s) = \hat{\delta}_{k,j}\theta_j \frac{\partial \mathbf{f}(\mathbf{y}(s|\boldsymbol{\theta},\boldsymbol{\zeta}),\boldsymbol{\zeta}^j)}{\partial \mathbf{x}} \boldsymbol{\psi}^k(s) \\ + \delta_{k,j}\mathbf{f}(\mathbf{y}(s|\boldsymbol{\theta},\boldsymbol{\zeta}),\boldsymbol{\zeta}^j), \quad s \in \tilde{\mathcal{I}}_j, \end{aligned} \tag{23}$$

and

$$\boldsymbol{\psi}^k(j) = \begin{cases} \mathbf{0}, & \text{if } j = 0, \\ \boldsymbol{\psi}^k(j^-), & \text{if } j = 1, \ldots, p-1, \end{cases} \tag{24}$$

where $(\boldsymbol{\theta},\boldsymbol{\zeta}) \in \Theta \times \mathcal{Z}$. Let $\boldsymbol{\psi}^k(\cdot|\boldsymbol{\theta},\boldsymbol{\zeta})$ denote the unique absolutely continuous solution of (23)-(24).

For each $k = 1, \ldots, p$ and $i = 1, \ldots, r$, define another auxiliary system as follows:

$$\begin{aligned} \dot{\boldsymbol{\phi}}^{k,i}(s) = \hat{\delta}_{k,j}\theta_j \frac{\partial \mathbf{f}(\mathbf{y}(s|\boldsymbol{\theta},\boldsymbol{\zeta}),\boldsymbol{\zeta}^j)}{\partial \mathbf{x}} \boldsymbol{\phi}^{k,i}(s) \\ + \delta_{k,j}\theta_j \frac{\partial \mathbf{f}(\mathbf{y}(s|\boldsymbol{\theta},\boldsymbol{\zeta}),\boldsymbol{\zeta}^j)}{\partial u_i}, \quad s \in \tilde{\mathcal{I}}_j, \end{aligned} \tag{25}$$

and

$$\boldsymbol{\phi}^{k,i}(j) = \begin{cases} \mathbf{0}, & \text{if } j = 0, \\ \boldsymbol{\phi}^{k,i}(j^-), & \text{if } j = 1, \ldots, p-1, \end{cases} \tag{26}$$

where $(\boldsymbol{\theta},\boldsymbol{\zeta}) \in \Theta \times \mathcal{Z}$. Let $\boldsymbol{\phi}^{k,i}(\cdot|\boldsymbol{\theta},\boldsymbol{\zeta})$ denote the unique absolutely continuous solution of (25)-(26).

We will show later that the partial derivatives of $\tilde{J}^p$ can be computed by solving systems (23)-(24) and (25)-(26). To do this, we will need the following lemma, which follows immediately from the results in Loxton et al. (2008) and Vincent & Grantham (1981).

Lemma 5.1. Let $(\boldsymbol{\theta},\boldsymbol{\zeta}) \in \Theta \times \mathcal{Z}$. Then for each $s \geq 0$,

$$\frac{\partial \mathbf{y}(s|\boldsymbol{\theta},\boldsymbol{\zeta})}{\partial \theta_k} = \boldsymbol{\psi}^k(s|\boldsymbol{\theta},\boldsymbol{\zeta}), \quad k = 1, \ldots, p-1,$$

and

$$\frac{\partial \mathbf{y}(s|\boldsymbol{\theta},\boldsymbol{\zeta})}{\partial \zeta_i^k} = \boldsymbol{\phi}^{k,i}(s|\boldsymbol{\theta},\boldsymbol{\zeta}), \quad k = 1, \ldots, p, \quad i = 1, \ldots, r.$$

Define

$$\mathcal{F} := \{\, (\boldsymbol{\theta},\boldsymbol{\zeta}) \in \Theta \times \mathcal{Z} : \; S^p(\boldsymbol{\theta},\boldsymbol{\zeta}) > p-1 \,\}.$$

We impose the following assumption on Problem 1.

Assumption 5.1. For each admissible control $\mathbf{u} \in \mathcal{U}$,

$$\frac{\partial \Phi(\mathbf{x}(t|\mathbf{u}))}{\partial \mathbf{x}} \mathbf{f}(\mathbf{x}(t|\mathbf{u}),\mathbf{u}(t))\Big|_{t=T(\mathbf{u})} \neq 0,$$

where $\Phi : \mathbb{R}^n \to \mathbb{R}$ is the function defining the stopping surface in Section 2.

Let $\bar{\Phi}(t) := \Phi(\mathbf{x}(t|\mathbf{u}))$. Then

$$\dot{\bar{\Phi}}(t) = \frac{d}{dt}\{\Phi(\mathbf{x}(t|\mathbf{u}))\} = \frac{\partial \Phi(\mathbf{x}(t|\mathbf{u}))}{\partial \mathbf{x}} \mathbf{f}(\mathbf{x}(t|\mathbf{u}),\mathbf{u}(t)).$$

Assumption 5.1 requires that $\dot{\bar{\Phi}} \neq 0$ at the terminal time $t = T(\mathbf{u})$. In other words, we require that the state trajectory does not approach the stopping surface tangentially.

We now show that, under Assumption 5.1, the partial derivatives of S^p exist at every point in $\mathcal{F}$.

Theorem 5.1. For each $(\boldsymbol{\theta},\boldsymbol{\zeta}) \in \mathcal{F}$,

$$\begin{aligned} \frac{\partial S^p}{\partial \theta_k} = -\left\{ \frac{\partial \Phi(\mathbf{y}(S^p|\boldsymbol{\theta},\boldsymbol{\zeta}))}{\partial \mathbf{x}} \boldsymbol{\psi}^k(S^p|\boldsymbol{\theta},\boldsymbol{\zeta}) \right\} \\ \div \left\{ \frac{\partial \Phi(\mathbf{y}(S^p|\boldsymbol{\theta},\boldsymbol{\zeta}))}{\partial \mathbf{x}} \mathbf{f}(\mathbf{y}(S^p|\boldsymbol{\theta},\boldsymbol{\zeta}),\boldsymbol{\zeta}^p) \right\}, \end{aligned}$$

and

$$\begin{aligned} \frac{\partial S^p}{\partial \zeta_i^k} = -\left\{ \frac{\partial \Phi(\mathbf{y}(S^p|\boldsymbol{\theta},\boldsymbol{\zeta}))}{\partial \mathbf{x}} \boldsymbol{\phi}^{k,i}(S^p|\boldsymbol{\theta},\boldsymbol{\zeta}) \right\} \\ \div \left\{ \frac{\partial \Phi(\mathbf{y}(S^p|\boldsymbol{\theta},\boldsymbol{\zeta}))}{\partial \mathbf{x}} \mathbf{f}(\mathbf{y}(S^p|\boldsymbol{\theta},\boldsymbol{\zeta}),\boldsymbol{\zeta}^p) \right\}, \end{aligned}$$

where for simplicity we write S^p instead of $S^p(\boldsymbol{\theta},\boldsymbol{\zeta})$.

Proof. Define a function $G : (p-1,\infty) \times \mathbb{R}^{p-1} \times \mathbb{R}^{pr} \to \mathbb{R}$ as follows:

$$G(\gamma,\mathbf{v},\mathbf{w}) := \Phi(\mathbf{y}(\gamma|\mathbf{v},\mathbf{w})).$$

Clearly, $(S^p,\boldsymbol{\theta},\boldsymbol{\zeta}) \in (p-1,\infty) \times \mathbb{R}^{p-1} \times \mathbb{R}^{pr}$ and

$$G(S^p,\boldsymbol{\theta},\boldsymbol{\zeta}) = 0.$$

By the implicit function theorem (see Nocedal & Wright (2006)), S^p has partial derivatives at $(\boldsymbol{\theta},\boldsymbol{\zeta})$ if G is continuously differentiable and $\partial G/\partial \gamma$ is non-zero at $(S^p,\boldsymbol{\theta},\boldsymbol{\zeta})$. We now show that these two conditions hold.

Differentiating G with respect to γ yields

$$\begin{aligned}\frac{\partial G(\gamma,\mathbf{v},\mathbf{w})}{\partial\gamma} &= \frac{\partial\Phi(\mathbf{y}(\gamma|\mathbf{v},\mathbf{w}))}{\partial\mathbf{x}}\dot{\mathbf{y}}(\gamma|\mathbf{v},\mathbf{w}) \\ &= \frac{\partial\Phi(\mathbf{y}(\gamma|\mathbf{v},\mathbf{w}))}{\partial\mathbf{x}}\mathbf{f}(\mathbf{y}(\gamma|\mathbf{v},\mathbf{w}),\mathbf{w}^p), \quad (27)\end{aligned}$$

where the last equality follows from equation (7) with $k = p$ (recall that $\theta_p = 1$).

