CHAPTER 3

OPTIMAL CONTROL

"What is now proved was once only imagined."

—William Blake.

3.1 INTRODUCTION

After more than three hundred years of evolution, optimal control theory has been formulated as an extension of the calculus of variations. Based on the theoretical foundation laid by several generations of mathematicians, optimal control has developed into a well-established research area and finds its applications in many scientific fields, ranging from mathematics and engineering to biomedical and management sciences. The *maximum principle*, developed in the late 1950s by Pontryagin and his coworkers [41], is among the biggest successes in optimal control. This principle as well as other results in optimal control apply to any problems of the calculus of variations discussed earlier in Chapter 2 (and gives equivalent results, as one would expect). This extension is most easily seen by considering the prototypical problem of the calculus of variations that consists of choosing a continuously differentiable function $\mathbf{x}(t)$, $t_1 \leq t \leq t_2$, to

$$\begin{aligned} \text{minimize:} \quad & \int_{t_1}^{t_2} \ell(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\, \mathrm{d}t \\ \text{subject to:} \quad & \mathbf{x}(t_1) \;=\; \mathbf{x}_1. \end{aligned}$$

Indeed, the above problem is readily transformed into the equivalent problem of finding a continuously differentiable function $\mathbf{u}(t)$, $t_1 \leq t \leq t_2$, to

$$\begin{aligned} \text{minimize:} \quad & \int_{t_1}^{t_2} \ell(t, \mathbf{x}(t), \mathbf{u}(t))\, \mathrm{d}t \\ \text{subject to:} \quad & \dot{\mathbf{x}}(t) \;=\; \mathbf{u}(t); \quad \mathbf{x}(t_1) \;=\; \mathbf{x}_1. \end{aligned}$$

Optimal control refers to this latter class of problems. In optimal control problems, variables are separated into two classes, namely the *state* (or *phase*) variables and the *control* variables. The evolution of the former is dictated by the latter, via a set of differential equations. Further, the control as well as the state variables are generally subject to constraints, which make many problems in optimal control *non-classical*, since problems with path constraints can hardly be handled in the classical calculus of variations. That is, the problem of optimal control can then be stated as: *"Determine the control signals that will cause a system to satisfy the physical constraints and, at the same time, minimize (or maximize) some performance criterion."* A precise mathematical formulation of optimal control problems shall be given in §3.2 below.

Despite its successes, however, optimal control theory is by no means complete, especially when it comes to the question of whether an optimal control exists for a given problems. The existence problem is of crucial importance, since it does not make much sense to seek a solution if none exists. As just an example, consider the problem of steering a system, from a prescribed initial state, to a fixed target, e.g., in minimum time. To find out whether an optimal control exists, one may start by investigating whether a *feasible* control can be found, i.e., one that satisfies the physical constraints. This latter question is closely related to system *controllability* in classical control theory, i.e., the ability to transfer the system from any initial state to any desired final state in a finite time. Should the system be uncontrollable, it is then likely that no successful control may be found for some initial states. And even though the system can be shown to be controllable, there may not exist an optimal control in the prescribed class of controls. The difficult problem of the existence of an optimal control shall be further discussed in §3.3.

Another important topic is to actually find an optimal control for a given problem, i.e., give a 'recipe' for operating the system in such a way that it satisfies the constraints in an optimal manner. Similar to the previous chapters on NLP and on the calculus of variations, our goal shall be to derive algebraic conditions that are either *necessary* or *sufficient* for optimality. These conditions are instrumental for singling out a small class of candidates for an optimal control. First, we shall investigate the application of variational methods to obtain necessary conditions of optimality for problems without state or control path constraints in §3.4; this will develop familiarity with the new notation and tools. Then, we shall consider methods based on so-called *maximum principles*, such as Pontryagin maximum principle, to address optimal control problems having path constraints in §3.5.

Finally, as most real-world problems are too complex to allow for an analytical solution, computational algorithms are inevitable in solving optimal control problems. As a result, several successful families of algorithms have been developed over the years. We shall present both *direct* and *indirect* approaches to solving optimal control problems in §3.6.

3.2 PROBLEM STATEMENT

The formulation of an optimal control problem requires several steps: the class of admissible controls is discussed in §3.2.1; the mathematical description (or model) of the system to be

controlled is considered in §3.2.2; the specification of a performance criterion is addressed in §3.2.3; then, the statement of physical constraints that should be satisfied is described in §3.2.4. Optimal criteria are discussed next in §3.2.5. Finally, we close the section with a discussion on open-loop and closed-loop optimal control laws in §3.2.6.

3.2.1 Admissible Controls

We shall consider the behavior of a system whose state at any instant of time is characterized by $n_x \geq 1$ real numbers $x_1, \ldots, x_{n_x}$ (for example, these may be coordinates and velocities). The vector space of the system under consideration is called the *phase* space. It is assumed that the system can be controlled, i.e., the system is equipped with *controllers* whose position dictates its future evolution. These controllers are characterized by points $\mathbf{u} = (u_1, \ldots, u_{n_u}) \in \mathbb{R}^{n_u}$, $n_u \geq 1$, namely the *control variables*.

In the vast majority of optimal control problems, the values that can be assumed by the control variables are restricted to a certain *control region* U, which may be any set in $\mathbb{R}^{n_u}$. In applications, the case where U is a closed region in $\mathbb{R}^{n_u}$ is important. For example, the control region U may be a *hypercube*,

$$|u_j| \leq 1, \quad j = 1, \ldots, n_u.$$

The physical meaning of choosing a closed and bounded control region is clear. The quantity of fuel being supplied to a motor, temperature, current, voltage, etc., which cannot take on arbitrarily large values, may serve as control variables. More general relations, such as

$$\phi(\mathbf{u}) = 0,$$

may also exist among the control variables.

We shall call every function $\mathbf{u}(\cdot)$, defined on some time interval $t_0 \leq t \leq t_\mathrm{f}$, a *control*. A control is an element of a (normed) linear space of real-vector-valued functions. Throughout this chapter, we shall consider the class of continuous controls or, more generally, piecewise continuous controls (see Fig. 3.1.):

Definition 3.1 (Piecewise Continuous Functions). *A real-valued function $u(t)$, $t_0 \leq t \leq t_\mathrm{f}$, is said to be* piecewise continuous*, denoted $u \in \hat{\mathcal{C}}[t_0, t_\mathrm{f}]$, if there is a finite (irreducible) partition $t_0 = \theta_0 < \theta_1 < \cdots < \theta_N < \theta_{N+1} = t_\mathrm{f}$ such that u may be regarded as a function in $\mathcal{C}[\theta_k, \theta_{k+1}]$ for each $k = 0, 1, \ldots, N$.*

That is, the class $\hat{\mathcal{C}}[t_0, t_\mathrm{f}]^{n_u}$ of n_u-dimensional vector-valued analogue of $\hat{\mathcal{C}}[t_0, t_\mathrm{f}]$, consists of those controls $\mathbf{u}$ with components $u_j \in \hat{\mathcal{C}}[t_0, t_\mathrm{f}]$, $j = 1, \ldots, n_u$. The discontinuities of one such control are by definition those of any of its components u_j.

Note that piecewise continuous controls correspond to the assumption of *inertia-less* controllers, since the values of $\mathbf{u}(t)$ may jump instantaneously when a discontinuity is met. This class of controls appears to be the most interesting for the practical applications of the theory, although existence of an optimal control is not guaranteed in general, as shall be seen later in §3.3.

The specification of the control region together with a class of controls leads naturally to the definition of an admissible control:

Definition 3.2 (Admissible Control). *A piecewise continuous control $\mathbf{u}(\cdot)$, defined on some time interval $t_0 \leq t \leq t_\mathrm{f}$, with range in the control region U,*

$$\mathbf{u}(t) \in U, \quad \forall t \in [t_0, t_\mathrm{f}],$$

is said to be an admissible control.

We shall denote by $\mathcal{U}[t_0, t_f]$ the class of admissible controls on $[t_0, t_f]$. It follows from Definition 3.1 that every admissible control $\mathbf{u} \in \mathcal{U}[t_0, t_f]$ is bounded.

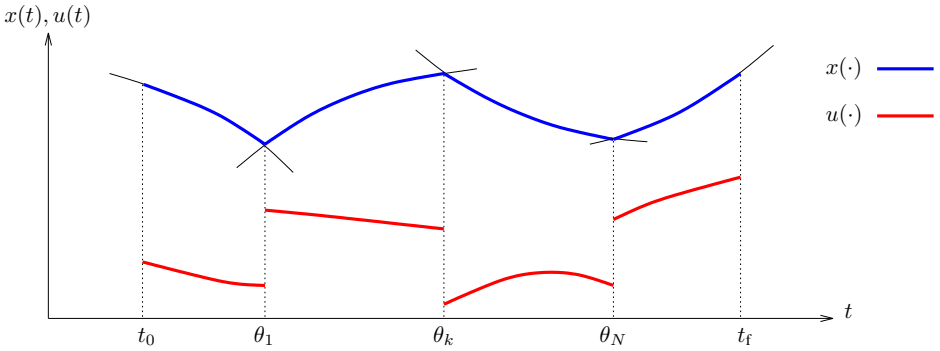


Figure 3.1. Illustration of a piecewise continuous control $u \in \hat{\mathcal{C}}[t_0, t_f]$ (red line), and the corresponding piecewise continuously differentiable response $x \in \hat{\mathcal{C}}^1[t_0, t_f]$ (blue line).

3.2.2 Dynamical System

A nontrivial part of any control problem is modeling the system. The objective is to obtain the *simplest* mathematical description that adequately predicts the response of the physical system to *all* admissible controls. We shall restrict our discussion herein to systems described by *ordinary differential equations* (ODEs) in state-space form,

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)); \quad \mathbf{x}(t_0) = \mathbf{x}_0. \tag{3.1}$$

Here, $t \in \mathbb{R}$ stands for the independent variable, usually called *time*; in the case where $\mathbf{f}$ does not depend explicitely on t, the system is said to be *autonomous*. The vector $\mathbf{u}(t) \in U$ represents the *control* (or *input* or *manipulated*) variables at time instant t. The vector $\mathbf{x}(t) \in \mathbb{R}^{n_x}$, $n_x \geq 1$, represents the *state* (or *phase*) variables, which characterize the behavior of the system at any time instant t. A solution $\mathbf{x}(t; \mathbf{x}_0, \mathbf{u}(\cdot))$ of (3.1) is called a *response* of the system, corresponding to the control $\mathbf{u}(\cdot)$, for the initial condition $\mathbf{x}(t_0) = \mathbf{x}_0$.

It shall also be assumed that $\mathbf{f}$ is continuous in the variables $t, \mathbf{x}, \mathbf{u}$ and continuously differentiable with respect to $\mathbf{x}$; in other words, the functions $\mathbf{f}(t, \mathbf{x}, \mathbf{u})$ and $\mathbf{f}_{\mathbf{x}}(t, \mathbf{x}, \mathbf{u}) := \frac{\partial \mathbf{f}}{\mathbf{x}}(t, \mathbf{x}, \mathbf{u})$ are defined and continuous, say on $[t_0, t_f] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u}$. This additional assumption ensures that a solution of (3.1) exists and is unique (at least locally) by Theorem A.46. [1] Further, the response $\mathbf{x}(t; \mathbf{x}_0, \mathbf{u}(\cdot))$ is piecewise continuously differentiable (see Definition 2.28, p. 82, and Fig. 3.1.) in its maximum interval of existence.

3.2.3 Performance Criterion

A *performance criterion* (also called *cost functional*, or simply *cost*) must be specified for evaluating the performance of a system quantitatively. By analogy to the problems of the

[1] See Appendix A.5 for a summary of local existence and uniqueness theorems for the solutions of nonlinear ODEs, as well as theorems on their continuous dependence and differentiability with respect to parameters.

calculus of variations (see §2.2.1, p. 63), the cost functional $\mathcal{J} : \mathcal{U}[t_0, t_\mathrm{f}] \to \mathbb{R}$ may be defined in the so-called *Lagrange form*,

$$\mathcal{J}(\mathbf{u}) := \int_{t_0}^{t_\mathrm{f}} \ell(t, \mathbf{x}(t), \mathbf{u}(t))\, \mathrm{d}t. \tag{3.2}$$

In this chapter, we shall assume that the Lagrangian $\ell(t, \mathbf{x}, \mathbf{u})$ is defined and continuous, together with its partial derivatives $\ell_{\mathbf{x}}(t, \mathbf{x}, \mathbf{u})$ on $\mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u}$. Moreover, either the initial time t_0 and final time t_f may be considered a fixed or a free variable in the optimization problem.

The objective functional may as well be specified in the *Mayer form*,

$$\mathcal{J}(\mathbf{u}) := \varphi(t_0, \mathbf{x}(t_0), t_\mathrm{f}, \mathbf{x}(t_\mathrm{f})), \tag{3.3}$$

with $\varphi : \mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R} \times \mathbb{R}^{n_x} \to \mathbb{R}$ being a real-valued function. Again, it shall be assumed throughout that $\varphi(t_0, \mathbf{x}_0, t_\mathrm{f}, \mathbf{x}_\mathrm{f})$ and $\varphi_{\mathbf{x}}(t_0, \mathbf{x}_0, t_\mathrm{f}, \mathbf{x}_\mathrm{f})$ exist and are continuous on $\mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R} \times \mathbb{R}^{n_x}$.

More generally, we may consider the cost functional in the *Bolza form*, which corresponds to the sum of an integral term and a terminal term as

$$\mathcal{J}(\mathbf{u}) := \varphi(t_0, \mathbf{x}(t_0), t_\mathrm{f}, \mathbf{x}(t_\mathrm{f})) + \int_{t_0}^{t_\mathrm{f}} \ell(t, \mathbf{x}(t), \mathbf{u}(t))\, \mathrm{d}t. \tag{3.4}$$

Interestingly enough, Mayer, Lagrange and Bolza problem formulations can be shown to be theoretically equivalent:

- Lagrange problems can be reduced to Mayer problems by introducing an additional state x_ℓ, the new state vector $\tilde{\mathbf{x}} := (x_\ell, x_1, \ldots, x_{n_x})^\mathsf{T}$, and an additional differential equation
$$\dot{x}_\ell(t) = \ell(t, \mathbf{x}(t), \mathbf{u}(t)); \quad x_\ell(t_0) = 0.$$
Then, the cost functional (3.2) is transformed into one of the Mayer form (3.3) with $\varphi(t_0, \tilde{\mathbf{x}}(t_0), t_\mathrm{f}, \tilde{\mathbf{x}}(t_\mathrm{f})) := x_\ell(t_\mathrm{f})$.