By Lemma 5.1,

$$\begin{aligned}\frac{\partial G(\gamma,\mathbf{v},\mathbf{w})}{\partial v_k} &= \frac{\partial\Phi(\mathbf{y}(\gamma|\mathbf{v},\mathbf{w}))}{\partial\mathbf{x}}\frac{\partial\mathbf{y}(\gamma|\mathbf{v},\mathbf{w})}{\partial v_k} \\ &= \frac{\partial\Phi(\mathbf{y}(\gamma|\mathbf{v},\mathbf{w}))}{\partial\mathbf{x}}\boldsymbol{\psi}^k(\gamma|\mathbf{v},\mathbf{w}) \quad (28)\end{aligned}$$

and

$$\begin{aligned}\frac{\partial G(\gamma,\mathbf{v},\mathbf{w})}{\partial w_i^k} &= \frac{\partial\Phi(\mathbf{y}(\gamma|\mathbf{v},\mathbf{w}))}{\partial\mathbf{x}}\frac{\partial\mathbf{y}(\gamma|\mathbf{v},\mathbf{w})}{\partial w_i^k} \\ &= \frac{\partial\Phi(\mathbf{y}(\gamma|\mathbf{v},\mathbf{w}))}{\partial\mathbf{x}}\boldsymbol{\phi}^{k,i}(\gamma|\mathbf{v},\mathbf{w}). \quad (29)\end{aligned}$$

Equations (27)-(29) show that G is continuously differentiable.

Now, since $\mu(\cdot|\boldsymbol{\theta})$ is strictly increasing on $(p-1,\infty)$, it follows from Theorem 4.3 that

$$\mu(p-1|\boldsymbol{\theta}) < \mu(S^p|\boldsymbol{\theta}) = T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}).$$

Thus,

$$T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}) \in \mathcal{I}_p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta})). \quad (30)$$

By Theorems 4.1 and 4.3,

$$\begin{aligned}\mathbf{y}(S^p|\boldsymbol{\theta},\boldsymbol{\zeta}) &= \mathbf{x}^p(t|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta})\big|_{t=\mu(S^p|\boldsymbol{\theta})} \\ &= \mathbf{x}^p(t|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta})\big|_{t=T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta})}. \quad (31)\end{aligned}$$

From (27), (30), (31), and Assumption 5.1, we obtain

$$\begin{aligned}\frac{\partial G(S^p,\boldsymbol{\theta},\boldsymbol{\zeta})}{\partial\gamma} &= \frac{\partial\Phi(\mathbf{y}(S^p|\boldsymbol{\theta},\boldsymbol{\zeta}))}{\partial\mathbf{x}}\mathbf{f}(\mathbf{y}(S^p|\boldsymbol{\theta},\boldsymbol{\zeta}),\boldsymbol{\zeta}^p) \\ &= \frac{\partial\Phi(\mathbf{x}^p(t|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}))}{\partial\mathbf{x}}\mathbf{f}(\mathbf{x}^p(t|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}),\boldsymbol{\zeta}^p)\Big|_{t=T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta})} \\ &\neq 0,\end{aligned}$$

as required. Thus, S^p has partial derivatives at $(\boldsymbol{\theta},\boldsymbol{\zeta})$. The formulae for $\partial S^p/\partial\theta_k$ and $\partial S^p/\partial\zeta_i^k$ can be easily obtained by differentiating $\Phi(\mathbf{y}(S^p|\boldsymbol{\theta},\boldsymbol{\zeta})) = 0$ and then applying Lemma 5.1. □

Now, by Lemma 5.1 and Theorem 5.1,

$$\begin{aligned}\frac{\partial\tilde{J}^p(\boldsymbol{\theta},\boldsymbol{\zeta})}{\partial\theta_k} &= \frac{\partial\Psi(\mathbf{y}(S^p|\boldsymbol{\theta},\boldsymbol{\zeta}))}{\partial\mathbf{x}}\boldsymbol{\psi}^k(S^p|\boldsymbol{\theta},\boldsymbol{\zeta}) \\ &\quad - \rho\frac{\partial\Phi(\mathbf{y}(S^p|\boldsymbol{\theta},\boldsymbol{\zeta}))}{\partial\mathbf{x}}\boldsymbol{\psi}^k(S^p|\boldsymbol{\theta},\boldsymbol{\zeta}),\end{aligned} \quad (32)$$

where

$$\begin{aligned}\rho := &\left\{\frac{\partial\Psi(\mathbf{y}(S^p|\boldsymbol{\theta},\boldsymbol{\zeta}))}{\partial\mathbf{x}}\mathbf{f}(\mathbf{y}(S^p|\boldsymbol{\theta},\boldsymbol{\zeta}),\boldsymbol{\zeta}^p)\right\} \\ &\div\left\{\frac{\partial\Phi(\mathbf{y}(S^p|\boldsymbol{\theta},\boldsymbol{\zeta}))}{\partial\mathbf{x}}\mathbf{f}(\mathbf{y}(S^p|\boldsymbol{\theta},\boldsymbol{\zeta}),\boldsymbol{\zeta}^p)\right\}.\end{aligned}$$

Similarly,

$$\begin{aligned}\frac{\partial\tilde{J}^p(\boldsymbol{\theta},\boldsymbol{\zeta})}{\partial\zeta_i^k} &= \frac{\partial\Psi(\mathbf{y}(S^p|\boldsymbol{\theta},\boldsymbol{\zeta}))}{\partial\mathbf{x}}\boldsymbol{\phi}^{k,i}(S^p|\boldsymbol{\theta},\boldsymbol{\zeta}) \\ &\quad - \rho\frac{\partial\Phi(\mathbf{y}(S^p|\boldsymbol{\theta},\boldsymbol{\zeta}))}{\partial\mathbf{x}}\boldsymbol{\phi}^{k,i}(S^p|\boldsymbol{\theta},\boldsymbol{\zeta}).\end{aligned} \quad (33)$$

Equations (32) and (33) are only applicable when $(\boldsymbol{\theta},\boldsymbol{\zeta}) \in \mathcal{F}$. However, this is not a major restriction because, as we now show, any pair outside $\mathcal{F}$ can be projected onto $\mathcal{F}$ without sacrificing cost.

Theorem 5.2. Let $(\boldsymbol{\theta},\boldsymbol{\zeta}) \notin \mathcal{F}$ be a given pair. Then there exists a corresponding pair $(\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\zeta}}) \in \mathcal{F}$ such that

$$\tilde{J}^p(\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\zeta}}) = \tilde{J}^p(\boldsymbol{\theta},\boldsymbol{\zeta}).$$

Proof. Since $0 < S^p(\boldsymbol{\theta},\boldsymbol{\zeta}) \leq p-1$, there exists an integer $\varsigma \in \{1,\ldots,p-1\}$ such that $S^p(\boldsymbol{\theta},\boldsymbol{\zeta}) \in (\varsigma-1,\varsigma]$. We consider two cases: (i) $\varsigma = 1$; and (ii) $\varsigma \geq 2$. For case (i), define $\bar{\boldsymbol{\theta}} \in \Theta$ and $\bar{\boldsymbol{\zeta}} \in \mathcal{Z}$ as follows:

$$\bar{\theta}_k := \frac{\mu(S^p(\boldsymbol{\theta},\boldsymbol{\zeta})|\boldsymbol{\theta})}{p}, \quad k = 1,\ldots,p-1, \quad (34)$$

and

$$\bar{\boldsymbol{\zeta}}^k := \boldsymbol{\zeta}^1, \quad k = 1,\ldots,p. \quad (35)$$

Recall that $\mu(\cdot|\boldsymbol{\theta})$ is increasing on $[0,1]$ because $\theta_1 > 0$. Thus, since $S^p(\boldsymbol{\theta},\boldsymbol{\zeta}) \in (0,1]$,

$$0 < \mu(S^p(\boldsymbol{\theta},\boldsymbol{\zeta})|\boldsymbol{\theta}) \leq \mu(1|\boldsymbol{\theta}) = \theta_1.$$

Therefore, by Theorem 4.3,

$$0 < T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}) \leq \theta_1. \quad (36)$$

But clearly,

$$\mathbf{u}^p(t|\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}),\bar{\boldsymbol{\zeta}}) = \mathbf{u}^p(t|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}) = \boldsymbol{\zeta}^1, \quad t \in [0,\theta_1). \quad (37)$$

Combining (36) and (37) gives

$$\mathbf{u}^p(t|\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}),\bar{\boldsymbol{\zeta}}) = \mathbf{u}^p(t|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}), \quad 0 \le t < T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}).$$