- Conversely, Mayer problems can be reduced to Lagrange problems by introducing an additional state variable x_ℓ, the new state vector $\tilde{\mathbf{x}} := (x_\ell, \mathbf{x}^\mathsf{T})^\mathsf{T}$, and an additional differential equation
$$\dot{x}_\ell(t) = 0; \quad x_\ell(t_0) = \frac{1}{t_\mathrm{f} - t_0} \varphi(t_0, \mathbf{x}(t_0), t_\mathrm{f}, \mathbf{x}(t_\mathrm{f})).$$
That is, the functional (3.3) can be rewritten in the Lagrange form (3.2) with $\ell(t, \tilde{\mathbf{x}}(t), \mathbf{u}(t)) := x_\ell(t)$.

- Finally, the foregoing transformations can be used to rewrite Bolza problems (3.4) in either the Mayer form or the Lagrange form, while it shall be clear that Mayer and Lagrange problems are special Bolza problems with $\ell(t, \tilde{\mathbf{x}}(t), \mathbf{u}(t)) := 0$ and $\varphi(t_0, \tilde{\mathbf{x}}(t_0), t_\mathrm{f}, \tilde{\mathbf{x}}(t_\mathrm{f})) := 0$, respectively.

3.2.4 Physical Constraints

A great variety of constraints may be imposed in an optimal control problem. These constraints restrict the range of values that can be assumed by both the control and the state variables. One usually distinguishes between *point constraints* and *path constraints*; optimal control problems may also contain *isoperimetric constraints*. All these constraints can be of equality or inequality type.

Point Constraints. These constraints are used routinely in optimal control problems, especially *terminal* constraints (i.e., point constraints defined at terminal time). As just an example, an inequality terminal constraint of the form

$$\psi(t_{\mathrm{f}}, \mathbf{x}(t_{\mathrm{f}})) \leq 0$$

may arise in a stabilization problems, e.g., for forcing the system's response to belong to a given target set at terminal time; another typical example is that of a process changeover where the objective is to bring the system from its actual steady state to a new steady state,

$$\psi'(t_{\mathrm{f}}, \mathbf{x}(t_{\mathrm{f}})) = 0.$$

Isoperimetric Constraints. Like problems of the calculus of variations, optimal control problems may have constraints involving the integral of a given functional over the time interval $[t_0, t_{\mathrm{f}}]$ (or some subinterval of it),

$$\int_{t_0}^{t_{\mathrm{f}}} h(t, \mathbf{x}(t), \mathbf{u}(t))\, \mathrm{d}t \leq C.$$

Clearly, a problem with isoperimetric constraints can be readily reformulated into an equivalent problem with point constraints only by invoking the transformation used previously in §3.2.3 for rewriting a Lagrange problem into the Mayer form.

Path Constraints. This last type of constraints is encountered in many optimal control problems. Path constraints may be defined for restricting the range of values taken by mixed functions of both the control and the state variables. Moreover, such restrictions can be imposed over the entire time interval $[t_0, t_{\mathrm{f}}]$ or any (nonempty) time subinterval, e.g., for safety reasons. For example, a path constraint could be define as

$$\phi(t, \mathbf{x}(t), \mathbf{u}(t)) \leq 0, \quad \forall t \in [t_0, t_{\mathrm{f}}],$$

hence restricting the points in phase space to a certain region $X \subset \mathbb{R}^{n_x}$ at all times. In general, a distinction is made between those path constraints depending explicitly on the control variables, and those depending only on the state variables ("pure" state constraints) such as

$$x_k(t) \ \leq \ x^U, \quad \forall t \in [t_0, t_{\mathrm{f}}],$$

for some $k \in \{1, \ldots, n_x\}$. This latter type of constraints being much more problematic to handle.

Constrained optimal control problems lead naturally to the concepts of feasible control and feasible pair:

Definition 3.3 (Feasible Control, Feasible Pair). *An admissible control $\bar{\mathbf{u}}(\cdot) \in \mathcal{U}[t_0, t_{\mathrm{f}}]$ is said to be* feasible*, provided that (i) the response $\bar{\mathbf{x}}(\cdot; \mathbf{x}_0, \mathbf{u}(\cdot))$ is defined on the entire interval $t_0 \leq t \leq t_{\mathrm{f}}$, and (ii) $\bar{\mathbf{u}}(\cdot)$ and $\bar{\mathbf{x}}(\cdot; \mathbf{x}_0, \mathbf{u}(\cdot))$ satisfy* all *of the physical (point and path) constraints during this time interval; the pair $(\bar{\mathbf{u}}(\cdot), \bar{\mathbf{x}}(\cdot))$ is then called a* feasible pair. *The* set of feasible controls*, $\Omega[t_0, t_{\mathrm{f}}]$, is defined as*

$$\Omega[t_0, t_{\mathrm{f}}] \ := \ \{\mathbf{u}(\cdot) \in \mathcal{U}[t_0, t_{\mathrm{f}}] : \mathbf{u}(\cdot) \textit{ feasible}\}\,.$$

Example 3.4 (Simple Car Control Problem). Consider the control problem to drive a car, initially park at p_0, to a fixed pre-assigned destination p_f in a straight line (see Fig.3.2.). Here, t denotes time and $p(t)$ represents the position of the car at a given t.

To keep the problem as simple as possible, we approximate the car by a unit point mass that can be accelerated by using the throttle or decelerated by using the brake; the control $u(t)$ thus represents the force on the car due to either accelerating ($u(t) \geq 0$) or decelerating ($u(t) \leq 0$) the car. Here, the control region U is specified as

$$U := \{u \in \mathbb{R} : u^L \ \leq \ u(t) \ \leq \ u^U\},$$

with $u^L < 0 < u^U$, based on the acceleration and braking capabilities of the vehicle.

As the state, we choose the 2-vector $\mathbf{x}(t) := (p(t), \dot{p}(t))$; the physical reason for using a 2-vector is that we want to know (i) where we are, and (ii) how fast we are going. By neglecting friction, the dynamics of our system can be expressed based on Newton's second law of motion as $\ddot{p}(t) = u(t)$. Rewriting this equation in the vector form, we get

$$\dot{\mathbf{x}}(t) \ = \ \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t). \tag{3.5}$$

This is a mathematical model of the process in state form. Moreover, assuming that the car starts from rest, we have

$$\mathbf{x}(t_0) \ = \ \begin{bmatrix} p_0 \\ 0 \end{bmatrix}. \tag{3.6}$$

The control problem being to bring the car at p_f *at rest*, we impose terminal constraints as

$$\mathbf{x}(t_\mathrm{f}) - \begin{bmatrix} p_\mathrm{f} \\ 0 \end{bmatrix} \ = \ \mathbf{0}.$$

In addition, if the car starts with G litters of gas and there are no service stations on the way, another constraints is

$$\int_{t_0}^{t_\mathrm{f}} [k_1 u(t) + k_2 x_2(t)] \, \mathrm{d}t \ \leq \ G,$$

which assumes that the rate of gas consumption is proportional to both acceleration and speed with constants of proportionality k_1 and k_2.

Finally, we turn to the selection of a performance measure. Suppose that the objective is to make the car reach point p_f as quickly as possible; then, the performance measure $\mathcal{J}$ is given by

$$\mathcal{J} \ := \ t_\mathrm{f} - t_0 \ = \ \int_{t_0}^{t_\mathrm{f}} \mathrm{d}t.$$

An alternative criterion could be to minimize the amount of fuel expended.

3.2.5 Optimality Criteria

Having defined a performance criterion, the set of physical constraints to be satisfied, and the set of admissible controls, one can then state the optimal control problem as follows:

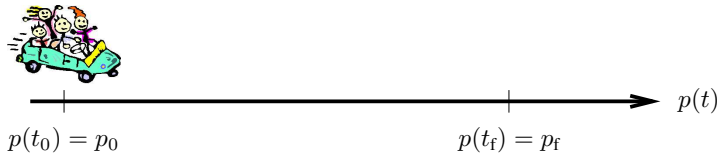


Figure 3.2. A simple control problem.

"find an admissible control $\mathbf{u}^\star \in \mathcal{U}[t_0, t_f]$ *which satisfies the physical constraints in such a manner that the cost functional* $\mathcal{J}(\mathbf{u}^\star)$ *has a minimum value."*

Similar to problems of the calculus of variations (see §2.3), we shall say that $\mathcal{J}$ assumes its minimum value at $\mathbf{u}^\star$ provided that

$$\mathcal{J}(\mathbf{u}^\star) \leq \mathcal{J}(\mathbf{u}), \quad \forall \mathbf{u} \in \Omega[t_0, t_f].$$

This assignment is *global* in nature and does not require consideration of a norm.

On the other hand, a description of the local minima of $\mathcal{J}$, namely

$$\exists \delta > 0 \text{ such that } \mathcal{J}(\mathbf{u}^\star) \leq \mathcal{J}(\mathbf{u}), \quad \forall \mathbf{u} \in \mathcal{B}_\delta(\mathbf{u}^\star) \cap \Omega[t_0, t_f],$$

requires that a norm (or, more generally, a distance) be specified. Having chosen the class of controls to be piecewise continuous functions, a possible choice is

$$\|\mathbf{u}\|_\infty := \sup_{t \in \bigcup_{k=0}^{N} (\theta_k, \theta_{k+1})} \|\mathbf{u}(t)\|,$$

with $t_0 = \theta_0 < \theta_1 < \cdots < \theta_N < \theta_{N+1} = t_f$ being a suitable partition for $\mathbf{u}$. This norm appears to be a natural choice, since $(\hat{\mathcal{C}}[t_0, t_f]^{n_u}, \|\cdot\|_\infty)$ is a Banach space. Under the additional assumption that the controls are continuously differentiable between two successive discontinuities $[\theta_k, \theta_{k+1}]$, $k = 0, \ldots, N$ (see Definition 3.1), another possible norm is

$$\|\mathbf{u}\|_{1,\infty} := \sup_{t \in \bigcup_{k=0}^{N} (\theta_k, \theta_{k+1})} \|\mathbf{u}(t)\| + \sup_{t \in \bigcup_{k=0}^{N} (\theta_k, \theta_{k+1})} \|\dot{\mathbf{u}}(t)\|.$$

3.2.6 Open-Loop vs. Closed-Loop Optimal Control

One of the ultimate goals of optimal control is *synthesize* an optimal control law, which can be used at *any* time t and for *any* (feasible) state value at t:

Definition 3.5 (Closed-Loop Optimal Control). *If a functional relation of the form*

$$\mathbf{u}^\star(t) = \boldsymbol{\omega}(t, \mathbf{x}(t)) \tag{3.7}$$

can be found for the optimal control at time t, then $\boldsymbol{\omega}$ *is called a* closed-loop optimal control *for the problem. (The terms* optimal feedback control *or* optimal control law *are also often used.)*

In general, the question of the very existence of a synthesizing control is rather complicated. Interestingly enough, this question has a positive answer for linear ODE systems under certain additional assumptions of an extremely general character (see §3.5). In this case, an optimal feedback can be found in the form of a linear time-varying control law,

$$\mathbf{u}^\star(t) \;=\; -\mathsf{K}(t)\,\mathbf{x}(t).$$

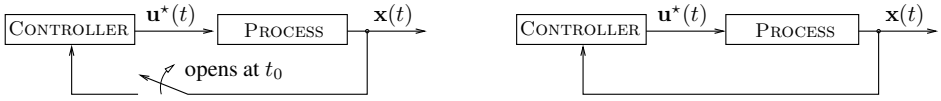
Obtaining a so-called open-loop optimal control law is much easier from a practical viewpoint:

Definition 3.6 (Open-Loop Optimal Control). *If the optimal control law is determined as a function of time for a specified initial state value, that is,*

$$\mathbf{u}^\star(t) \;=\; \boldsymbol{\omega}\left(t, \mathbf{x}(t_0)\right), \tag{3.8}$$

then the optimal control is said to be in open-loop *form.*

Therefore, an open-loop optimal control is optimal only for a *particular* initial state value, whereas, if an optimal control law is known, the optimal control history can be generated from *any* initial state. Conceptually, it is helpful to think off the difference between an optimal control law and an open-loop optimal control as shown in Fig. 3.3. below.



a. Open-loop optimal control b. Closed-loop optimal control

Figure 3.3. Open-loop *vs.* closed-loop optimal control.

Throughout this chapter, we shall mostly consider open-loop optimal control problems, also referred to as dynamic optimization problems in the literature. Note that open-loop optimal controls are rarely applied directly in practice due to the presence of *uncertainty* (such as model mismatch, process disturbance and variation in initial condition), which can make the system operate sub-optimally or, worse, lead to infeasible operation due to constraints violation. Yet, the knowledge of an open-loop optimal control law for a given process can provide valuable insight on how to improve system operation as well as some idea on how much can be gained upon optimization. Moreover, open-loop optimal controls are routinely used in a number of feedback control algorithms such as *model predictive control* (MPC) and *repeated optimization* [1, 42]. The knowledge of an open-loop optimal solution is also pivotal to many implicit optimization schemes including *NCO tracking* [51, 52].

3.3 EXISTENCE OF AN OPTIMAL CONTROL

Since optimal control problems encompass problems of the calculus of variations, difficulties are to be expected regarding the existence of a solution (see §2.4). Apart from the rather obvious case where no feasible control exists for the problem, the absence of an optimal control mostly lies in the fact that many feasible control sets of interest fail to be compact.

Observe first that when the ODEs have finite escape time for certain admissible controls, the cost functional can be unbounded. One particular sufficient condition for the solutions of ODEs to be extended indefinitely, as given by Theorem A.52 in Appendix A.5.1, is that the responses satisfy an a priori bound

$$\|\mathbf{x}(t; \mathbf{x}_0, \mathbf{u}(\cdot))\| \leq \alpha, \quad \forall t \geq t_0,$$

for every feasible control. In particular, the responses of a linear system $\dot{\mathbf{x}} = \mathsf{A}(t, \mathbf{u})\mathbf{x} + \mathbf{b}(t, \mathbf{u})$, from a fixed initial state, cannot have a finite escape time.

Observe next that when the time interval is unbounded, the corresponding set of feasible controls is itself unbounded, and hence not compact. Therefore, care should always be taken so that the operation be restricted to a compact time interval, e.g., $[t_0, T]$, where T is chosen so large that $\Omega[t_0, t_\mathrm{f}] \neq \emptyset$ for some $t_\mathrm{f} \in [t_0, T]$. These considerations are illustrated in an example below.