This implies

$$\mathbf{x}^p(t|\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}),\bar{\boldsymbol{\zeta}}) = \mathbf{x}^p(t|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}), \quad 0 \le t \le T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}),$$

and

$$T^p(\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}),\bar{\boldsymbol{\zeta}}) = T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}).$$

Consequently, by Theorem 4.4,

$$\tilde{J}^p(\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\zeta}}) = J^p(\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}),\bar{\boldsymbol{\zeta}}) = J^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}) = \tilde{J}^p(\boldsymbol{\theta},\boldsymbol{\zeta}).$$

We now show that $(\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\zeta}}) \in \mathcal{F}$. First, note that

$$\begin{aligned}\mu(p-1|\bar{\boldsymbol{\theta}}) &= \sum_{j=1}^{p-1}\bar{\theta}_j \\ &< \mu(S^p(\boldsymbol{\theta},\boldsymbol{\zeta})|\boldsymbol{\theta}) \\ &= T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}) \\ &= T^p(\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}),\bar{\boldsymbol{\zeta}}) \\ &= \mu(S^p(\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\zeta}})|\bar{\boldsymbol{\theta}}).\end{aligned}$$

Since $\mu(\cdot|\bar{\boldsymbol{\theta}})$ is non-decreasing, this implies

$$p-1 < S^p(\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\zeta}}).$$

Thus, $(\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\zeta}})$ defined by (34) and (35) is the required pair.

We now consider case (ii) when $\varsigma \ge 2$. Suppose that $\theta_\varsigma = 0$. Then it follows from equation (7) with $k = \varsigma$ that

$$\dot{\mathbf{y}}(s|\boldsymbol{\theta},\boldsymbol{\zeta}) = \mathbf{0}, \qquad s \in (\varsigma-1,\varsigma). \tag{38}$$

Thus, by inclusion (10),

$$\mathbf{y}(\varsigma-1|\boldsymbol{\theta},\boldsymbol{\zeta}) = \mathbf{y}(s|\boldsymbol{\theta},\boldsymbol{\zeta})\big|_{s=S^p(\boldsymbol{\theta},\boldsymbol{\zeta})} \in \Omega.$$

Since $\varsigma - 1 > 0$, this implies

$$S^p(\boldsymbol{\theta},\boldsymbol{\zeta}) \le \varsigma - 1 < S^p(\boldsymbol{\theta},\boldsymbol{\zeta}),$$

which is a contradiction. Hence, θ_ς must be strictly positive. This means that $\mu(\cdot|\boldsymbol{\theta})$ is strictly increasing on the interval $[\varsigma-1,\varsigma]$.

Now, define $\bar{\boldsymbol{\theta}} \in \Theta$ and $\bar{\boldsymbol{\zeta}} \in \mathcal{Z}$ as follows:

$$\bar{\theta}_k := \begin{cases} \dfrac{\theta_1}{p-\varsigma+1}, & \text{if } k = 1,\ldots,p-\varsigma+1, \\ \theta_{k-p+\varsigma}, & \text{if } k = p-\varsigma+2,\ldots,p-1, \end{cases} \tag{39}$$

and

$$\bar{\boldsymbol{\zeta}}^k := \begin{cases} \boldsymbol{\zeta}^1, & \text{if } k = 1,\ldots,p-\varsigma+1, \\ \boldsymbol{\zeta}^{k-p+\varsigma}, & \text{if } k = p-\varsigma+2,\ldots,p. \end{cases} \tag{40}$$

Recall that $S^p(\boldsymbol{\theta},\boldsymbol{\zeta}) \in (\varsigma-1,\varsigma]$. Thus, since $\mu(\cdot|\boldsymbol{\theta})$ is strictly increasing on $[\varsigma-1,\varsigma]$,

$$\begin{aligned}\mu(\varsigma-1|\boldsymbol{\theta}) &< \mu(S^p(\boldsymbol{\theta},\boldsymbol{\zeta})|\boldsymbol{\theta}) \\ &= T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}) \\ &\le \mu(\varsigma|\boldsymbol{\theta}) \\ &= \theta_1 + \cdots + \theta_\varsigma.\end{aligned} \tag{41}$$

We also have

$$\mathbf{u}^p(t|\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}),\bar{\boldsymbol{\zeta}}) = \mathbf{u}^p(t|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}),\ t \in [0,\theta_1+\cdots+\theta_\varsigma). \tag{42}$$

Combining (41) and (42) yields

$$\mathbf{u}^p(t|\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}),\bar{\boldsymbol{\zeta}}) = \mathbf{u}^p(t|\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}),\ 0 \le t < T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}).$$

Thus, as in the proof of case (i),

$$T^p(\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}),\bar{\boldsymbol{\zeta}}) = T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta})$$

and

$$\tilde{J}^p(\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\zeta}}) = \tilde{J}^p(\boldsymbol{\theta},\boldsymbol{\zeta}).$$

We now need to show that $(\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\zeta}}) \in \mathcal{F}$. Recalling again that $\varsigma - 1 < S^p(\boldsymbol{\theta},\boldsymbol{\zeta})$, we have

$$\begin{aligned}\mu(p-1|\bar{\boldsymbol{\theta}}) &= \sum_{k=1}^{p-\varsigma+1}\frac{\theta_1}{p-\varsigma+1} + \sum_{k=p-\varsigma+2}^{p-1}\theta_{k-p+\varsigma} \\ &= \sum_{k=1}^{\varsigma-1}\theta_k \\ &< \mu(S^p(\boldsymbol{\theta},\boldsymbol{\zeta})|\boldsymbol{\theta}) \\ &= T^p(\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}),\boldsymbol{\zeta}) \\ &= T^p(\tilde{\boldsymbol{\tau}}(\bar{\boldsymbol{\theta}}),\bar{\boldsymbol{\zeta}}) \\ &= \mu(S^p(\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\zeta}})|\bar{\boldsymbol{\theta}}).\end{aligned}$$

This shows that $p-1 < S^p(\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\zeta}})$ (recall that $\mu(\cdot|\boldsymbol{\theta})$ is non-decreasing). □

If $(\boldsymbol{\theta},\boldsymbol{\zeta}) \in \mathcal{F}$, then we can use equations (32) and (33) to compute the partial derivatives of $\tilde{J}^p$ at $(\boldsymbol{\theta},\boldsymbol{\zeta})$. However, if $(\boldsymbol{\theta},\boldsymbol{\zeta}) \notin \mathcal{F}$, then equations (32) and (33) are not applicable. In this case, we can use the method shown in the proof of Theorem 5.2 to generate a new point $(\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\zeta}}) \in \mathcal{F}$ having the same cost as $(\boldsymbol{\theta},\boldsymbol{\zeta})$. The partial derivatives of $\tilde{J}^p$ can then be computed at this new point.

We now propose the following algorithm for solving Problem 3.

Algorithm 5.1. Input an integer $p \ge 2$ and an initial pair $(\boldsymbol{\theta},\boldsymbol{\zeta}) \in \Theta \times \mathcal{Z}$.

1. If $S^p(\boldsymbol{\theta},\boldsymbol{\zeta}) \le p-1$, then find $\varsigma \in \{1,\ldots,p-1\}$ such that $S^p(\boldsymbol{\theta},\boldsymbol{\zeta}) \in (\varsigma-1,\varsigma]$. Otherwise, go to Step 4.

2. If $\varsigma = 1$, then define $(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\zeta}}) \in \mathcal{F}$ according to equations (34) and (35). Otherwise, define $(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\zeta}}) \in \mathcal{F}$ according to equations (39) and (40).
3. Set $(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\zeta}}) \to (\boldsymbol{\theta}, \boldsymbol{\zeta})$.
4. Compute $\partial \tilde{J}^p(\boldsymbol{\theta}, \boldsymbol{\zeta})/\partial \theta_k$ and $\partial \tilde{J}^p(\boldsymbol{\theta}, \boldsymbol{\zeta})/\partial \zeta_i^k$ using equations (32) and (33).
5. If $(\boldsymbol{\theta}, \boldsymbol{\zeta})$ is optimal, then stop. Otherwise, use the derivative information obtained in Step 4 to compute a descent direction.
6. Perform a line search along this direction to obtain a new pair $(\boldsymbol{\theta}', \boldsymbol{\zeta}') \in \Theta \times \mathcal{Z}$.
7. Set $(\boldsymbol{\theta}', \boldsymbol{\zeta}') \to (\boldsymbol{\theta}, \boldsymbol{\zeta})$ and return to Step 1.

Note that Steps 5 and 6 of Algorithm 5.1 can be implemented using a standard nonlinear programming method such as sequential quadratic programming or the conjugate gradient method (see Luenberger & Ye (2008) and Nocedal & Wright (2006)). After using Algorithm 5.1 to solve Problem 3, we can easily generate a solution of Problem 2 using the Equivalence Theorem.

6 Examples

We implemented Algorithm 5.1 as a Fortran program for solving Problem 2. Our program uses the optimization code NLPQLP (Schittkowski (2007)) to test optimality, compute the search direction, and perform the line search in Steps 5 and 6 of Algorithm 5.1. The differential equation solver LSODAR (Hindmarsh (1982)) is used to solve the state and auxiliary systems. Note that LSODAR automatically terminates once the stopping criterion is satisfied.

6.1 Van der Pol Oscillator

A Van der Pol oscillator can be modelled by the following dynamic system:

$$\dot{x}_1 = x_2, \tag{43a}$$

$$\dot{x}_2 = -x_1 + x_2(1 - x_1^2) + u, \tag{43b}$$

and

$$x_1(0) = 1, \quad x_2(0) = 1, \tag{44}$$

where the control $u : [0, \infty) \to \mathbb{R}$ is subject to the bound constraints

$$-1 \leq u(t) \leq 1, \quad t \geq 0. \tag{45}$$

We define the terminal time T for system (43)-(44) to be the first time at which the following stopping criterion is satisfied:

$$x_2(T) = 0. \tag{46}$$

Our optimal control problem is defined as follows: *Choose the control* $u : [0, \infty) \to \mathbb{R}$ *to maximize* $x_1(T)$ *subject to the dynamic system* (43)-(44)*, the control constraints* (45)*, and the stopping criterion* (46)*.*

Using our program with $p = 2$, we obtained an optimal terminal time of $T = 1.1699$ and an optimal objective function value of $x_1(T) = 1.6394$. The optimal control is shown in Figure 1; the corresponding state trajectories are shown in Figure 2. Note that the optimal control is almost constant.

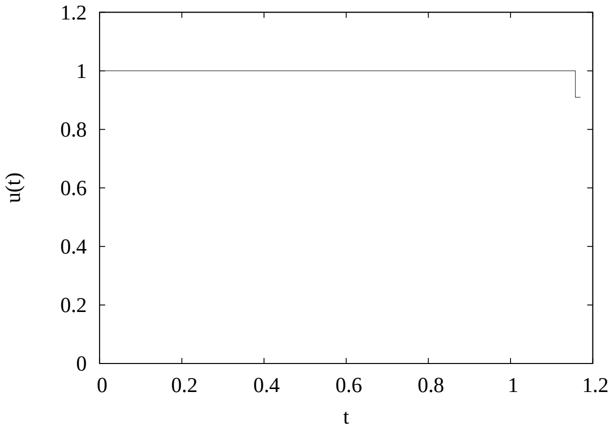

Fig. 1. Optimal control for Example 6.1.

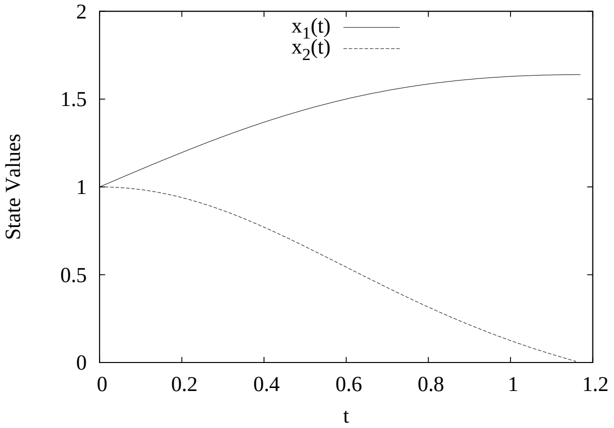

Fig. 2. Optimal state trajectories for Example 6.1.

We now compare our results with those produced by the software package MISER 3.3 (Jennings, Fisher, Teo & Goh (2004)). MISER works by partitioning the time horizon into multiple subintervals, and then approximating the control by a constant value on each subinterval. MISER assumes that the terminal time T is fixed. Thus, before applying MISER, we use the substitution $s = t/T$ to map the time horizon $[0, T]$ into the fixed interval $[0, 1]$. Equations (43)-(44) then become

$$\dot{x}_1 = Tx_2, \tag{47a}$$

$$\dot{x}_2 = T(-x_1 + x_2(1 - x_1^2) + u), \tag{47b}$$

and

$$x_1(0) = 1, \quad x_2(0) = 1, \tag{48}$$

where T is now a free decision variable. The control con-

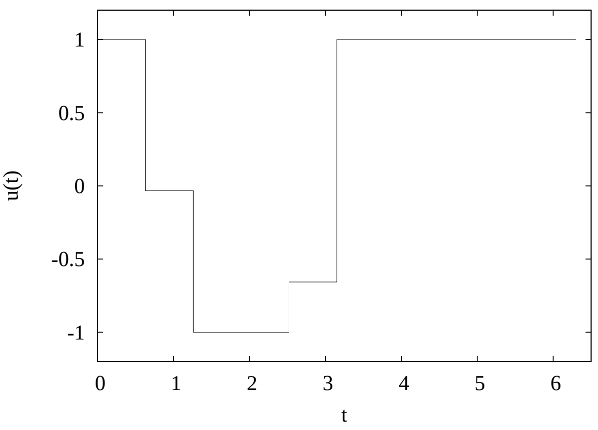

Fig. 3. Optimal control for Example 6.1 computed by MISER.

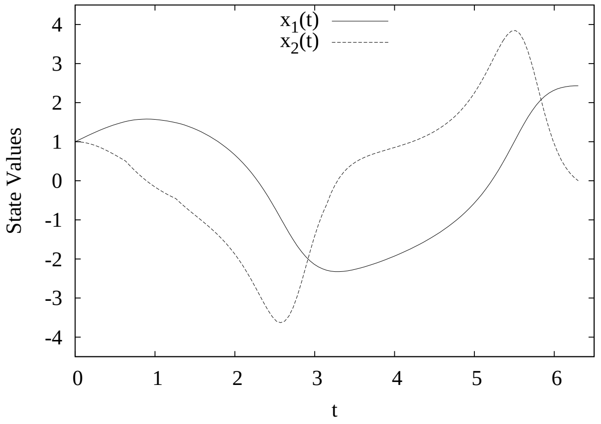

Fig. 4. State trajectories for Example 6.1 computed by MISER.

straints and stopping criterion become

$$-1 \leq u(s) \leq 1, \quad s \in [0, 1], \tag{49}$$

and

$$x_2(1) = 0. \tag{50}$$

The new optimal control problem with fixed terminal time is defined as follows: *Choose* T *and* $u : [0, 1] \to \mathbb{R}$ *to maximize* $x_1(1)$ *subject to the dynamic system* (47)-(48)*, the control constraints* (49)*, and the terminal state constraint* (50)*.*

Using MISER 3.3 with 10 subintervals and an initial guess of $u = 0$, we obtained an optimal terminal time of $T = 6.3013$ and an optimal objective function value of $x_1(T) = 2.4320$. The optimal control is shown in Figure 3, while the corresponding state trajectories are shown in Figure 4. Note that the control and state trajectories are plotted with respect to the original time variable t.

We see from Figure 4 that $T = 6.3013$ is *not* the first time at which $x_2 = 0$; we also have $x_2 = 0$ near $t = 1$ and $t = 3$. Thus, the state trajectory produced by MISER violates the stopping criterion (46). This is because MISER views $x_2(T) = 0$ as a boundary condition, rather than a stopping condition. In fact, the classical optimal control methods underpinning MISER cannot guarantee that T is the first time at which the stopping criterion is satisfied. Consequently, such methods produce invalid results when applied to this problem. Algorithm 5.1 produces the correct optimal solution.

6.2 Time-optimal Control of a Stirred Tank Mixer

Lee et al. (1997) have considered the following dynamic model for a stirred tank mixer:

$$\dot{x}_1 = \frac{(1 - x_1)u_1 + (2 - x_1)u_2}{x_2}, \tag{51a}$$

$$\dot{x}_2 = -0.02\sqrt{x_2} + u_1 + u_2, \tag{51b}$$

and

$$x_1(0) = 0.8, \quad x_2(0) = 0.7, \tag{52}$$

where x_1 is the output concentration, x_2 is the volume of liquid in the tank, and u_1 and u_2 are input flow rates.