Example 3.7. Consider the car control problem described in Example 3.4, with the objective to find $(u(\cdot), t_\mathrm{f}) \in \hat{\mathcal{C}}[t_0, \infty) \times [t_0, \infty)$ such that p_f is reached within minimal amount of fuel expended:

$$\begin{aligned} \min_{u(\cdot), t_\mathrm{f}} \quad & \mathcal{J}(u, t_\mathrm{f}) := \int_{t_0}^{t_\mathrm{f}} [u(t)]^2 \, \mathrm{d}t \\ \text{s.t.} \quad & \text{Equations } (3.5, 3.6) \\ & x_1(t_\mathrm{f}) - p_\mathrm{f} = 0 \\ & 0 \leq u(t) \leq 1, \forall t. \end{aligned}$$

The state trajectories being continuous and $p_\mathrm{f} > p_0$, $\mathbf{u}(t) \equiv 0$ is infeasible and $\mathcal{J}(u) > 0$ for every feasible control.

Now, consider the sequence of (constant) admissible controls $u^k(t) = \frac{1}{k}$, $t \geq t_0$, $k \geq 1$. The response $\mathbf{x}^k(t)$, $t \geq t_0$, is easily calculated as

$$\begin{aligned} x_1(t) =& \frac{1}{2k}(t - t_0)^2 + p_0(t - t_0) \\ x_2(t) =& \frac{1}{k}(t - t_0) + p_0, \end{aligned}$$

and the target p_f is first reached at

$$t_\mathrm{f}^k = t_0 + 4k \left(\sqrt{p_0^2 + \frac{2}{k} p_\mathrm{f}} - p_0 \right).$$

Hence, $u^k \in \Omega[t_0, t_\mathrm{f}^k]$, and we have

$$\mathcal{J}(u^k) = \int_{t_0}^{t_\mathrm{f}^k} \frac{1}{k^2} \, \mathrm{d}t = \frac{4}{k} \left(\sqrt{p_0^2 + \frac{2}{k} p_\mathrm{f}} - p_0 \right) \longrightarrow 0,$$

as $k \to +\infty$. That is, $\inf \mathcal{J}(u) = 0$, i.e., the problem does not have a minimum.

Even when the time horizon is finite and the system response is bounded for every feasible control, an optimal control may not exist. This is because, similar to problems of the calculus of variations, the sets of interest are too big to be compact. This is illustrated subsequently in an example.

Example 3.8. Consider the problem to minimize the functional

$$\mathcal{J}(u) := \int_0^1 \sqrt{x(t)^2 + u(t)^2} \, \mathrm{d}t,$$

for $u \in \mathcal{C}[0,1]$, subject to the terminal constraint

$$x(1) = 1,$$

where the response $x(t)$ is given by the linear initial value problem,

$$\dot{x}(t) = u(t); \quad x(0) = 0.$$

Observe that the above optimal control problem is equivalent to that of minimizing the functional

$$\mathcal{J}(x) := \int_0^1 \sqrt{x(t)^2 + \dot{x}(t)^2}\, \mathrm{d}t,$$

on $\mathcal{D} := \{x \in \mathcal{C}^1[0,1] : x(0) = 0, x(1) = 1\}$. But since this variational problem does not have a solution (see Example 2.5, p. 68), it should be clear that the former optimal control problem does not have a solution either.

For a optimal control to exist, additional restrictions must be placed on the class of admissible controls. Two such possible classes of controls are:

(i) the class $\mathcal{U}_\lambda[t_0, t_\mathrm{f}] \subset \mathcal{U}[t_0, t_\mathrm{f}]$ of controls which satisfy a Lipschitz condition:

$$\|\mathbf{u}(t) - \mathbf{u}(s)\| \;\leq\; \lambda\|t - s\| \quad \forall t, s \in [t_0, t_\mathrm{f}];$$

(ii) the class $\mathcal{U}_r[t_0, t_\mathrm{f}] \subset \mathcal{U}[t_0, t_\mathrm{f}]$ of piecewise constant controls with at most r points of discontinuity.

Existence can also be guaranteed without restricting the controls, provided that certain convexity assumptions hold.

3.4 VARIATIONAL APPROACH

Sufficient condition for an optimal control problem to have a solution, such as those discussed in the previous section, while reassuring, are not at all useful in helping us *find* solutions. In this section (and the next section), we shall describe a set of conditions which any optimal control must necessarily satisfy. For many optimal control problems, such conditions allow to single out a small subset of controls, sometimes even a single control. There is thus reasonable chance of finding an optimal control, *if one exists*, among these candidates. However, it should be reemphasized that *necessary conditions may delineate a nonempty set of candidates, even though an optimal control does not exist for the problem*.

In this section, we shall consider optimal control problems having no restriction on the control variables (i.e., the control region U corresponds to $\mathbb{R}^{n_u}$) as well as on the state variables. More general optimal control problems with control and state path constraints, shall be considered later on in §3.5.

3.4.1 Euler-Lagrange Equations

We saw in §2.5.2 that the simplest problem of the calculus of variations has both its endpoints fixed. In optimal control, on the other hand, the simplest problem involves a free value of

the state variables at the right endpoint (terminal time):

$$\text{minimize:} \quad \int_{t_0}^{t_\mathrm{f}} \ell(t, \mathbf{x}(t), \mathbf{u}(t))\, \mathrm{d}t \tag{3.9}$$

$$\text{subject to:} \quad \dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)); \quad \mathbf{x}(t_0) = \mathbf{x}_0, \tag{3.10}$$

with fixed initial time t_0 and terminal time t_f. A control function $\mathbf{u}(t)$, $t_0 \leq t \leq t_\mathrm{f}$, together with the initial value problem (3.10), determines the response $\mathbf{x}(t)$, $t_0 \leq t \leq t_\mathrm{f}$ (provided it exists). Thus, we may speak of finding a control, since the corresponding response is implied.

To develop first-order necessary conditions in problems of the calculus of variations (Euler's equation), we constructed a one-parameter family of comparison trajectories $\mathbf{x}(t) + \eta\boldsymbol{\xi}(t)$, with $\boldsymbol{\xi} \in \mathcal{C}^1[t_0, t_\mathrm{f}]^{n_x}$ such that $\|\boldsymbol{\xi}\|_{1,\infty} \leq \delta$ (see Theorem 2.14 and proof). However, when it comes to optimal control problems such as (3.9,3.10), a variation of the state trajectory $\mathbf{x}$ cannot be explicitely related to a variation of the control $\mathbf{u}$, in general, because the state and control variables are implicitly related by the (nonlinear) differential equation (3.10). Instead, we shall consider a one-parameter family of comparison trajectories $\mathbf{u}(t) + \eta\boldsymbol{\omega}(t)$, with $\boldsymbol{\omega} \in \mathcal{C}[t_0, t_\mathrm{f}]^{n_u}$ such that $\|\boldsymbol{\omega}\|_\infty \leq \delta$. Then, analogous to the proof of Theorem 2.14, we shall use the geometric characterization for a local minimizer of a functional on a subset of a normed linear space as given by Theorem 2.13. These considerations lead to the following:

Theorem 3.9 (First-Order Necessary Conditions). *Consider the problem to minimize the functional*

$$\mathcal{J}(\mathbf{u}) := \int_{t_0}^{t_\mathrm{f}} \ell(t, \mathbf{x}(t), \mathbf{u}(t))\, \mathrm{d}t, \tag{3.11}$$

subject to

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)); \quad \mathbf{x}(t_0) = \mathbf{x}_0, \tag{3.12}$$

for $\mathbf{u} \in \mathcal{C}[t_0, t_\mathrm{f}]^{n_u}$, with fixed endpoints $t_0 < t_\mathrm{f}$, where ℓ and $\mathbf{f}$ are continuous in $(t, \mathbf{x}, \mathbf{u})$ and have continuous first partial derivatives with respect to $\mathbf{x}$ and $\mathbf{u}$ for all $(t, \mathbf{x}, \mathbf{u}) \in [t_0, t_\mathrm{f}] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u}$. Suppose that $\mathbf{u}^\star \in \mathcal{C}[t_0, t_\mathrm{f}]^{n_u}$ is a (local) minimizer for the problem, and let $\mathbf{x}^\star \in \mathcal{C}^1[t_0, t_\mathrm{f}]^{n_x}$ denote the corresponding response. Then, there is a vector function $\boldsymbol{\lambda}^\star \in \mathcal{C}^1[t_0, t_\mathrm{f}]^{n_x}$ such that the triple $(\mathbf{u}^\star, \mathbf{x}^\star, \boldsymbol{\lambda}^\star)$ satisfies the system

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)); \qquad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{3.13}$$

$$\dot{\boldsymbol{\lambda}}(t) = -\ell_{\mathbf{x}}(t, \mathbf{x}(t), \mathbf{u}(t)) - \mathbf{f}_{\mathbf{x}}(t, \mathbf{x}(t), \mathbf{u}(t))^{\mathsf{T}} \boldsymbol{\lambda}(t); \qquad \boldsymbol{\lambda}(t_\mathrm{f}) = \mathbf{0} \tag{3.14}$$

$$\mathbf{0} = \ell_{\mathbf{u}}(t, \mathbf{x}(t), \mathbf{u}(t)) + \mathbf{f}_{\mathbf{u}}(t, \mathbf{x}(t), \mathbf{u}(t))^{\mathsf{T}} \boldsymbol{\lambda}(t). \tag{3.15}$$

for $t_0 \leq t \leq t_\mathrm{f}$. These equations are known collectively as the Euler-Lagrange equations*, and* (3.14) *is often referred to as the* adjoint equation *(or the* costate equation*).*

Proof. Consider a one-parameter family of comparison controls $\mathbf{v}(t; \eta) := \mathbf{u}^\star(t) + \eta\boldsymbol{\omega}(t)$, where $\boldsymbol{\omega}(t) \in \mathcal{C}[t_0, t_\mathrm{f}]^{n_u}$ is some fixed function, and η is a (scalar) parameter. Based on the continuity and differentiability properties of $\mathbf{f}$, we know that there exists $\bar{\eta} > 0$ such that the response $\mathbf{y}(t; \eta) \in \mathcal{C}^1[t_0, t_\mathrm{f}]^{n_x}$ associated to $\mathbf{v}(t; \eta)$ through (3.12) exists, is unique, and is differentiable with respect to η, for all $\eta \in \mathcal{B}_{\bar{\eta}}(0)$ and for all $t \in [t_0, t_\mathrm{f}]$ (see Appendix A.5). Clearly, $\eta = 0$ provides the optimal response $\mathbf{y}(t; 0) \equiv \mathbf{x}^\star(t)$, $t_0 \leq t \leq t_\mathrm{f}$.

Since the control $\mathbf{v}(t;\eta)$ is admissible and its associated response is $\mathbf{y}(t,\eta)$, we have

$$\begin{aligned}\mathcal{J}(\mathbf{v}(\cdot;\eta)) &= \int_{t_0}^{t_\mathrm{f}} \Big[\ell(t,\mathbf{y}(t;\eta),\mathbf{v}(t;\eta)) + \boldsymbol{\lambda}(t)^\mathsf{T}\left[\mathbf{f}(t,\mathbf{y}(t;\eta),\mathbf{v}(t;\eta)) - \dot{\mathbf{y}}(t;\eta)\right]\Big]\,\mathrm{d}t \\ &= \int_{t_0}^{t_\mathrm{f}} \Big[\ell(t,\mathbf{y}(t;\eta),\mathbf{v}(t;\eta)) + \boldsymbol{\lambda}(t)^\mathsf{T}\mathbf{f}(t,\mathbf{y}(t;\eta),\mathbf{v}(t;\eta)) + \dot{\boldsymbol{\lambda}}(t)^\mathsf{T}\mathbf{y}(t;\eta)\Big]\,\mathrm{d}t \\ &\quad - \boldsymbol{\lambda}(t_\mathrm{f})^\mathsf{T}\mathbf{y}(t_\mathrm{f};\eta) + \boldsymbol{\lambda}(t_0)^\mathsf{T}\mathbf{y}(t_0;\eta),\end{aligned}$$

for any $\boldsymbol{\lambda} \in \mathcal{C}^1[t_0,t_\mathrm{f}]^{n_x}$ and for each $\eta \in \mathcal{B}_{\bar{\eta}}(0)$. Based on the differentiability properties of ℓ and $\mathbf{y}$, and by Theorem 2.A.59, we have

$$\begin{aligned}\frac{\partial}{\partial\eta}\mathcal{J}(\mathbf{v}(\cdot;\eta)) &= \int_{t_0}^{t_\mathrm{f}} \Big[\ell_\mathbf{u}(t,\mathbf{y}(t;\eta),\mathbf{v}(t;\eta)) + \mathbf{f}_\mathbf{u}(t,\mathbf{y}(t;\eta),\mathbf{v}(t;\eta))^\mathsf{T}\boldsymbol{\lambda}(t)\Big]^\mathsf{T}\boldsymbol{\omega}(t)\,\mathrm{d}t \\ &\quad + \int_{t_0}^{t_\mathrm{f}} \Big[\ell_\mathbf{x}(t,\mathbf{y}(t;\eta),\mathbf{v}(t;\eta)) + \mathbf{f}_\mathbf{x}(t,\mathbf{y}(t;\eta),\mathbf{v}(t;\eta))^\mathsf{T}\boldsymbol{\lambda}(t) + \dot{\boldsymbol{\lambda}}(t)\Big]^\mathsf{T}\mathbf{y}_\eta(t;\eta)\,\mathrm{d}t \\ &\quad - \boldsymbol{\lambda}(t_\mathrm{f})^\mathsf{T}\mathbf{y}_\eta(t_\mathrm{f};\eta) + \boldsymbol{\lambda}(t_0)^\mathsf{T}\mathbf{y}_\eta(t_0;\eta),\end{aligned}$$

for any $\boldsymbol{\omega} \in \mathcal{C}[t_0,t_\mathrm{f}]^{n_u}$ and any $\boldsymbol{\lambda} \in \mathcal{C}^1[t_0,t_\mathrm{f}]^{n_x}$. Taking the limit as $\eta \to 0$, and since $\mathbf{y}_\eta(t_0;\eta) = 0$, we get

$$\begin{aligned}\delta\mathcal{J}(\mathbf{u}^\star;\boldsymbol{\omega}) &= \int_{t_0}^{t_\mathrm{f}} \Big[\ell_\mathbf{u}(t,\mathbf{x}^\star(t),\mathbf{u}^\star(t)) + \mathbf{f}_\mathbf{u}(t,\mathbf{x}^\star(t),\mathbf{u}^\star(t))^\mathsf{T}\boldsymbol{\lambda}(t)\Big]^\mathsf{T}\boldsymbol{\omega}(t)\,\mathrm{d}t \\ &\quad + \int_{t_0}^{t_\mathrm{f}} \Big[\ell_\mathbf{x}(t,\mathbf{x}^\star(t),\mathbf{u}^\star(t)) + \mathbf{f}_\mathbf{x}(t,\mathbf{x}^\star(t),\mathbf{u}^\star(t))^\mathsf{T}\boldsymbol{\lambda}(t) + \dot{\boldsymbol{\lambda}}(t)\Big]^\mathsf{T}\mathbf{y}_\eta(t;0)\,\mathrm{d}t \\ &\quad - \boldsymbol{\lambda}(t_\mathrm{f})^\mathsf{T}\mathbf{y}_\eta(t_\mathrm{f};0),\end{aligned}$$

which is finite for each $\boldsymbol{\omega} \in \mathcal{C}[t_0,t_\mathrm{f}]^{n_u}$ and each $\boldsymbol{\lambda} \in \mathcal{C}^1[t_0,t_\mathrm{f}]^{n_x}$, since the integrand is continuous on $[t_0,t_\mathrm{f}]$. That is, $\delta\mathcal{J}(\mathbf{u}^\star;\boldsymbol{\omega})$ exists for each $\boldsymbol{\omega} \in \mathcal{C}[t_0,t_\mathrm{f}]^{n_u}$ and each $\boldsymbol{\lambda} \in \mathcal{C}^1[t_0,t_\mathrm{f}]^{n_x}$.