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

$$0 \leq u_1(t) \leq 0.03, \quad 0 \leq u_2(t) \leq 0.01, \quad t \geq 0. \tag{53}$$

Furthermore, the desired terminal state is

$$x_1(T) = 1.25, \quad x_2(T) = 1. \tag{54}$$

The aim is to choose the input flow rates appropriately so that the system is transferred from the initial state (52) to the terminal state (54) in minimum time. Thus, we have the following time-optimal control problem: *Choose the input flow rates* u_1 *and* u_2 *to minimize the final time* T *subject to the dynamic system* (51)-(52)*, the control constraints* (53) *and the terminal state constraints* (54)*.*

We now show how to convert this time-optimal control problem into the form of Problem 1. First, since the terminal value of x_1 is 1.25, we define the stopping time T to be the first time at which

$$x_1(T) = 1.25. \tag{55}$$

We also introduce a new state variable x_3, where

$$\dot{x}_3(t) = 1, \quad x_3(0) = 0. \tag{56}$$

Clearly, $x_3(t) = t$.

The objective function should penalize *both* the final time and the deviation of the second state from its desired terminal value (note that the first state will always

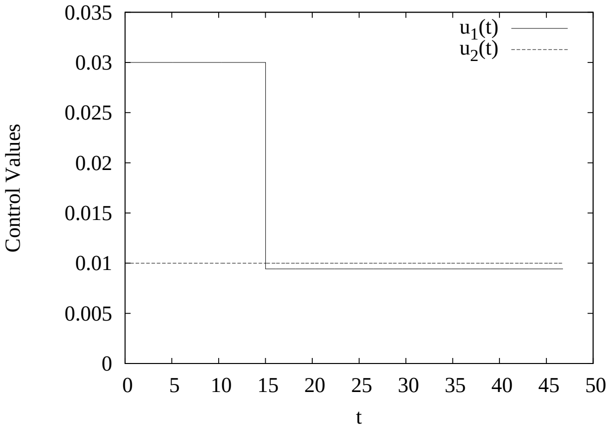

Fig. 5. Optimal controls for Example 6.2.

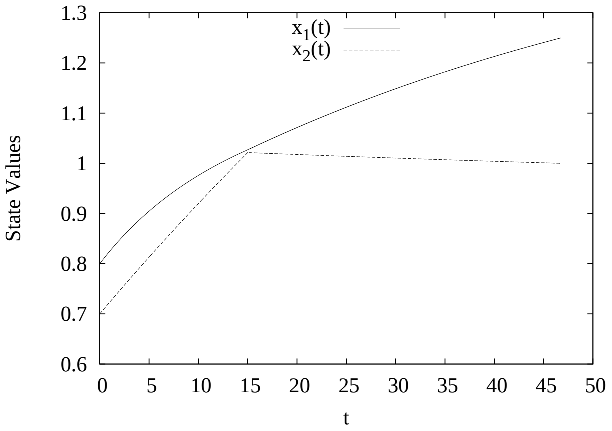

Fig. 6. Optimal state trajectories for Example 6.2.

be equal to 1.25 at the terminal time). Thus, we introduce the following objective function:

$$J = T + \gamma(x_2(T) - 1)^2 = x_3(T) + \gamma(x_2(T) - 1)^2,$$

where $\gamma > 0$ is a large penalty parameter.

On this basis, the time-optimal control problem given above can be converted into the following problem: *Choose the input flow rates u_1 and u_2 to minimize J subject to the dynamic system* (51)-(52) *and* (56)*, the control constraints* (53)*, and the stopping criterion* (55). This new problem is in the form of Problem 1 and can be solved using Algorithm 5.1.

Using our Fortran program with $p = 4$ and $\gamma = 10^6$, we solved this problem to obtain a minimum terminal time of $T = 46.7754$. This is similar to the terminal time reported in Lee et al. (1997). The optimal controls and optimal state trajectories are shown in Figures 5 and 6, respectively. Note that the optimal control only has one switch, even through $p - 1 = 3$ switches are allowed.

6.3 Optimal Control of a Hang Glider – Part I

Bulirsch, Nerz, Pesch & von Stryk (1993) consider the problem of maximizing the range of a hang glider in the presence of a thermal updraft. The glider's motion is described by the following differential equations:

$$\dot{x}_1 = x_3, \tag{57a}$$

$$\dot{x}_2 = x_4, \tag{57b}$$

$$\dot{x}_3 = \frac{1}{m}(-L\sin\eta - D\cos\eta), \tag{57c}$$

$$\dot{x}_4 = \frac{1}{m}(L\cos\eta - D\sin\eta) - g, \tag{57d}$$

where $m := 100$ is the mass of the glider (kg), $g := 9.8$ is the gravitational acceleration (ms^{-2}), and the functions η, L, and D are defined by

$$\begin{aligned} &\sin(\eta) = \frac{x_4 - \alpha(x_1)}{v(x_1, x_3, x_4)}, \quad \cos(\eta) = \frac{x_3}{v(x_1, x_3, x_4)}, \\ &L = \tfrac{1}{2}\gamma S u v(x_1, x_3, x_4)^2, \\ &D = \tfrac{1}{2}\gamma S(c_0 + c_1 u^2) v(x_1, x_3, x_4)^2, \\ &v(x_1, x_3, x_4) = \sqrt{x_3^2 + (x_4 - \alpha(x_1))^2}. \end{aligned}$$

Here, $\alpha(x_1)$ is the velocity profile of the thermal updraft. For a stable airmass, $\alpha(x_1) = 0$. The constants in the model are defined as follows:

$$\gamma := 1.13, \quad S := 14, \quad c_0 := 0.034, \quad c_1 := 0.069662.$$

In equations (57), x_1 is the glider's horizontal position (m), x_2 is the glider's altitude (m), x_3 is the glider's horizontal speed (ms^{-1}), x_4 is the glider's vertical speed (ms^{-1}), and u is a control function representing the lift coefficient.

The initial conditions for the state variables are:

$$\begin{aligned} x_1(0) = 0, \quad x_2(0) = 1000, \\ x_3(0) = 13.23, \quad x_4(0) = -1.288. \end{aligned} \tag{58}$$

Furthermore, the lift coefficient is subject to the following bound constraints:

$$0 \le u(t) \le 1.4, \quad t \ge 0. \tag{59}$$

We consider the glider's trajectory as it descends from an altitude of 1000 metres to 900 metres. Thus, the terminal time T is the first time at which the following stopping criterion is satisfied:

$$x_2(T) = 900. \tag{60}$$

Our range maximization problem is defined as follows: *Choose the lift coefficient u to maximize the range $x_1(T)$*

subject to the dynamic system (57)-(58)*, the control constraints* (59)*, and the stopping criterion* (60)*.*

Vanderbei (2001) solved this problem analytically for the simple case when $\alpha(x_1) = 0$ (i.e. stable airmass). The optimal control is static in this case:

$$u(t) = \sqrt{\frac{c_0}{c_1}} = 0.69862, \quad t \geq 0. \qquad (61)$$

Bulirsch et al. (1993) consider a more complicated problem in which the glider experiences a thermal updraft located about 250 metres from its initial position. The velocity profile of the updraft is

$$\alpha(x_1) = 2.5(1 - \beta(x_1)) \exp(-\beta(x_1)), \qquad (62)$$

where

$$\beta(x_1) = (0.01x_1 - 2.5)^2.$$

In the presence of this updraft, an uncontrolled glider ($u = 0$) will achieve a range of only $x_1(T) = 54.05$ with a corresponding terminal time of $T = 4.63$.

To solve the range maximization problem, we first ran our Fortran program with $p = 2$ and an initial guess of $u = 0$. We then re-solved the problem with $p = 3$, using the optimal solution for $p = 2$ as the initial guess. We repeated this procedure for $p = 4$, $p = 5$, and $p = 6$. The $p = 5$ and $p = 6$ solutions are almost identical, so larger values of p are unlikely to yield improved results. The optimal terminal time is $T = 102.45$ and the corresponding maximum range is $x_1(T) = 1240.41$. The optimal control is shown in Figure 7. The optimal gliding trajectory and corresponding speed plots are shown in Figures 8 and 9, respectively. Note that further improvements may be possible with a piecewise linear approximation for the lift coefficient.

We simulated system (57)-(58) for $\alpha(x_1)$ given by (62) and $u(t)$ given by the analytical solution in (61). The resulting flight trajectory is shown in Figure 10. This trajectory achieves a final range of $x_1(T) = 1199.86$, which is less than the range achieved by Algorithm 5.1. The corresponding terminal time is $T = 90.74$.

The control shown in Figure 7 is an *open-loop control.* We conclude this example by considering the problem of designing an optimal *feedback control.* To this end, we assume that the control is a linear function of the total speed, where the total speed is given by

$$s(t) = \sqrt{x_3^2(t) + x_4^2(t)}.$$

Thus,

$$\begin{aligned} u(t) &= ks(t) \\ &= k\sqrt{x_3(t)^2 + x_4(t)^2}, \quad t \geq 0, \qquad (63) \end{aligned}$$

where k is a feedback gain constant. Such control laws are commonly used for gliders.

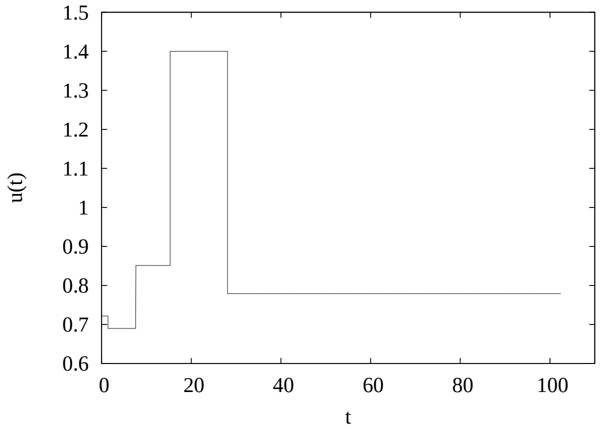

Fig. 7. Optimal lift coefficient for Example 6.3.

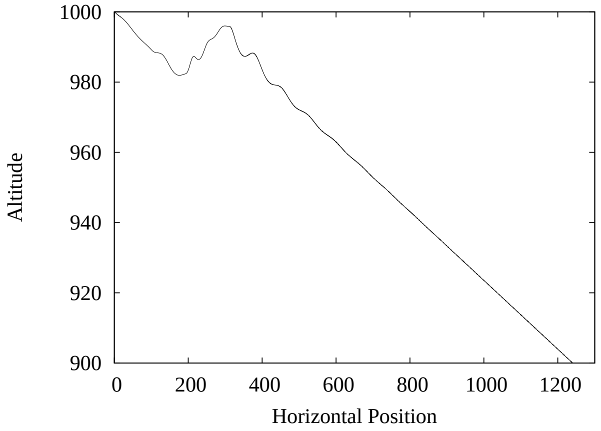

Fig. 8. Optimal flight trajectory for Example 6.3.

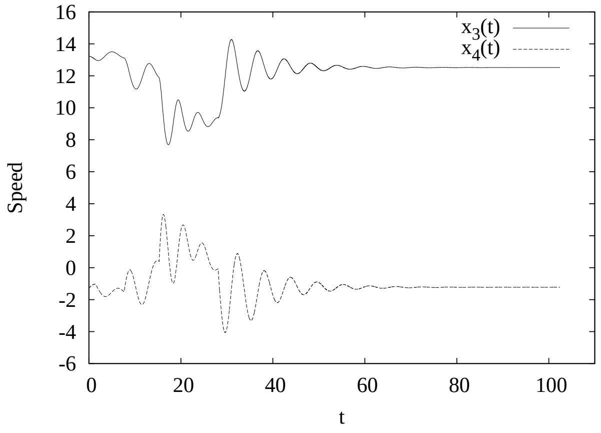

Fig. 9. Speed plots for Example 6.3.

When the lift coefficient u is defined by (63), L and D

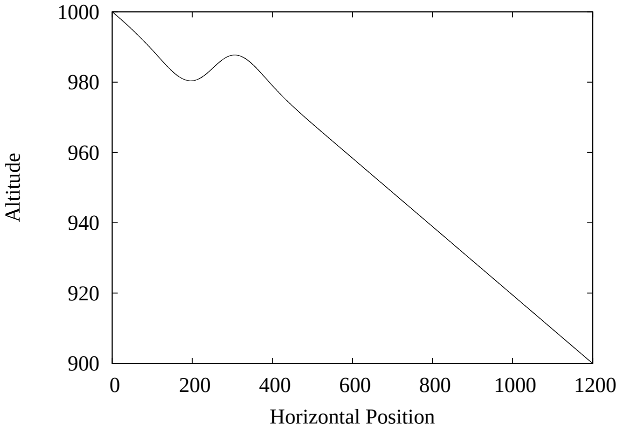

Fig. 10. Flight trajectory corresponding to the analytical solution (61) in Example 6.3.

in equations (57) become:

$$\begin{aligned} L &= \tfrac{1}{2}\gamma S k v(x_1, x_3, x_4)^2 \sqrt{x_3^2 + x_4^2}, \\ D &= \tfrac{1}{2}\gamma S(c_0 + c_1 k^2 x_3^2 + c_1 k^2 x_4^2) v(x_1, x_3, x_4)^2, \end{aligned} \tag{64}$$

where k is now a decision variable to be chosen optimally. Our goal is to choose k judiciously so that the glider's range is maximized. To ensure that the bounds (59) are satisfied, we impose the following constraints:

$$0 \leq k \leq 0.2. \tag{65}$$

We now consider the following optimal feedback control problem: *Choose the feedback gain constant k to maximize the range $x_1(T)$ subject to the dynamic system* (57)-(58) *with L and D defined by* (64)*, the constraints* (65)*, and the stopping criterion* (60)*.*

Algorithm 5.1 can be easily modified to solve the problem given above. In fact, this problem is simpler than the optimal control problems considered previously (its decision variable is a real number rather than a function). Using a modified version of Algorithm 5.1, we obtained an optimal feedback gain constant of $k = 0.06435$. The corresponding optimal terminal time is $T = 98.05$ and the maximum range is $x(T) = 1211.82$. The optimal feedback control is shown in Figure 11.

6.4 Optimal Control of a Hang Glider – Part II

Lin et al. (2011), Teo et al. (1989), and Teo et al. (1987) consider the following model for a hang glider:

$$\dot{x}_1 = x_3 \cos(x_4), \tag{66a}$$

$$\dot{x}_2 = x_3 \sin(x_4), \tag{66b}$$

$$\dot{x}_3 = -(k_1 + k_2 u^2) x_3^2 - g \sin(x_4), \tag{66c}$$

$$\dot{x}_4 = k_3 x_3 u - \frac{g}{x_3} \cos(x_4), \tag{66d}$$

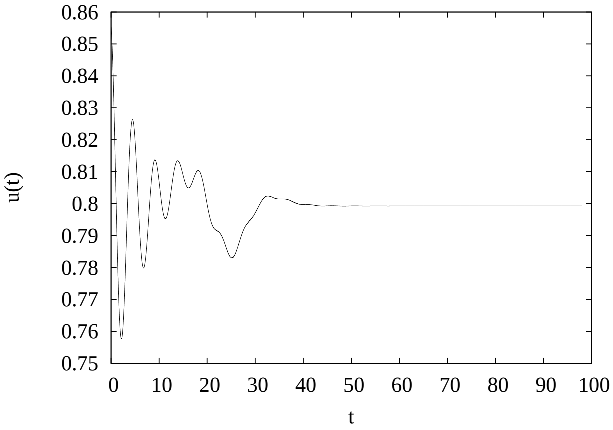

Fig. 11. Optimal feedback control for Example 6.3.

and

$$x_1(0) = x_2(0) = 0, \;\; x_3(0) = 370, \;\; x_4(0) = 1.5, \tag{67}$$

where x_1 is the glider's horizontal position (m), x_2 is the glider's altitude (m), x_3 is the glider's speed (ms^{-1}), x_4 is the angle between the glider's velocity vector and the horizon (rad), u is the glider's angle of attack (rad), $g := 9.8$ is the gravitational acceleration (ms^{-2}), and the constants k_1, k_2, and k_3 are defined by

$$\begin{aligned} k_1 &:= 3.289 \times 10^{-5}, \\ k_2 &:= 1.133 \times 10^{-3}, \\ k_3 &:= 3.289 \times 10^{-3}. \end{aligned}$$

The angle of attack (the control variable) is subject to the following bound constraints:

$$-0.2 \leq u(t) \leq 0.2, \quad t \geq 0. \tag{68}$$

Let T denote the glider's crash time. Then $T > 0$ is the first time at which the following stopping criterion is satisfied:

$$x_2(T) = 0. \tag{69}$$

This equation defines a stopping surface for (66)-(67).

Simulating system (66)-(67) for the uncontrolled case when $u = 0$ results in a terminal time of $T = 68.35$ and a final range of $x_1(T) = 1477.53$.

We now consider the following optimal control problem: *Choose the angle of attack $u : [0, \infty) \to \mathbb{R}$ to maximize the range $x_1(T)$ subject to the dynamics* (66)-(67)*, the control constraints* (68)*, and the stopping criterion* (69)*.* We refer to this problem as Problem A.