Now, $\mathbf{u}^\star$ being a local minimizer, by Theorem 2.13,

$$0 = \int_{t_0}^{t_\mathrm{f}} \Big[\ell_\mathbf{x}^\star + \mathbf{f}_\mathbf{x}^{\star\mathsf{T}}\boldsymbol{\lambda}(t) + \dot{\boldsymbol{\lambda}}(t)\Big]^\mathsf{T}\mathbf{y}_\eta(t;0) + \Big[\ell_\mathbf{u}^\star + \mathbf{f}_\mathbf{u}^{\star\mathsf{T}}\boldsymbol{\lambda}(t)\Big]^\mathsf{T}\boldsymbol{\omega}(t)\,\mathrm{d}t - \boldsymbol{\lambda}(t_\mathrm{f})^\mathsf{T}\mathbf{y}_\eta(t_\mathrm{f};0),$$

for each $\boldsymbol{\omega} \in \mathcal{C}[t_0,t_\mathrm{f}]^{n_u}$ and each $\boldsymbol{\lambda} \in \mathcal{C}^1[t_0,t_\mathrm{f}]^{n_x}$, where the compressed notations $\ell_z^\star := \ell_z(t,\mathbf{x}^\star(t),\mathbf{u}^\star(t))$ and $\mathbf{f}_z^\star := \mathbf{f}_z(t,\mathbf{x}^\star(t),\mathbf{u}^\star(t))$ are used.

Because the effect of a variation of the control on the course of the response is hard to determine (i.e., $\mathbf{y}_\eta(t;0)$), we choose $\boldsymbol{\lambda}^\star(t)$, $t_0 \le t \le t_\mathrm{f}$, so as to obey the differential equation

$$\dot{\boldsymbol{\lambda}}(t) = -\,\mathbf{f}_\mathbf{x}^{\star\mathsf{T}}\boldsymbol{\lambda}(t) - \ell_\mathbf{x}^\star, \tag{3.16}$$

with the terminal condition $\boldsymbol{\lambda}(t_\mathrm{f}) = \mathbf{0}$. Note that (3.16) being a linear system of ODEs, and from the regularity assumptions on ℓ and $\mathbf{f}$, the solution $\boldsymbol{\lambda}^\star$ exists and is unique over $[t_0,t_\mathrm{f}]$ (see Theorem A.50, p. xiv), i.e., $\boldsymbol{\lambda} \in \mathcal{C}^1[t_0,t_\mathrm{f}]^{n_x}$. That is, the condition

$$0 = \int_{t_0}^{t_\mathrm{f}} \Big[\ell_\mathbf{u}^\star + \mathbf{f}_\mathbf{u}^{\star\mathsf{T}}\boldsymbol{\lambda}^\star(t)\Big]^\mathsf{T}\boldsymbol{\omega}(t)\,\mathrm{d}t,$$

must hold for any $\boldsymbol{\omega} \in \mathcal{C}[t_0, t_\mathrm{f}]^{n_u}$. In particular, for $\boldsymbol{\omega}(t)$ such that $\omega_i(t) := \ell^\star_{u_i} + \mathbf{f}^{\star\,\mathsf{T}}_{u_i} \boldsymbol{\lambda}^\star(t)$ and $\omega_j(t) = 0$ for $j \neq i$, we get

$$0 = \int_{t_0}^{t_\mathrm{f}} \left[\ell^\star_{u_i} + \mathbf{f}^{\star\,\mathsf{T}}_{u_i} \boldsymbol{\lambda}^\star(t) \right]^2 \mathrm{d}t,$$

for each $i = 1, \ldots, n_x$, which in turn implies the necessary condition that

$$0 = \ell_{u_i}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t)) + \mathbf{f}_{u_i}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t))^\mathsf{T} \boldsymbol{\lambda}^\star(t), \quad i = 1, \ldots, n_x,$$

for each $t \in [t_0, t_\mathrm{f}]$. □

Some comments are in order before we look at an example.

- The optimality conditions consist of n_u algebraic equations (3.15), together with $2 \times n_x$ ODEs (3.13,3.14) and their respective boundary conditions. Hence, the Euler-Lagrange equations provide a complete set of necessary conditions. However, the boundary conditions for (3.13) and (3.14) are split, i.e., some are given at $t = t_0$ and others at $t = t_\mathrm{f}$. Such problems are known as *two-point boundary value problems* (TPBVPs) and are notably more difficult to solve than IVPs.

- In the special case where $\mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) := \mathbf{u}(t)$, with $n_u = n_x$, (3.15) gives

  $$\boldsymbol{\lambda}^\star(t) = -\ell_\mathbf{u}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t)).$$

  Then, from (3.14), we obtain the Euler equation

  $$\frac{\mathrm{d}}{\mathrm{d}t} \ell_\mathbf{u}(t, \mathbf{x}^\star(t), \dot{\mathbf{x}}^\star(t)) = \ell_\mathbf{x}(t, \mathbf{x}^\star(t), \dot{\mathbf{x}}^\star(t)),$$

  together with the natural boundary condition

  $$[\ell_\mathbf{u}(t, \mathbf{x}^\star(t), \dot{\mathbf{x}}^\star(t))]_{t=t_\mathrm{f}} = \mathbf{0}.$$

  Hence, the Euler-Lagrange equations encompass the necessary conditions of optimality derived previously for problems of the calculus of variations (see §2.5).

- It is convenient to introduce the *Hamiltonian function* $\mathcal{H} : \mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R}^{n_x} \to \mathbb{R}$ associated with the optimal control problem (3.9,3.10), by adjoining the right-hand side of the differential equations to the cost integrand as

  $$\mathcal{H}(t, \mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}) = \ell(t, \mathbf{x}, \mathbf{u}) + \boldsymbol{\lambda}^\mathsf{T} \mathbf{f}(t, \mathbf{x}, \mathbf{u}). \tag{3.17}$$

  Thus, Euler-Lagrange equations (3.13–3.15) can be rewritten as

  $$\dot{\mathbf{x}}(t) = \mathcal{H}_{\boldsymbol{\lambda}}; \qquad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{3.13'}$$

  $$\dot{\boldsymbol{\lambda}}(t) = -\mathcal{H}_\mathbf{x}; \qquad \boldsymbol{\lambda}(t_\mathrm{f}) = \mathbf{0} \tag{3.14'}$$

  $$\mathbf{0} = \mathcal{H}_\mathbf{u}, \tag{3.15'}$$

  for $t_0 \leq t \leq t_\mathrm{f}$. Note that a necessary condition for the triple $(\mathbf{u}^\star, \mathbf{x}^\star, \boldsymbol{\lambda}^\star)$ to give a local minimum of $\mathcal{J}$ is that $\mathbf{u}^\star(t)$ be a stationary point of the Hamiltonian function with $\mathbf{x}^\star(t)$ and $\boldsymbol{\lambda}^\star(t)$, at each $t \in [t_0, t_\mathrm{f}]$. In some cases, one can express $\mathbf{u}(t)$ as a

function of $\mathbf{x}(t)$ and $\boldsymbol{\lambda}(t)$ from (3.15'), and then substitute into (3.13',3.14') to get a TPBVP in the variables $\mathbf{x}$ and $\boldsymbol{\lambda}$ only (see Example 3.10 below).

- The variation of the Hamiltonian function along an optimal trajectory is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H} \;= \mathcal{H}_t + \mathcal{H}_{\mathbf{x}}^{\mathsf{T}}\dot{\mathbf{x}} + \mathcal{H}_{\mathbf{u}}^{\mathsf{T}}\dot{\mathbf{u}} + \mathbf{f}^{\mathsf{T}}\dot{\boldsymbol{\lambda}} \;=\; \mathcal{H}_t + \mathcal{H}_{\mathbf{u}}^{\mathsf{T}}\dot{\mathbf{u}} + \mathbf{f}^{\mathsf{T}}\left[\mathcal{H}_{\mathbf{x}} + \dot{\boldsymbol{\lambda}}\right] \;=\; \mathcal{H}_t.$$

  Hence, if neither ℓ nor $\mathbf{f}$ depend explicitly on t, we get $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H} \equiv 0$, hence $\mathcal{H}$ is constant along an optimal trajectory; in other words, $\mathcal{H}$ yields a *first integral* to the TPBVP (3.13'–3.15').

- The Euler-Lagrange equations (3.13'–3.15') are necessary conditions both for a minimization *and* for a maximization problem. Yet, in a minimization problem, $\mathbf{u}^\star(t)$ must minimize $\mathcal{H}(t, \mathbf{x}^\star(t), \cdot, \boldsymbol{\lambda}^\star(t))$, i.e., the condition $\mathcal{H}_{\mathbf{uu}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t), \boldsymbol{\lambda}^\star(t)) \geq 0$ is also necessary. On the other hand, the additional condition $\mathcal{H}_{\mathbf{uu}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t), \boldsymbol{\lambda}^\star(t)) \leq 0$ is necessary in a maximization problem. These latter conditions have not yet been established and shall be discussed later on.

Example 3.10. Consider the optimal control problem

$$\text{minimize:}\;\; \mathcal{J}(u) \;:=\; \int_0^1 \left[\tfrac{1}{2}u(t)^2 - x(t)\right]\,\mathrm{d}t \tag{3.18}$$

$$\text{subject to:}\;\; \dot{x}(t) \;=\; 2\left[1 - u(t)\right];\quad x(0) \;=\; 1. \tag{3.19}$$

To find candidate optimal controls for the problem (3.18,3.19), we start by forming the Hamiltonian function

$$\mathcal{H}(x, u, \lambda) \;=\; \tfrac{1}{2}u^2 - x + 2\lambda(1 - u).$$

Candidate solutions $(u^\star, x^\star, \lambda^\star)$ are those satisfying the Euler-Lagrange equations, i.e.,

$$\begin{aligned} \dot{x}^\star(t) \;&= \mathcal{H}_\lambda \;=\; 2\left[1 - u^\star(t)\right]; && x^\star(0) \;= 1 \\ \dot{\lambda}^\star(t) \;&= -\,\mathcal{H}_x \;=\; 1; && \lambda^\star(1) \;= 0 \\ 0 \;&= \mathcal{H}_u \;=\; u^\star(t) - 2\lambda^\star(t). \end{aligned}$$

The adjoint equation trivially yields

$$\lambda^\star(t) \;=\; t - 1,$$

and from the optimality condition, we get

$$u^\star(t) \;=\; 2(t - 1).$$

(Note that $u^\star$ is indeed a candidate *minimum* solution for the problem since $\mathcal{H}_{uu} = 1 > 0$ for each $0 \leq t \leq 1$.) Finally, substituting the optimal control candidate back into (3.19) yields

$$\dot{x}^\star(t) \;=\; 6 - 4t;\quad x^\star(0) \;=\; 1.$$

Integrating the latter equation, and drawing the results together, we obtain

$$u^\star(t) \;= 2(t - 1) \tag{3.20}$$

$$x^\star(t) \;=\; -\,2t^2 + 6t + 1 \tag{3.21}$$

$$\lambda^\star(t) \;= t - 1. \tag{3.22}$$

It is also readily verified that $\mathcal{H}$ is constant along the optimal trajectory,

$$\mathcal{H}(t, x^\star(t), u^\star(t), \lambda^\star(t)) \;=\; -5.$$

Finally, we illustrate the optimality of the control $u^\star$ by considering the modified controls $v(t;\eta) := u^\star(t)+\eta\omega(t)$, and their associated responses $y(t;\eta)$. The perturbed cost function reads:

$$\begin{aligned} \mathcal{J}(v(t;\eta)) \;&:=\; \int_0^1 \left[\frac{1}{2}[u^\star(t)+\eta\omega(t)]^2 - y(t;\eta)\right] \mathrm{d}t \\ \text{s.t.}\;\; \dot{y}(t;\eta) \;&=\; 2\,[1-u^\star(t)-\eta\omega(t)]\,; \quad x(0) \;=\; 1. \end{aligned}$$

The cost function $\mathcal{J}(v(t;\eta))$ is represented in Fig. 3.4. for different perturbations $\omega(t) = t^k$, $k = 0, \ldots, 4$. Note that the minimum of $\mathcal{J}(v(t;\eta))$ is always attained at $\eta = 0$.

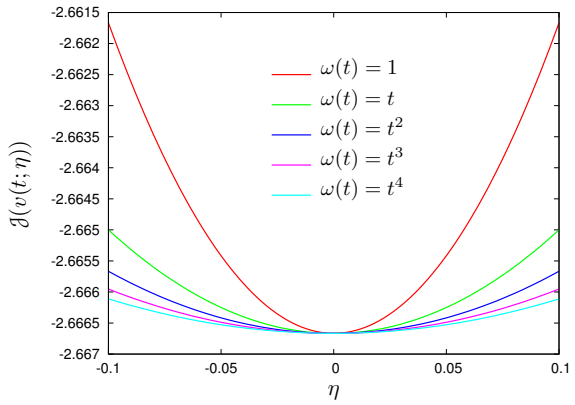


Figure 3.4. Function $\mathcal{J}(v(t;\eta))$ for various perturbations $\omega(t) = t^k$, $k = 0, \ldots, 4$, in Example 3.10.