We used our Fortran program to solve Problem A for $p = 2, 3, 4, 5, 6$. As in Example 6.3, the optimal solution for each value of p was used as the starting point for the next value of p. The optimal terminal time for

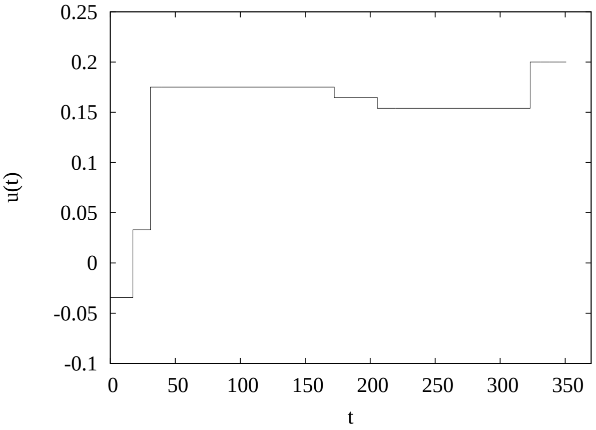

Fig. 12. Optimal control for Problem A in Example 6.4.

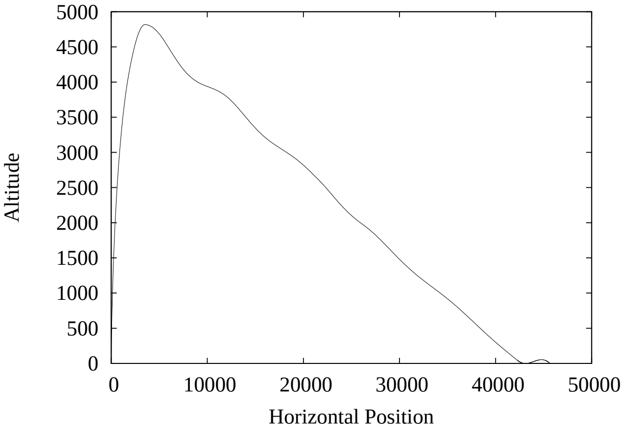

Fig. 13. Optimal trajectory for Problem A in Example 6.4.

$p = 6$ is $T = 350.79$ and the corresponding maximum range is $x_1(T) = 45,650.27$. The optimal angle of attack is shown in Figure 12 and the optimal state variables are shown in Figures 13-15. Note that, as expected, the flight trajectory produced by Algorithm 5.1 has significantly longer range than the uncontrolled trajectory. Note also that Lin et al. (2011), Teo et al. (1989), and Teo et al. (1987) report ranges of over 47,000 metres, but the optimal controls in these references violate the bound constraints (68). These constraints are an essential part of the problem formulation, as the angle of attack is always bounded in practice. Therefore, they cannot be omitted.

In Problem A, the glider's initial speed of 370 metres per second exceeds Mach 1. The same initial speed is used in Lin et al. (2011), Teo et al. (1989), and Teo et al. (1987). Since the speed changes from supersonic to subsonic during the time horizon (see Figure 15), the aerodynamic parameters k_1, k_2, and k_3 in model (66)-(67) are unlikely to be constant. Thus, we now change the initial speed in Problem A from $x_3(0) = 370$ to $x_3(0) = 250$, which is less than Mach 1. The new range maximization problem is referred to as Problem B.

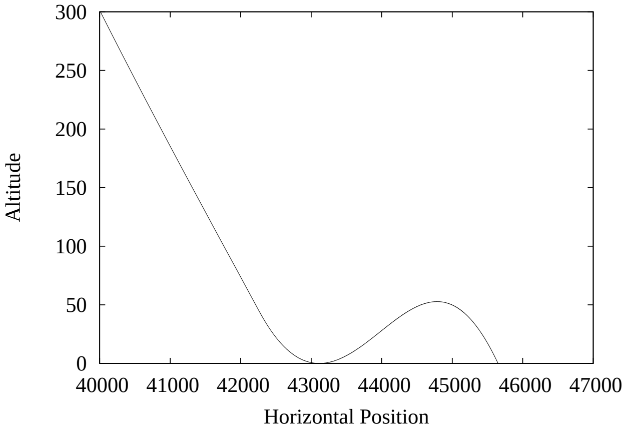

Fig. 14. The optimal trajectory in Figure 13 skims the ground near the terminal time.

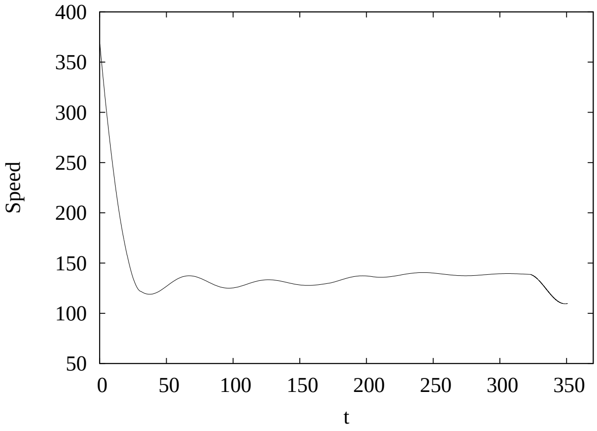

Fig. 15. Speed plot for Problem A in Example 6.4.

The dynamic system for Problem B is (66)-(67) with $x_3(0) = 370$ replaced by $x_3(0) = 250$. Simulating this system for the uncontrolled case when $u = 0$ results in a terminal time of $T = 48.49$ and a final range of $x_1(T) = 778.76$. We now use our Fortran program to solve Problem B for increasing values of p (as before, we use the solution for each value of p as the starting point for the next value of p). The optimal solution of Problem B for $p = 6$ has a terminal time of $T = 156.73$ and a maximum range of $x_1(T) = 20,006.89$. The optimal control is shown in Figure 16 and the corresponding flight trajectory is shown in Figures 17 and 18. The glider's speed is shown in Figure 19. Note that the speed is always subsonic (below Mach 1), and hence the assumption of constant aerodynamic parameters is more realistic here than for Problem A.

Figures 17 and 18 show that the optimal trajectory skims the ground near the terminal time (the glider's altitude is less than one centimetre at $t \approx 130$). This trajectory is clearly not robust, as small disturbances could easily cause the glider to crash prematurely. A more practical control scheme would keep the glider's altitude above

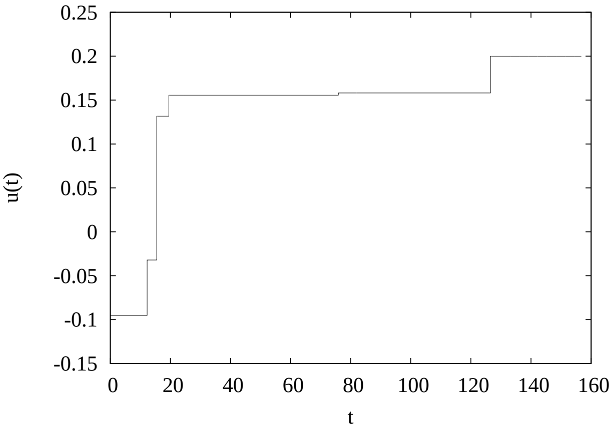

Fig. 16. Optimal control for Problem B in Example 6.4.

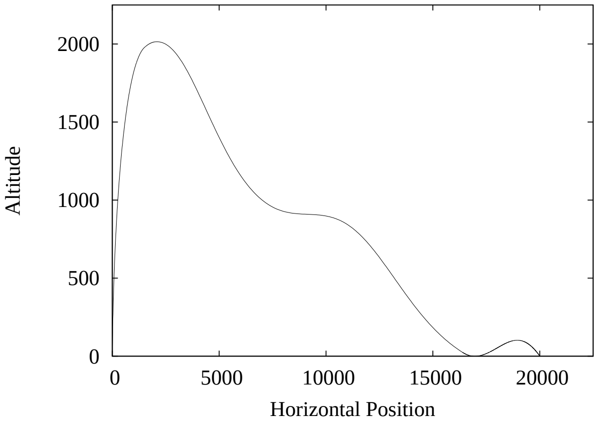

Fig. 17. Optimal trajectory for Problem B in Example 6.4.

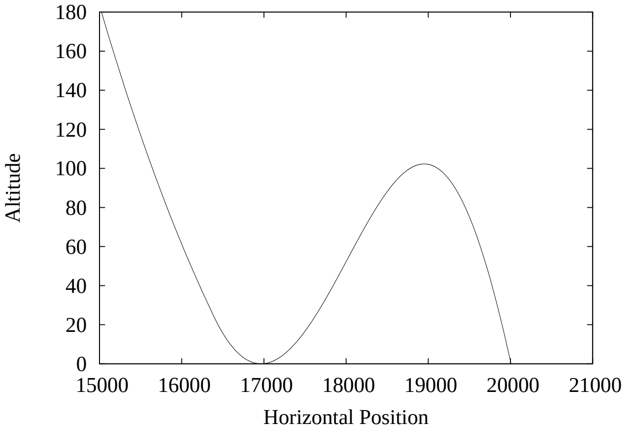

Fig. 18. The optimal trajectory in Figure 17 skims the ground near the terminal time.

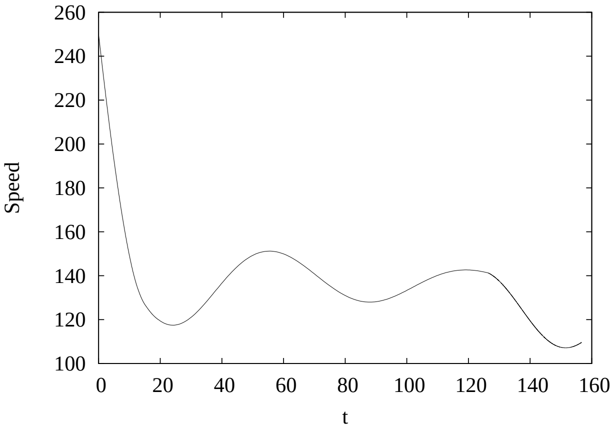

Fig. 19. Speed plot for Problem B in Example 6.4.

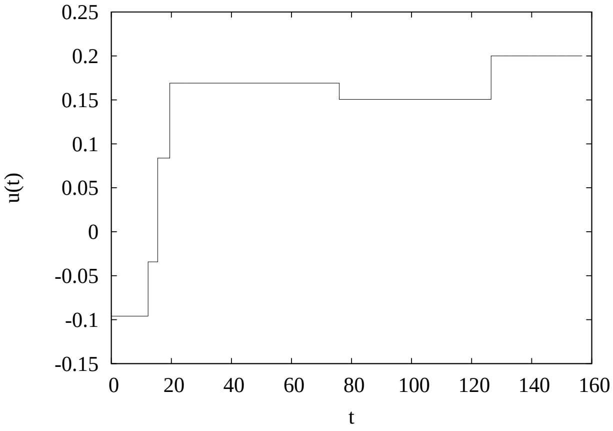

Fig. 20. Optimal control for Problem C in Example 6.4.

(say) 20 metres for as long as possible. Accordingly, we replace the stopping criterion (69) with

$$x_2(T) = 20. \tag{70}$$

As we will see, this stopping criterion prevents the glider from skimming the ground, as the control will try and keep the glider at least 20 metres above the ground for as long as possible. Note that T here is actually defined as the *second* time at which equation (70) is satisfied (otherwise, the flight trajectory will terminate shortly after launch).

Let Problem C refer to the modified range maximization problem with (70) as the new stopping criterion. We solved Problem C for $p = 6$, using the optimal solution of Problem B as the initial guess. The optimal terminal time is $T = 156.83$ and the corresponding maximum range is $x_1(T) = 19,987.63$. The optimal control is shown in Figure 20, and the corresponding flight trajectory is shown in Figures 21 and 22. Note that, unlike Problem B's optimal trajectory, Problem C's optimal trajectory is at least 20 metres above the ground when it dives and relaunches near the terminal time.

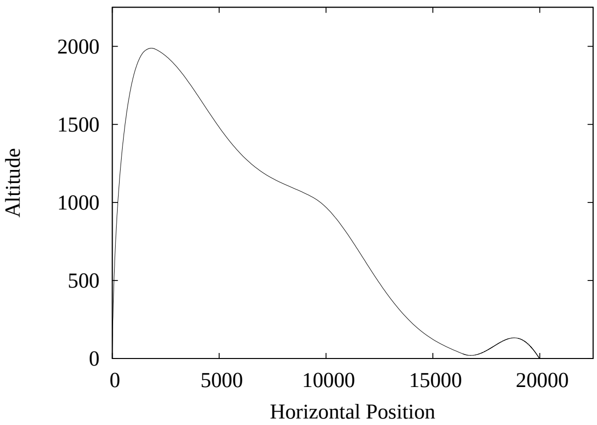

Fig. 21. Optimal trajectory for Problem C in Example 6.4.

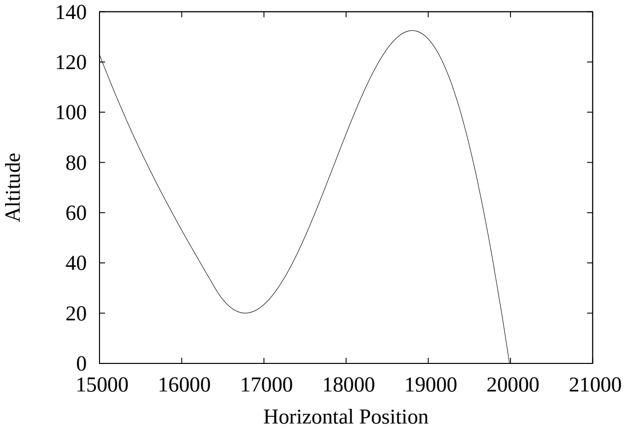

Fig. 22. Close-up of the end of the trajectory in Figure 21.

References

[1] N. U. Ahmed. *Elements of Finite-Dimensional Systems and Control Theory.* Longman Scientific and Technical, Essex, 1988.

[2] R. Bulirsch, E. Nerz, H. J. Pesch, and O. von Stryk. Combining direct and indirect methods in optimal control: Range maximization of a hang glider. In R. Bulirsch, A. Miele, J. Stoer, and K. H. Well, editors, *Optimal Control – Calculus of Variations, Optimal Control Theory and Numerical Methods*, volume 111, pages 273–288. Birkhäuser, 1993.

[3] M. Chyba, T. Haberkorn, S. B. Singh, R. N. Smith, and S. K. Choi. Increasing underwater vehicle autonomy by reducing energy consumption. *Ocean Engineering*, 36:62–73, 2009.

[4] 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*, 4:247–270, 2008.

[5] W. W. Hager. Runge-Kutta methods in optimal control and the transformed adjoint system. *Numerische Mathematik*, 87:247–282, 2000.

[6] A. C. Hindmarsh. Large ordinary differential equation systems and software. *IEEE Control Systems Magazine*, 2:24–30, 1982.

[7] L. S. Jennings, M. E. Fisher, K. L. Teo, and C. J. Goh. *MISER3 optimal control software: Theory and user manual.* University of Western Australia, Perth, July 2004.

[8] C. Y. Kaya and J. M. Martínez. Euler discretization and inexact restoration for optimal control. *Journal of Optimization Theory and Applications*, 134:191–206, 2007.

[9] C. Y. Kaya and J. L. Noakes. Computational method for time-optimal switching control. *Journal of Optimization Theory and Applications*, 117:69–92, 2003.

[10] H. W. J. Lee, K. L. Teo, V. Rehbock, and L. S. Jennings. Control parametrization enhancing technique for time optimal control problems. *Dynamic Systems and Applications*, 6:243–262, 1997.

[11] 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*, 7:63–81, 2011.

[12] R. Loxton, K. L. Teo, and V. Rehbock. Optimal control problems with multiple characteristic time points in the objective and constraints. *Automatica*, 44:2923–2929, 2008.

[13] D. G. Luenberger and Y. Ye. *Linear and Nonlinear Programming.* Springer, New York, 3rd edition, 2008.

[14] R. Luus. *Iterative Dynamic Programming.* Chapman and Hall, Boca Raton, 2000.

[15] J. Nocedal and S. J. Wright. *Numerical Optimization.* Springer, New York, 2nd edition, 2006.

[16] K. Schittkowski. *NLPQLP: A Fortran implementation of a sequential quadratic programming algorithm with distributed and non-monotone line search – User's guide, version 2.24.* University of Bayreuth, Bayreuth, June 2007.

[17] K. L. Teo, C. J. Goh, and C. C. Lim. A computational method for a class of dynamical optimization problems in which the terminal time is conditionally free. *IMA Journal of Mathematical Control and Information*, 6:81–95, 1989.

[18] K. L. Teo, C. J. Goh, and K. H. Wong. *A Unified Computational Approach to Optimal Control Problems.* Longman Scientific and Technical, Essex, 1991.

[19] K. L. Teo, G. Jepps, E. J. Moore, and S. Hayes. A computational method for free time optimal control problems, with application to maximizing the range of an aircraft-like projectile. *Journal of the Australian Mathematical Society – Series B*, 28:393–413, 1987.

[20] R. J. Vanderbei. Case studies in trajectory optimization: Trains, planes, and other pastimes. *Optimization and Engineering*, 2:215–243, 2001.

[21] T. L. Vincent and W. J. Grantham. *Optimality in Parametric Systems.* John Wiley, New York, 1981.

[22] O. von Stryk. Numerical solution of optimal control problems by direct collocation. In R. Bulirsch, A. Miele, J. Stoer, and K. H. Well, editors, *Optimal Control – Calculus of Variations, Optimal Control Theory and Numerical Methods*, volume 111, pages 129–143. Birkhäuser, 1993.