3.4.2 Mangasarian Sufficient Conditions

In essence, searching for a control $\mathbf{u}^\star$ that minimizes the performance measure $\mathcal{J}$ means that $\mathcal{J}(\mathbf{u}^\star) \leq \mathcal{J}(\mathbf{u})$, for all admissible $\mathbf{u}$. That is, what we want to determine is the *global minimum* value of $\mathcal{J}$, not merely *local minima*. (Remind that there may actually be several global optimal controls, i.e., distinct controls achieving the global minimum of $\mathcal{J}$.) Conditions under which the necessary conditions (3.13'–3.15') are also sufficient for optimality, i.e., provide a global optimal control are given by the following:

Theorem 3.11 (Mangasarian Sufficient Condition). *Consider the problem to minimize the functional*

$$\mathcal{J}(\mathbf{u}) := \int_{t_0}^{t_\mathrm{f}} \ell(t, \mathbf{x}(t), \mathbf{u}(t))\, \mathrm{d}t, \tag{3.23}$$

subject to

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)); \quad \mathbf{x}(t_0) = \mathbf{x}_0, \tag{3.24}$$

for $\mathbf{u} \in \mathcal{C}[t_0, t_\mathrm{f}]^{n_u}$, with fixed endpoints $t_0 < t_\mathrm{f}$, where ℓ and $\mathbf{f}$ are continuous in $(t, \mathbf{x}, \mathbf{u})$, have continuous first partial derivatives with respect to $\mathbf{x}$ and $\mathbf{u}$, and are [strictly] jointly convex in $\mathbf{x}$ and $\mathbf{u}$, for all $(t, \mathbf{x}, \mathbf{u}) \in [t_0, t_\mathrm{f}] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u}$. Suppose that the triple $\mathbf{u}^\star \in \mathcal{C}[t_0, t_\mathrm{f}]^{n_u}$, $\mathbf{x}^\star \in \mathcal{C}^1[t_0, t_\mathrm{f}]^{n_x}$ and $\boldsymbol{\lambda}^\star \in \mathcal{C}^1[t_0, t_\mathrm{f}]^{n_x}$ satisfies the Euler-Lagrange equations (3.13–3.15). Suppose also that

$$\boldsymbol{\lambda}^\star(t) \ \geq \mathbf{0}, \tag{3.25}$$

for all $t \in [t_0, t_\mathrm{f}]$. Then, $\mathbf{u}^\star$ is a [strict] global minimizer for the problem (3.23,3.24).

Proof. ℓ being jointly convex in $(\mathbf{x}, \mathbf{u})$, for any feasible control $\mathbf{u}$ and its associated response $\mathbf{x}$, we have

$$\begin{aligned}
\mathcal{J}(\mathbf{u}) - \mathcal{J}(\mathbf{u}^\star) \ &= \ \int_{t_0}^{t_\mathrm{f}} \left[\ell(t, \mathbf{x}(t), \mathbf{u}(t)) - \ell(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t))\right] \, \mathrm{d}t \\
&\geq \ \int_{t_0}^{t_\mathrm{f}} \left({\ell_{\mathbf{x}}^\star}^\mathsf{T} \left[\mathbf{x}(t) - \mathbf{x}^\star(t)\right] + {\ell_{\mathbf{u}}^\star}^\mathsf{T} \left[\mathbf{u}(t) - \mathbf{u}^\star(t)\right] \right) \, \mathrm{d}t,
\end{aligned}$$

with the usual compressed notation. Since the triple $(\mathbf{u}^\star, \mathbf{x}^\star, \boldsymbol{\lambda}^\star)$ satisfies the Euler-Lagrange equations (3.13–3.15), we obtain

$$\begin{aligned}
\mathcal{J}(\mathbf{u}) - \mathcal{J}(\mathbf{u}^\star) \ \geq \ \int_{t_0}^{t_\mathrm{f}} \Bigg(&- \left[{\mathbf{f}_{\mathbf{x}}^\star}^\mathsf{T} \boldsymbol{\lambda}^\star(t) + \dot{\boldsymbol{\lambda}}^\star(t) \right]^\mathsf{T} \left[\mathbf{x}(t) - \mathbf{x}^\star(t)\right] \\
&- \left[{\mathbf{f}_{\mathbf{u}}^\star}^\mathsf{T} \boldsymbol{\lambda}^\star(t) \right]^\mathsf{T} \left[\mathbf{u}(t) - \mathbf{u}^\star(t)\right] \Bigg) \, \mathrm{d}t,
\end{aligned}$$

Integrating by part the term in $\dot{\boldsymbol{\lambda}}^\star(t)$, and rearranging the terms, we get

$$\begin{aligned}
\mathcal{J}(\mathbf{u}) - \mathcal{J}(\mathbf{u}^\star) \ \geq \ \int_{t_0}^{t_\mathrm{f}} &\boldsymbol{\lambda}^\star(t)^\mathsf{T} \left(\mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) - \mathbf{f}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t)) - \mathbf{f}_{\mathbf{x}}^\star \left[\mathbf{x}(t) - \mathbf{x}^\star(t)\right] \right. \\
&\left. - \mathbf{f}_{\mathbf{u}}^\star \left[\mathbf{u}(t) - \mathbf{u}^\star(t)\right]\right) \, \mathrm{d}t \\
&- \boldsymbol{\lambda}^\star(t_\mathrm{f})^\mathsf{T} \left[\mathbf{x}(t_\mathrm{f}) - \mathbf{x}^\star(t_\mathrm{f})\right] + \boldsymbol{\lambda}^\star(t_0)^\mathsf{T} \left[\mathbf{x}(t_0) - \mathbf{x}^\star(t_0)\right].
\end{aligned}$$

Note that the integrand is positive due to (3.25) and the joint convexity of $\mathbf{f}$ in $(\mathbf{x}, \mathbf{u})$; the remaining two terms are equal to zero due to the optimal adjoint boundary conditions and the prescribed state initial conditions, respectively. That is,

$$\mathcal{J}(\mathbf{u}) \ \geq \ \mathcal{J}(\mathbf{u}^\star),$$

for each feasible control. □

Remark 3.12. In the special case where $\mathbf{f}$ is linear in $(\mathbf{x}, \mathbf{u})$, the result holds *without* any sign restriction for the costates $\boldsymbol{\lambda}^\star(t)$. Further, if ℓ is jointly convex in $(\mathbf{x}, \mathbf{u})$ and φ is convex in $\mathbf{x}$, while $\mathbf{f}$ is jointly concave in $(\mathbf{x}, \mathbf{u})$ and $\boldsymbol{\lambda}^\star(t) \leq \mathbf{0}$, then the necessary conditions are also sufficient for optimality.

Remark 3.13. The Mangasarian sufficient conditions have limited applicability for, in most practical problems, either the terminal cost, the integral cost, or the differential equations fail to be convex or concave. [2]

[2] A less restrictive sufficient condition, known as the Arrow sufficient condition, 'only' requires that the function $\mathcal{M}(t, \mathbf{x}, \boldsymbol{\lambda}^\star) := \min_{\mathbf{u} \in U} \mathcal{H}(t, \mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}^\star)$, i.e., the minimized Hamiltonian with respect to $\mathbf{u} \in U$, be a convex function in $\mathbf{x}$. See, e.g., [49] for a survey of sufficient conditions in optimal control theory.

Example 3.14. Consider the optimal control problem (3.18,3.19) in Example 3.10. The integrand is jointly convex in (u, x) on $\mathbb{R}^2$, and the right-hand side of the differential equation is linear in u (and independent of x). Moreover, the candidate solution $(u^\star(t), x^\star(t), \lambda^\star)$ given by (3.20–3.22) satisfies the Euler-Lagrange equations (3.13–3.15), for each $t \in [0, 1]$. Therefore, $u^\star(t)$ is a global minimizer for the problem (irrespective of the sign of the adjoint variable due to the linearity of (3.19), see Remark 3.12).

3.4.3 Piecewise Continuous Extremals

It may sometimes occur that a continuous control $\mathbf{u} \in \mathcal{C}[t_0, t_\mathrm{f}]^{n_u}$ satisfying the Euler-Lagrange equations cannot be found for a particular optimal control problem. It is then natural to wonder whether such problems have extremals in the larger class of piecewise continuous controls $\hat{\mathcal{C}}[t_0, t_\mathrm{f}]^{n_u}$ (see Definition 3.1). It is also natural to seek improved results in the class of piecewise continuous controls, even though a continuous control satisfying the Euler-Lagrange equations could be found. Discontinuous controls give rise to discontinuities in the slope of the response (i.e., $\mathbf{x} \in \hat{\mathcal{C}}^1[t_0, t_\mathrm{f}]$), and are referred to as *corner points* (or simply *corners*) by analogy to classical problems of the calculus of variations (see §2.6). The purpose of this subsection is to summarize the conditions that must hold at the corner points of an optimal solution.

Consider an optimal control problem of the form (3.9,3.10), and suppose that $\hat{\mathbf{u}}^\star \in \hat{\mathcal{C}}[t_0, t_\mathrm{f}]$ is an optimal control for that problem, with associated response $\hat{\mathbf{x}}^\star$ and adjoint $\hat{\boldsymbol{\lambda}}^\star$. Then, at every possible corner point $\theta \in (t_0, t_\mathrm{f})$ of $\hat{\mathbf{u}}^\star$, we have

$$\hat{\mathbf{x}}^\star(\theta^-) = \hat{\mathbf{x}}^\star(\theta^+) \tag{3.26}$$

$$\hat{\boldsymbol{\lambda}}^\star(\theta^-) = \hat{\boldsymbol{\lambda}}^\star(\theta^+) \tag{3.27}$$

$$\mathcal{H}(\theta^-, \hat{\mathbf{x}}^\star(\theta), \hat{\mathbf{u}}^\star(\theta^-), \hat{\boldsymbol{\lambda}}^\star(\theta)) = \mathcal{H}(\theta^+, \hat{\mathbf{x}}^\star(\theta), \hat{\mathbf{u}}^\star(\theta^+), \hat{\boldsymbol{\lambda}}^\star(\theta)), \tag{3.28}$$

where θ^- and θ^+ denote the time just before and just after the corner, respectively; $z(\theta^-)$ and $z(\theta^+)$ denote the left and right limit values of a quantity z at θ, respectively.

Remark 3.15 (Link to the Weierstrass-Erdmann Corner Conditions). It is readily shown that the corner conditions (3.27) and (3.28) are equivalent to the Weierstrass-Erdmann conditions (2.23) and (2.24) in classical problems of the calculus of variations.

Although corners in optimal control trajectories are more common in problems having inequality path constraints (either input or state constraints), the following example illustrates that problems without path inequality constraint may also exhibit corners.

Example 3.16. Consider the optimal control problem

$$\text{minimize:} \quad \mathcal{J}(u) := \int_0^1 \left[u(t)^2 - u(t)^4 - x(t)\right] \mathrm{d}t \tag{3.29}$$

$$\text{subject to:} \quad \dot{x}(t) = -u(t); \quad x(0) = 1 \tag{3.30}$$

The Hamiltonian function for this problem reads

$$\mathcal{H}(x, u, \lambda) = u^2 - u^4 - x - \lambda u.$$

Candidate optimal solutions $(u^\star, x^\star, \lambda^\star)$ are those satisfying the Euler-Lagrange equations

$$\begin{aligned} \dot{x}(t) &= \mathcal{H}_\lambda = u(t); & x(0) &= 1 \\ \dot{\lambda}(t) &= -\mathcal{H}_x = 1; & \lambda^\star(1) &= 0 \qquad (3.31) \\ 0 &= \mathcal{H}_u = 2u(t) - 4u(t)^3 - \lambda(t). & & \qquad (3.32) \end{aligned}$$

The adjoint equation (3.31) has solution $\lambda^\star(t) = t - 1$, $0 \le t \le 1$, which upon substitution into (3.32) yields

$$2u^\star(t) - 4u^\star(t)^3 = t - 1.$$

Values of the control variable $u(t)$, $0 \le t \le 1$, satisfying the former condition are shown in Fig. 3.16 below. Note that for there is a unique solution $u_1(t)$ to the Euler-Lagrange equations for $0 \le t \lessapprox 0.455$, then 3 possible solutions $u_1(t)$, $u_2(t)$, and $u_3(t)$ exist for $0.455 \lessapprox t \le 1$. That is, candidate optimal controls start with the value $u^\star(t) = u_1(t)$ until $t \approx 0.455$. Then, the optimal control may be discontinuous and switch between $u_1(t)$, $u_2(t)$, and $u_3(t)$. Note, however, that a discontinuity can only occur at those time instants where the corner condition (3.28) is satisfied.

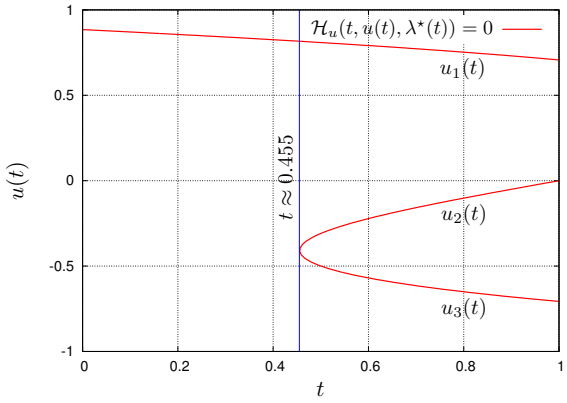


Figure 3.5. Values of the control variable $u(t)$, $0 \le t \le 1$, satisfying the Euler-Lagrange equations in Example 3.16.

3.4.4 Interpretation of the Adjoint Variables

We saw in Chapter 1 on nonlinear programming that the Lagrange multiplier associated to a particular constraint can be interpreted as the sensitivity of the objective function to a change in that constraint (see Remark 1.61, p. 30). Our objective in this subsection is to obtain a useful interpretation of the adjoint variables $\boldsymbol{\lambda}(t)$ which are associated to the state variables $\mathbf{x}(t)$.

Throughout this subsection, we consider an optimal control problem of the following form

$$\text{minimize:} \quad \mathcal{J}(\mathbf{u}) = \int_{t_0}^{t_\mathrm{f}} \ell(t, \mathbf{x}(t), \mathbf{u}(t))\, \mathrm{d}t \qquad (3.33)$$

$$\text{subject to:} \quad \dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)); \quad \mathbf{x}(t_0) = \mathbf{x}_0, \qquad (3.34)$$

with fixed initial time t_0 and terminal time t_f, where ℓ and $\mathbf{f}$ are continuous in $(t, \mathbf{x}, \mathbf{u})$, and have continuous first partial derivatives with respect to $\mathbf{x}$ and $\mathbf{u}$, for all $(t, \mathbf{x}, \mathbf{u}) \in [t_0, t_\mathrm{f}] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u}$. Let $\mathcal{V}(\mathbf{x}_0, t_0)$ denote the minimum value of $\mathcal{J}(\mathbf{u})$, for a given initial state $\mathbf{x}_0$ at t_0. For simplicity, suppose that $\mathbf{u}^\star \in \mathcal{C}[t_0, t_\mathrm{f}]^{n_u}$ is the unique control providing this minimum, and let $\mathbf{x}^\star \in \mathcal{C}^1[t_0, t_\mathrm{f}]^{n_x}$ and $\boldsymbol{\lambda}^\star \in \mathcal{C}^1[t_0, t_\mathrm{f}]^{n_x}$ denote the corresponding response and adjoint trajectories, respectively.

Now, consider a modification of the optimal control problem (3.33,3.34) in which the initial state is $\mathbf{x}_0 + \boldsymbol{\xi}$, with $\boldsymbol{\xi} \in \mathbb{R}^{n_x}$. Suppose that a unique optimal control, $\mathbf{v}(t; \boldsymbol{\xi})$, exists for the perturbed problem for each $\boldsymbol{\xi} \in \mathcal{B}_\delta(\mathbf{0})$, with $\delta > 0$, and let $\mathbf{y}(t; \boldsymbol{\xi})$ denote the corresponding optimal response, i.e.,

$$\dot{\mathbf{y}}(t; \boldsymbol{\xi}) = \mathbf{f}(t, \mathbf{y}(t; \boldsymbol{\xi}), \mathbf{v}(t; \boldsymbol{\xi})); \quad \mathbf{y}(t_0; \boldsymbol{\xi}) = \mathbf{x}_0 + \boldsymbol{\xi}.$$

Clearly, $\mathbf{v}(t; \mathbf{0}) = \mathbf{u}^\star(t)$ and $\mathbf{y}(t; \mathbf{0}) = \mathbf{x}^\star(t)$, for $t_0 \le t \le t_\mathrm{f}$. Suppose further that the functions $\mathbf{v}(t; \boldsymbol{\xi})$ and $\mathbf{y}(t; \boldsymbol{\xi})$ are continuously differentiable with respect to $\boldsymbol{\xi}$ on $\mathcal{B}_\delta(\mathbf{0})$.

Appending the differential equation (3.34) in (3.33) with the adjoint variable $\boldsymbol{\lambda}^\star$, we have

$$\begin{aligned}\mathcal{V}(\mathbf{y}(t_0; \boldsymbol{\xi}), t_0) &:= \int_{t_0}^{t_\mathrm{f}} \ell(t, \mathbf{y}(t; \boldsymbol{\xi}), \mathbf{v}(t; \boldsymbol{\xi}))\, \mathrm{d}t \\ &= \int_{t_0}^{t_\mathrm{f}} \Big(\ell(t, \mathbf{y}(t; \boldsymbol{\xi}), \mathbf{v}(t; \boldsymbol{\xi})) + \boldsymbol{\lambda}^\star(t)^\mathsf{T} \left[\mathbf{f}(t, \mathbf{y}(t; \boldsymbol{\xi}), \mathbf{v}(t; \boldsymbol{\xi})) - \dot{\mathbf{y}}(t; \boldsymbol{\xi}) \right] \Big)\, \mathrm{d}t.\end{aligned}$$

Then, upon differentiation of $\mathcal{V}(\mathbf{x}_0 + \boldsymbol{\xi}, t_0)$ with respect to $\boldsymbol{\xi}$, we obtain

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\xi}} \mathcal{V}(\mathbf{y}(t_0; \boldsymbol{\xi}), t_0) = &\int_{t_0}^{t_\mathrm{f}} \Big[\ell_{\mathbf{u}}(t, \mathbf{y}(t; \boldsymbol{\xi}), \mathbf{v}(t; \boldsymbol{\xi})) + \boldsymbol{\lambda}^\star(t)^\mathsf{T} \mathbf{f}_{\mathbf{u}}(t, \mathbf{y}(t; \boldsymbol{\xi}), \mathbf{v}(t; \boldsymbol{\xi})) \Big]^\mathsf{T} \mathbf{v}_{\boldsymbol{\xi}}(t; \boldsymbol{\xi})\, \mathrm{d}t \\ &+ \int_{t_0}^{t_\mathrm{f}} \Big[\ell_{\mathbf{x}}(t, \mathbf{y}(t; \boldsymbol{\xi}), \mathbf{v}(t; \boldsymbol{\xi})) + \boldsymbol{\lambda}^\star(t)^\mathsf{T} \mathbf{f}_{\mathbf{x}}(t, \mathbf{y}(t; \boldsymbol{\xi}), \mathbf{v}(t; \boldsymbol{\xi})) + \dot{\boldsymbol{\lambda}}^\star(t) \Big]^\mathsf{T} \mathbf{y}_{\boldsymbol{\xi}}(t; \boldsymbol{\xi})\, \mathrm{d}t \\ &- \boldsymbol{\lambda}^\star(t_\mathrm{f})^\mathsf{T} \mathbf{y}_{\boldsymbol{\xi}}(t_\mathrm{f}; \boldsymbol{\xi}) + \boldsymbol{\lambda}^\star(t_0)^\mathsf{T} \mathbf{y}_{\boldsymbol{\xi}}(t_0; \boldsymbol{\xi}),\end{aligned}$$

and taking the limit as $\boldsymbol{\xi} \to \mathbf{0}$ yields

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\xi}} \mathcal{V}(\mathbf{v}(t_0; \boldsymbol{\xi}),\, t_0)\Big|_{\boldsymbol{\xi}=\mathbf{0}} = &\int_{t_0}^{t_\mathrm{f}} \Big[\ell_{\mathbf{u}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t)) + \boldsymbol{\lambda}^\star(t)^\mathsf{T} \mathbf{f}_{\mathbf{u}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t)) \Big]^\mathsf{T} \mathbf{v}_{\boldsymbol{\xi}}(t; \mathbf{0})\, \mathrm{d}t \\ &+ \int_{t_0}^{t_\mathrm{f}} \Big[\ell_{\mathbf{x}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t)) + \boldsymbol{\lambda}^\star(t)^\mathsf{T} \mathbf{f}_{\mathbf{x}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t)) + \dot{\boldsymbol{\lambda}}^\star(t) \Big]^\mathsf{T} \mathbf{y}_{\boldsymbol{\xi}}(t; \mathbf{0})\, \mathrm{d}t \\ &- \boldsymbol{\lambda}^\star(t_\mathrm{f})^\mathsf{T} \mathbf{y}_{\boldsymbol{\xi}}(t_\mathrm{f}; \mathbf{0}) + \boldsymbol{\lambda}^\star(t_0)^\mathsf{T} \mathbf{y}_{\boldsymbol{\xi}}(t_0; \mathbf{0}).\end{aligned}$$

Finally, noting that the triple $(\mathbf{u}^\star, \mathbf{x}^\star, \boldsymbol{\lambda}^\star)$ satisfies the Euler-Lagrange equations (3.13–3.15), and since $\mathbf{v}(t_0; \boldsymbol{\xi}) = I_{n_x}$, we are left with

$$\boldsymbol{\lambda}^\star(t_0) = \frac{\partial}{\partial \boldsymbol{\xi}} \mathcal{V}(\mathbf{v}(t_0; \boldsymbol{\xi}), t_0)\Big|_{\boldsymbol{\xi}=\mathbf{0}} = \mathcal{V}_{\mathbf{x}}(\mathbf{x}_0, t_0). \tag{3.35}$$

That is, *the adjoint variable $\boldsymbol{\lambda}(t_0)$ at initial time can be interpreted as the sensitivity of the cost functional to a change in the initial condition $\mathbf{x}_0$.* In other words, $\boldsymbol{\lambda}(t_0)$ represents the marginal valuation in the optimal control problem of the state at initial time.

The discussion thus far has only considered the adjoint variables at initial time. We shall now turn to the interpretation of the adjoint variables $\boldsymbol{\lambda}(t)$ at any time $t_0 \leq t \leq t_\mathrm{f}$. We start by proving the so-called *principle of optimality*, which asserts that *any restriction on* $[t_1, t_\mathrm{f}]$ *of an optimal control on* $[t_0, t_\mathrm{f}]$ *is itself optimal*, for any $t_1 \geq t_0$.

Lemma 3.17 (Principle of Optimality). *Let* $\mathbf{u}^\star \in \hat{\mathcal{C}}[t_0, t_\mathrm{f}]^{n_u}$ *be an optimal control for the problem to*

$$\textit{minimize:} \quad \int_{t_0}^{t_\mathrm{f}} \ell(t, \mathbf{x}(t), \mathbf{u}(t))\, \mathrm{d}t \tag{3.36}$$

$$\textit{subject to:} \quad \dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)); \quad \mathbf{x}(t_0) = \mathbf{x}_0, \tag{3.37}$$

and let $\mathbf{x}^\star \in \hat{\mathcal{C}}^1[t_0, t_\mathrm{f}]^{n_x}$ *denote the corresponding optimal response. Then, for any* $t_1 \in [t_0, t_\mathrm{f}]$*, the restriction* $\mathbf{u}^\star(t)$*,* $t_1 \leq t \leq t_\mathrm{f}$*, is an optimal control for the problem to*

$$\textit{minimize:} \quad \int_{t_1}^{t_\mathrm{f}} \ell(t, \mathbf{x}(t), \mathbf{u}(t))\, \mathrm{d}t \tag{3.38}$$

$$\textit{subject to:} \quad \dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)); \quad \mathbf{x}(t_1) = \mathbf{x}^\star(t_1). \tag{3.39}$$

Proof. Let $\mathcal{V}(x_0, t_0)$ denote the minimum values of the optimal control problem (3.36,3.37). Clearly,

$$\mathcal{V}(x_0, t_0) = \int_{t_0}^{t_1} \ell(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t))\, \mathrm{d}t + \int_{t_1}^{t_\mathrm{f}} \ell(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t))\, \mathrm{d}t.$$

By contradiction, suppose that the restriction $\mathbf{u}^\star(t)$, $t_1 \leq t \leq t_\mathrm{f}$, is not optimal for the problem (3.38,3.39). Then, there exists a (feasible) control $\mathbf{u}^\dagger(t)$, $t_1 \leq t \leq t_\mathrm{f}$, that imparts to the functional (3.38) the value

$$\int_{t_1}^{t_\mathrm{f}} \ell(t, \mathbf{x}^\dagger(t), \mathbf{u}^\dagger(t))\, \mathrm{d}t < \int_{t_1}^{t_\mathrm{f}} \ell(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t))\, \mathrm{d}t.$$

Further, by joining $\mathbf{u}^\star(t)$, $t_0 \leq t \leq t_1$, and $\mathbf{u}^\dagger(t)$, $t_1 \leq t \leq t_\mathrm{f}$, one obtains a piecewise continuous control that is feasible and satisfies

$$\int_{t_0}^{t_1} \ell(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t))\, \mathrm{d}t + \int_{t_1}^{t_\mathrm{f}} \ell(t, \mathbf{x}^\dagger(t), \mathbf{u}^\dagger(t))\, \mathrm{d}t < \mathcal{V}(x_0, t_0),$$

hence contradicting the optimality of $\mathbf{u}^\star$, $t_0 \leq t \leq t_\mathrm{f}$, for the problem (3.36,3.37). □

Back to the question of the interpretation of the adjoint variables, the method used to reach (3.35) at initial time can be applied to the restricted optimal control problem (3.38,3.39), which by Lemma 3.17 has optimal solution $\mathbf{u}^\star(t)$, $t_1 \leq t \leq t_\mathrm{f}$. This method leads to the result

$$\boldsymbol{\lambda}^\star(t_1) = \mathcal{V}_\mathbf{x}(\mathbf{x}^\star(t_1), t_1),$$

and since the time t_1 was arbitrary, we get

$$\boldsymbol{\lambda}^\star(t) = \mathcal{V}_\mathbf{x}(\mathbf{x}^\star(t), t), \quad t_0 \leq t \leq t_\mathrm{f}.$$

That is, *if there were an exogenous, tiny perturbation to the state variable at time* t *and if the control were modified optimally thereafter, the optimal cost value would change at the rate*

$\boldsymbol{\lambda}(t)$; said differently, $\boldsymbol{\lambda}(t)$ is the marginal valuation in the optimal control problem of the state variable at time t. In particular, the optimal cost value remains unchanged in case of an exogenous perturbation at terminal time t_f, i.e., the rate of change is $\mathcal{V}_\mathbf{x}(\mathbf{x}^\star(t_\mathrm{f}), t_\mathrm{f}) = 0$. This interpretation confirms the natural boundary conditions of the adjoint variables in the Euler-Lagrange equations (3.13–3.15).

3.4.5 General Terminal Constraints

So far, we have only considered optimal control problems with fixed initial time t_0 and terminal time t_f, and free terminal state variables $\mathbf{x}(t_\mathrm{f})$. However, many optimal control problems do not fall into this formulation. Often, the terminal time is free (e.g., in minimum time problems), and the state variables at final time are either fixed or constrained to lie on a smooth manifold.

In this subsection, we shall consider problems having end-point constraints of the form $\boldsymbol{\psi}(t_\mathrm{f}, \mathbf{x}(t_\mathrm{f})) = \mathbf{0}$, with t_f being specified or not. Besides terminal constraints, we shall also add a *terminal cost* (or *salvage term*) $\phi(t_\mathrm{f}, \mathbf{x}(t_\mathrm{f}))$ to the cost functional, so that the problem is now in the Bolza form. In the case of a free terminal time problem, t_f shall be considered an additional variable in the optimization problem. Similar to free end-point problems of the calculus of variations (see, e.g., §2.5.5 and §2.7.3), we shall then define the optimization horizon by extension on a "sufficiently" large interval $[t_0, T]$, and consider the linear space $\mathcal{C}[t_0, T]^{n_u} \times \mathbb{R}$, supplied with the norm $\|(\mathbf{u}, t)\|_\infty := \|\mathbf{u}\|_\infty + |t|$, as the class of admissible controls for the problem.

In order to obtain necessary conditions of optimality for problems with terminal constraints, the idea is to apply the method of Lagrange multipliers described in §2.7.1, by specializing the normed linear space $(\mathcal{X}, \|\cdot\|)$ to $(\mathcal{C}[t_0, T]^{n_u} \times \mathbb{R}, \|(\cdot,\cdot)\|_\infty)$ and considering the Gâteaux derivative $\delta\mathcal{J}(\mathbf{u}, t_\mathrm{f}; \boldsymbol{\omega}, \tau)$ at any point $(\mathbf{u}, t_\mathrm{f})$ and in any direction $(\boldsymbol{\omega}, \tau)$. One such set of necessary conditions is given by the following:

Theorem 3.18 (Necessary Conditions for Problems having Equality Terminal Constraints). *Consider the optimal control problem to*

$$\textit{minimize:}\quad \mathcal{J}(\mathbf{u}, t_\mathrm{f}) := \int_{t_0}^{t_\mathrm{f}} \ell(t, \mathbf{x}(t), \mathbf{u}(t))\,\mathrm{d}t + \phi(t_\mathrm{f}, \mathbf{x}(t_\mathrm{f})) \tag{3.40}$$

$$\textit{subject to:}\quad \mathcal{P}_k(\mathbf{u}, t_\mathrm{f}) := \psi_k(t_\mathrm{f}, \mathbf{x}(t_\mathrm{f})) = 0, \quad k = 1, \ldots, n_\psi \tag{3.41}$$

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)); \quad \mathbf{x}(t_0) = \mathbf{x}_0, \tag{3.42}$$

for $\mathbf{u} \in \mathcal{C}[t_0, T]^{n_u}$, with fixed initial time t_0 and free terminal time t_f; ℓ and $\mathbf{f}$ are continuous in $(t, \mathbf{x}, \mathbf{u})$ and have continuous first partial derivatives with respect to $\mathbf{x}$ and $\mathbf{u}$ for all $(t, \mathbf{x}, \mathbf{u}) \in [t_0, T] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u}$; ϕ and $\boldsymbol{\psi}$ are continuous and have continuous first partial derivatives with respect to t and $\mathbf{x}$ for all $(t, \mathbf{x}) \in [t_0, T] \times \mathbb{R}^{n_x}$. Suppose that $(\mathbf{u}^\star, t_\mathrm{f}^\star) \in \mathcal{C}[t_0, T]^{n_u} \times [t_0, T)$ is a (local) minimizer for the problem, and let $\mathbf{x}^\star \in \mathcal{C}^1[t_0, T]^{n_x}$ denote the corresponding response. Suppose further that

$$\begin{vmatrix} \delta\mathcal{P}_1(\mathbf{u}^\star, t_\mathrm{f}^\star; \bar{\boldsymbol{\omega}}_1, \bar{\tau}_1) & \cdots & \delta\mathcal{P}_1(\mathbf{u}^\star, t_\mathrm{f}^\star; \bar{\boldsymbol{\omega}}_{n_\psi}, \bar{\tau}_{n_\psi}) \\ \vdots & \ddots & \vdots \\ \delta\mathcal{P}_{n_\psi}(\mathbf{u}^\star, t_\mathrm{f}^\star; \bar{\boldsymbol{\omega}}_1, \bar{\tau}_1) & \cdots & \delta\mathcal{P}_{n_\psi}(\mathbf{u}^\star, t_\mathrm{f}^\star; \bar{\boldsymbol{\omega}}_{n_\psi}, \bar{\tau}_{n_\psi}) \end{vmatrix} \neq 0, \tag{3.43}$$

for n_ψ (independent) directions $(\bar{\boldsymbol{\omega}}_1, \bar{\tau}_1), \ldots, (\bar{\boldsymbol{\omega}}_{n_\psi}, \bar{\tau}_{n_\psi}) \in \mathcal{C}[t_0, T]^{n_u} \times \mathbb{R}$. Then, there exist a function $\boldsymbol{\lambda}^\star \in \mathcal{C}^1[t_0, T]^{n_x}$ and a vector $\boldsymbol{\nu}^\star \in \mathbb{R}^{n_\psi}$ such that $(\mathbf{u}^\star, \mathbf{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star, t_\mathrm{f}^\star)$

satisfies the Euler-Lagrange equations

$$\dot{\mathbf{x}}(t) = \mathcal{H}_{\boldsymbol{\lambda}}(t, \mathbf{x}(t), \mathbf{u}(t), \boldsymbol{\lambda}(t)); \qquad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{3.44}$$

$$\dot{\boldsymbol{\lambda}}(t) = -\mathcal{H}_{\mathbf{x}}(t, \mathbf{x}(t), \mathbf{u}(t), \boldsymbol{\lambda}(t)); \qquad \boldsymbol{\lambda}(t_{\mathrm{f}}) = \Phi_{\mathbf{x}}(t_{\mathrm{f}}, \mathbf{x}(t_{\mathrm{f}})) \tag{3.45}$$

$$\mathbf{0} = \mathcal{H}_{\mathbf{u}}(t, \mathbf{x}(t), \mathbf{u}(t), \boldsymbol{\lambda}(t)), \tag{3.46}$$

for all $t_0 \le t \le t_{\mathrm{f}}$, along with the conditions

$$\boldsymbol{\psi}(t_{\mathrm{f}}, \mathbf{x}(t_{\mathrm{f}})) = \mathbf{0} \tag{3.47}$$

$$\Phi_t(t_{\mathrm{f}}, \mathbf{x}(t_{\mathrm{f}})) + \mathcal{H}(t_{\mathrm{f}}, \mathbf{x}(t_{\mathrm{f}}), \mathbf{u}(t_{\mathrm{f}}), \boldsymbol{\lambda}(t_{\mathrm{f}})) = \mathbf{0}, \tag{3.48}$$

with $\Phi := \phi + \boldsymbol{\nu}^{\mathsf{T}}\boldsymbol{\psi}$ and $\mathcal{H} := \ell + \boldsymbol{\lambda}^{\mathsf{T}}\mathbf{f}$.

Proof. Consider a one-parameter family of comparison controls $\mathbf{v}(t;\eta) := \mathbf{u}^\star(t) + \eta\boldsymbol{\omega}(t)$, where $\boldsymbol{\omega}(t) \in \mathcal{C}[t_0, T]^{n_u}$ is some fixed function, and η is a (scalar) parameter. Let $\mathbf{y}(t;\eta) \in \mathcal{C}^1[t_0, T]^{n_x}$ be the response corresponding to $\mathbf{v}(t;\eta)$ through (3.42). In particular, $\eta = 0$ provides the optimal response $\mathbf{y}(t;0) \equiv \mathbf{x}^\star(t)$, $t_0 \le t \le t_{\mathrm{f}}$.

We show, as in the proof of Theorem 3.9, that the Gâteaux variation at $(\mathbf{u}^\star, t_{\mathrm{f}}^\star)$ in any direction $(\boldsymbol{\omega}, \tau) \in \mathcal{C}[t_0, T]^{n_u} \times \mathbb{R}$ of the cost functional $\mathcal{J}$ subject to the initial value problem (3.42) is given by

$$\begin{aligned}\delta\mathcal{J}(\mathbf{u}^\star, t_{\mathrm{f}}^\star; \boldsymbol{\omega}, \tau) = &\int_{t_0}^{t_{\mathrm{f}}^\star} \left[\ell_{\mathbf{u}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t)) + \mathbf{f}_{\mathbf{u}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t))^{\mathsf{T}}\boldsymbol{\lambda}^{(0)}(t)\right]^{\mathsf{T}} \boldsymbol{\omega}(t)\, \mathrm{d}t \\ &+ \left[\phi_t(t_{\mathrm{f}}^\star, \mathbf{x}^\star(t_{\mathrm{f}}^\star)) + \ell(t_{\mathrm{f}}^\star, \mathbf{x}^\star(t_{\mathrm{f}}^\star), \mathbf{u}^\star(t_{\mathrm{f}}^\star)) + \mathbf{f}(t_{\mathrm{f}}^\star, \mathbf{x}^\star(t_{\mathrm{f}}^\star), \mathbf{u}^\star(t_{\mathrm{f}}^\star))^{\mathsf{T}}\boldsymbol{\lambda}^{(0)}(t)\right] \tau,\end{aligned}$$

with the adjoint variables $\boldsymbol{\lambda}^{(0)}(t)$, $t_0 \le t \le t_{\mathrm{f}}^\star$, calculated as

$$\dot{\boldsymbol{\lambda}}^{(0)}(t) = -\ell_{\mathbf{x}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t)) - \mathbf{f}_{\mathbf{x}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t))^{\mathsf{T}}\boldsymbol{\lambda}^{(0)}(t); \quad \boldsymbol{\lambda}^{(0)}(t_{\mathrm{f}}^\star) = \phi_{\mathbf{x}}(t_{\mathrm{f}}^\star, \mathbf{x}^\star(t_{\mathrm{f}}^\star)).$$

On the other hand, the Gâteaux variations at $(\mathbf{u}^\star, t_{\mathrm{f}}^\star)$ in any direction $(\boldsymbol{\omega}, \tau) \in \mathcal{C}[t_0, T]^{n_u} \times \mathbb{R}$ of the functionals $\mathcal{P}_k$ subject to the initial value problem (3.42) is given by

$$\begin{aligned}\delta\mathcal{P}_k(\mathbf{u}^\star, t_{\mathrm{f}}^\star; \boldsymbol{\omega}, \tau) = &\int_{t_0}^{t_{\mathrm{f}}^\star} \left[\mathbf{f}_{\mathbf{u}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t))^{\mathsf{T}}\boldsymbol{\lambda}^{(k)}(t)\right]^{\mathsf{T}} \boldsymbol{\omega}(t)\, \mathrm{d}t \\ &+ \left[(\psi_k)_t(t_{\mathrm{f}}^\star, \mathbf{x}^\star(t_{\mathrm{f}}^\star)) + \mathbf{f}(t_{\mathrm{f}}^\star, \mathbf{x}^\star(t_{\mathrm{f}}^\star), \mathbf{u}^\star(t_{\mathrm{f}}^\star))^{\mathsf{T}}\boldsymbol{\lambda}^{(k)}(t)\right] \tau,\end{aligned}$$

with the adjoint variables $\boldsymbol{\lambda}^{(k)}(t)$, $t_0 \le t \le t_{\mathrm{f}}^\star$, given by

$$\dot{\boldsymbol{\lambda}}^{(k)}(t) = -\mathbf{f}_{\mathbf{x}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t))^{\mathsf{T}}\boldsymbol{\lambda}^{(k)}(t); \quad \boldsymbol{\lambda}^{(k)}(t_{\mathrm{f}}^\star) = (\psi_k)_{\mathbf{x}}(t_{\mathrm{f}}^\star, \mathbf{x}^\star(t_{\mathrm{f}}^\star)), \tag{3.49}$$

for each $k = 1, \ldots, n_\psi$.

Note that, based on the differentiability assumptions on ℓ, ϕ, $\boldsymbol{\psi}$ and $\mathbf{f}$, the Gâteaux derivatives $\delta\mathcal{J}(\mathbf{u}^\star, t_{\mathrm{f}}^\star; \boldsymbol{\omega}, \tau)$ and $\delta\mathcal{P}_k(\mathbf{u}^\star, t_{\mathrm{f}}^\star; \boldsymbol{\omega}, \tau)$ exist and are continuous in each direction $(\boldsymbol{\omega}, \tau) \in \mathcal{C}[t_0, T]^{n_u} \times \mathbb{R}$. Since $(\mathbf{u}^\star, t_{\mathrm{f}}^\star)$ gives a (local) minimum for (3.40–3.42) and condition (3.43) holds at $(\mathbf{u}^\star, t_{\mathrm{f}}^\star)$, by Theorem 2.47 (and Remark 2.49), there exists a vector $\boldsymbol{\nu}^\star \in \mathbb{R}^{n_\psi}$ such that

$$\begin{aligned}0 &= \delta\left(\mathcal{J} + \textstyle\sum_{k=1}^{n_\psi} \nu_k^\star \mathcal{P}_k\right)(\mathbf{u}^\star, t_{\mathrm{f}}^\star; \boldsymbol{\xi}, \tau) \\ &= \int_{t_0}^{t_{\mathrm{f}}^\star} \mathcal{H}_{\mathbf{u}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t), \boldsymbol{\lambda}^\star(t))^{\mathsf{T}}\boldsymbol{\omega}(t)\, \mathrm{d}t \\ &\quad + \left[\Phi_t(t_{\mathrm{f}}^\star, \mathbf{x}^\star(t_{\mathrm{f}}^\star)) + \mathcal{H}(t_{\mathrm{f}}^\star, \mathbf{x}^\star(t_{\mathrm{f}}^\star), \mathbf{u}^\star(t_{\mathrm{f}}^\star), \boldsymbol{\lambda}^\star(t_{\mathrm{f}}^\star))\right] \tau,\end{aligned}$$

for each $(\boldsymbol{\omega}, \tau) \in \mathcal{C}[t_0, T]^{n_u} \times \mathbb{R}$, where $\Phi := \phi + \boldsymbol{\nu}^{\star\mathsf{T}}\boldsymbol{\psi}$, $\boldsymbol{\lambda}^\star := \boldsymbol{\lambda}^{(0)} + \sum_{k=1}^{n_\psi} \nu_k^\star \boldsymbol{\lambda}^{(k)}$, and $\mathcal{H} := \ell + \boldsymbol{\lambda}^{\star\mathsf{T}}\mathbf{f}$. In particular, taking $\tau := 0$ and restricting attention to $\boldsymbol{\omega}(t)$ such that $\omega_i(t) := \mathcal{H}_{u_i}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t), \boldsymbol{\lambda}^\star(t))$ and $\omega_j(t) := 0$ for $j \neq i$, we get the necessary conditions of optimality

$$\mathcal{H}_{u_i}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t), \boldsymbol{\lambda}^\star(t)) = 0, \quad i = 1, \ldots, n_x,$$

for each $t_0 \leq t \leq t_\mathrm{f}^\star$. On the other hand, choosing $\boldsymbol{\omega}(t) := \mathbf{0}$, $t_0 \leq t \leq t_\mathrm{f}^\star$, and $\tau := \Phi_t(t_\mathrm{f}^\star, \mathbf{x}^\star(t_\mathrm{f}^\star)) + \mathcal{H}(t_\mathrm{f}^\star, \mathbf{x}^\star(t_\mathrm{f}^\star), \mathbf{u}^\star(t_\mathrm{f}^\star), \boldsymbol{\lambda}^\star(t_\mathrm{f}^\star))$, yields the transversal condition

$$\Phi_t(t_\mathrm{f}^\star, \mathbf{x}^\star(t_\mathrm{f}^\star)) + \mathcal{H}(t_\mathrm{f}^\star, \mathbf{x}^\star(t_\mathrm{f}^\star), \mathbf{u}^\star(t_\mathrm{f}^\star), \boldsymbol{\lambda}^\star(t_\mathrm{f}^\star)) = 0.$$

Finally, since the adjoint differential equations giving $\boldsymbol{\lambda}^{(0)}$ and $\boldsymbol{\lambda}^{(k)}$, $k = 1, \ldots, n_\psi$, are linear, the adjoint variables $\boldsymbol{\lambda}^\star$ must satisfy the following differential equations and corresponding terminal conditions

$$\dot{\boldsymbol{\lambda}}^\star(t) = -\mathcal{H}_{\mathbf{x}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t), \boldsymbol{\lambda}^\star(t)); \quad \boldsymbol{\lambda}^\star(t_\mathrm{f}^\star) = \Phi_{\mathbf{x}}(t_\mathrm{f}^\star, \mathbf{x}^\star(t_\mathrm{f}^\star)),$$

for all $t_0 \leq t \leq t_\mathrm{f}^\star$. □

Observe that the optimality conditions consist of n_u algebraic equations (3.46), together with $2 \times n_x$ ODEs (3.44,3.45) and their respective boundary conditions, which determine the optimal control $\mathbf{u}^\star$, response $\mathbf{x}^\star$, and adjoint trajectories $\boldsymbol{\lambda}^\star$. Here again, these equations yield a challenging TPBVP. In addition, the necessary conditions (3.47) determine the optimal Lagrange multiplier vector $\boldsymbol{\nu}^\star$. Finally, for problems with free terminal time, the transversal condition (3.48) determines the optimal terminal time $t_\mathrm{f}^\star$. We thus have a complete set of necessary conditions.

Remark 3.19 (Reachability Condition). One of the most difficult aspect in applying Theorem 3.18 is to verify that the terminal constraints satisfy the regularity condition (3.43). In order to gain insight into this condition, consider an optimal control problem of the form (3.40–3.42), with fixed terminal time t_f and a single terminal state constraint $\mathcal{P}(\mathbf{u}) := x_j(t_\mathrm{f}) - x_{\mathrm{f}j} = 0$, for some $j \in \{1, \ldots, n_x\}$. The Gâteaux variation of $\mathcal{P}$ at $\mathbf{u}^\star$ in any direction $\boldsymbol{\omega} \in \mathcal{C}[t_0, t_\mathrm{f}]^{n_u}$ is

$$\delta\mathcal{P}(\mathbf{u}^\star; \boldsymbol{\omega}) = \int_{t_0}^{t_\mathrm{f}} \left[\mathbf{f}_{\mathbf{u}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t))^{\mathsf{T}} \boldsymbol{\lambda}^{(j)}(t)\right]^{\mathsf{T}} \boldsymbol{\omega}(t)\, \mathrm{d}t,$$

where $\dot{\boldsymbol{\lambda}}^{(j)}(t) = -\mathbf{f}_{\mathbf{x}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t))^{\mathsf{T}} \boldsymbol{\lambda}^{(j)}(t)$, $t_0 \leq t \leq t_\mathrm{f}$, with terminal conditions $\lambda_j^{(j)}(t_\mathrm{f}) = 1$ and $\lambda_i^{(j)}(t_\mathrm{f}) = 0$ for $i \neq j$. By choosing $\boldsymbol{\omega}(t) := \mathbf{f}_{\mathbf{u}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t))^{\mathsf{T}} \boldsymbol{\lambda}^{(j)}(t)$, $t_0 \leq t \leq t_\mathrm{f}$, we obtain the following sufficient condition for the terminal constraint to be regular:

$$\int_{t_0}^{t_\mathrm{f}} \boldsymbol{\lambda}^{(j)}(t)^{\mathsf{T}} \mathbf{f}_{\mathbf{u}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t)) \mathbf{f}_{\mathbf{u}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t))^{\mathsf{T}} \boldsymbol{\lambda}^{(j)}(t)\, \mathrm{d}t \neq 0. \tag{3.50}$$

This condition can be seen as a *reachability condition*[3] for the system. In other words, if this condition does not hold, then it may *not* be possible to find a control $\mathbf{u}(t)$ so that the terminal condition $x_j(t_\mathrm{f}) = x_{\mathrm{f}j}$ be satisfied at final time.

[3]Reachability is defined as follow (see, e.g., [2]):

Definition 3.20 (Reachability). *A state $\mathbf{x}_\mathrm{f}$ is said to be* reachable *at time t_f, if for some finite $t_0 < t_\mathrm{f}$ there exists an input $\mathbf{u}(t)$, $t_0 \leq t \leq t_\mathrm{f}$, that transfers the state $\mathbf{x}(t)$ from the origin at t_0, to $\mathbf{x}_\mathrm{f}$ at time t_f.*

More generally, for optimal control problems such as (3.40–3.42) with fixed terminal time t_f and n_ψ terminal constraints $\psi_k(\mathbf{x}(t_\mathrm{f})) = 0$, the foregoing sufficient condition becomes

$$\operatorname{rank} \boldsymbol{\Psi} = n_\psi,$$

where $\boldsymbol{\Psi}$ is a $(n_\psi \times n_\psi)$ matrix defined by

$$\boldsymbol{\Psi}_{ij} := \int_{t_0}^{t_\mathrm{f}} \boldsymbol{\lambda}^{(i)}(t)^\mathsf{T} \mathbf{f}_\mathbf{u}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t)) \mathbf{f}_\mathbf{u}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t))^\mathsf{T} \boldsymbol{\lambda}^{(j)}(t)\, \mathrm{d}t, \quad 1 \le i, j \le n_\psi,$$

with the adjoint variables $\boldsymbol{\lambda}^{(k)}(t)$, $t_0 \le t \le t_\mathrm{f}$, defined by (3.49).

A summary of the necessary conditions of optimality encountered so far is given hereafter, before an example is considered.

Remark 3.21 (Summary of Necessary Conditions). Necessary conditions of optimality for the problem

$$\begin{aligned} \text{minimize:} \quad & \int_{t_0}^{t_\mathrm{f}} \ell(t, \mathbf{x}(t), \mathbf{u}(t))\, \mathrm{d}t + \phi(t_\mathrm{f}, \mathbf{x}(t_\mathrm{f})) \\ \text{subject to:} \quad & \boldsymbol{\psi}(t_\mathrm{f}, \mathbf{x}(t_\mathrm{f})) = \mathbf{0}, \\ & \dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)); \quad \mathbf{x}(t_0) = \mathbf{x}_0, \end{aligned}$$

are as follows:

- Euler-Lagrange Equations:

$$\left.\begin{aligned} \dot{\mathbf{x}} &= \mathcal{H}_{\boldsymbol{\lambda}} \\ \dot{\boldsymbol{\lambda}} &= -\mathcal{H}_\mathbf{x} \\ \mathbf{0} &= \mathcal{H}_\mathbf{u} \end{aligned}\right\}, \quad t_0 \le t \le t_\mathrm{f};$$

  with $\mathcal{H} := \ell + \boldsymbol{\lambda}^\mathsf{T} \mathbf{f}$.

- Legendre-Clebsch Condition:

$$\mathcal{H}_{\mathbf{uu}} \text{ semi-definite positive}, \quad t_0 \le t \le t_\mathrm{f};$$

- Transversal Conditions:

$$\begin{aligned} &\left[\mathcal{H} + \phi_t + \boldsymbol{\nu}^\mathsf{T} \boldsymbol{\psi}_t\right]_{t_\mathrm{f}} = 0, \text{ if } t_\mathrm{f} \text{ is free} \\ &\left[\boldsymbol{\lambda} - \phi_\mathbf{x} + \boldsymbol{\nu}^\mathsf{T} \boldsymbol{\psi}_\mathbf{x}\right]_{t_\mathrm{f}} = \mathbf{0} \\ &\left[\boldsymbol{\psi}\right]_{t_\mathrm{f}} = \mathbf{0}, \text{ and } \boldsymbol{\psi} \text{ satisfy a regularity condition;} \end{aligned}$$

Example 3.22. Consider the optimal control problem

$$\text{minimize:} \quad \mathcal{J}(u) := \int_0^1 \tfrac{1}{2} u(t)^2\, \mathrm{d}t \tag{3.51}$$

$$\text{subject to:} \quad \dot{x}(t) = u(t) - x(t); \quad x(0) = 1 \tag{3.52}$$

$$x(1) = 0. \tag{3.53}$$

We start by considering the reachability condition (3.50). The adjoint equation corresponding to the terminal constraint (3.53) is

$$\dot{\lambda}^{(1)}(t) = \lambda^{(1)}(t); \quad \lambda^{(1)}(1) = 1,$$

which has solution $\lambda^{(1)}(t) = \mathrm{e}^{t-1}$. That is, we have

$$\int_0^1 \boldsymbol{\lambda}^{(j)^\mathsf{T}} \mathbf{f_u} \mathbf{f_u}^\mathsf{T} \boldsymbol{\lambda}^{(j)} \, \mathrm{d}t = \int_0^1 \mathrm{e}^{2t-2} \, \mathrm{d}t = \frac{1-\mathrm{e}^2}{2} \neq 0.$$

Therefore, the terminal constraint (3.53) is regular and Theorem 3.18 applies.

The Hamiltonian function for the problem reads

$$\mathcal{H}(x,u,\lambda) \;=\; \tfrac{1}{2}u^2 + \lambda(u-x).$$

Candidate optimal solutions $(u^\star, x^\star, \lambda^\star)$ are those satisfying the Euler-Lagrange equations

$$\begin{aligned} \dot{x}(t) \;&= \mathcal{H}_\lambda \;=\; u(t) - x(t); & x(0) \;&= 1 \\ \dot{\lambda}(t) \;&= -\mathcal{H}_x \;=\; \lambda(t); & \lambda^\star(1) \;&= \nu \\ 0 \;&= \mathcal{H}_u \;=\; u(t) + \lambda(t). \end{aligned}$$

The adjoint equation has solution

$$\lambda^\star(t) \;=\; \nu^\star \mathrm{e}^{t-1},$$

and from the optimality condition, we get

$$u^\star(t) \;=\; -\nu^\star \mathrm{e}^{t-1}.$$

(Note that $u^\star$ is indeed a candidate *minimum* solution for the problem since $\mathcal{H}_{uu} = 1 > 0$ for each $0 \leq t \leq 1$.) Substituting the optimal control candidate back into (3.52) yields

$$\dot{x}^\star(t) \;=\; -\nu^\star \mathrm{e}^{t-1} - x(t); \quad x(0) \;=\; 1.$$

Upon integration of the state equation, and drawing the results together, we obtain

$$\begin{aligned} u^\star(t) \;&= -\nu^\star \mathrm{e}^{t-1} \\ x^\star(t) \;&= \mathrm{e}^{-t}\left[1 + \frac{\nu^\star}{2\mathrm{e}} - \frac{\nu^\star}{2}\mathrm{e}^{2t-1}\right] = \mathrm{e}^{-t} - \frac{\nu}{\mathrm{e}}\sinh(t) \\ \lambda^\star(t) \;&= \nu^\star \mathrm{e}^{t-1}. \end{aligned}$$

(One may also use the fact that $\mathcal{H} = const.$ along an optimal trajectory to obtain $x^\star(t)$.) Finally, the terminal condition $x^\star(1) = 0$ gives

$$\nu^\star = \frac{2}{\mathrm{e} - \mathrm{e}^{-1}} = \frac{1}{\sinh(1)}.$$

The optimal trajectories $u^\star(t)$ and $x^\star(t)$, $0 \leq t \leq 1$, are shown in Fig. 3.6. below.

Remark 3.23 (Problems having Inequality Terminal Constraints). In case the optimal control problem (3.40–3.42) has terminal *inequality* state constraints of the form

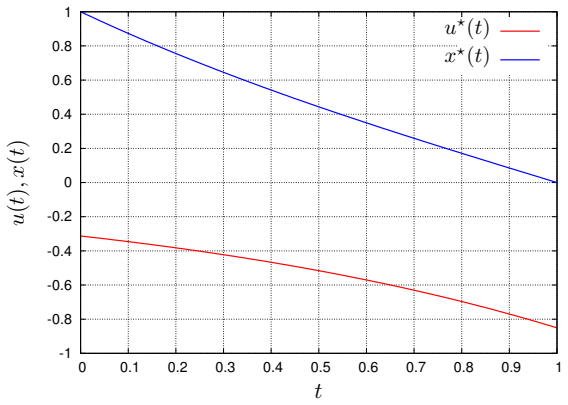


Figure 3.6. Optimal trajectories $u^\star(t)$ and $x^\star(t)$, $0 \le t \le 1$, in Example 3.22.

$\psi_k(t_\mathrm{f}, \mathbf{x}(t_\mathrm{f})) \le 0$ (in lieu of equality constraints), the conditions (3.44–3.46) and (3.48) remain necessary for optimality. On the other hand, the constraint conditions (3.47) shall be replaced by

$$\boldsymbol{\psi}(t_\mathrm{f}, \mathbf{x}(t_\mathrm{f})) \le \mathbf{0} \tag{3.54}$$

$$\boldsymbol{\nu} \ge \mathbf{0} \tag{3.55}$$

$$\boldsymbol{\nu}^\mathsf{T} \boldsymbol{\psi}(t_\mathrm{f}, \mathbf{x}(t_\mathrm{f})) = \mathbf{0}. \tag{3.56}$$

A proof is easily obtained upon invoking Theorem 2.51 instead of Theorem 2.47 in the proof of Theorem 3.18.

3.4.6 Application: Linear Time-Varying Systems with Quadratic Criteria

Consider the problem of bringing the state of a linear time-varying (LTV) system,

$$\dot{\mathbf{x}}(t) = \mathsf{A}(t)\mathbf{x}(t) + \mathsf{B}(t)\mathbf{u}(t), \tag{3.57}$$

with $\mathbf{x}(t) \in \mathbb{R}^{n_x}$ and $\mathbf{u}(t) \in \mathbb{R}^{n_u}$, from an initial state $\mathbf{x}(t_0) \neq \mathbf{0}$ to a terminal state

$$\mathbf{x}(t_\mathrm{f}) \approx \mathbf{0}, \qquad t_\mathrm{f} \text{ given},$$

using "acceptable" levels of the control $\mathbf{u}(t)$, and not exceeding "acceptable" levels of the state $\mathbf{x}(t)$ on the path.

A solution to this problem can be obtained by minimizing a performance index made up of a quadratic form in the terminal state plus an integral of quadratic forms in the state and controls:

$$\mathcal{J}(\mathbf{u}) := \int_{t_0}^{t_\mathrm{f}} \tfrac{1}{2} \left[\mathbf{u}(t)^\mathsf{T} \mathsf{Q}(t) \mathbf{u}(t) + \mathbf{x}(t)^\mathsf{T} \mathsf{R}(t) \mathbf{x}(t) \right] \, \mathrm{d}t + \tfrac{1}{2} \mathbf{x}(t_\mathrm{f})^\mathsf{T} \mathsf{S}_\mathrm{f} \mathbf{x}(t_\mathrm{f}), \tag{3.58}$$

where $\mathsf{S}_\mathrm{f} \succeq \mathbf{0}$, $\mathsf{R}(t) \succeq \mathbf{0}$, and $\mathsf{Q}(t) \succ \mathbf{0}$ are symmetric matrices. (In practice, these matrices must be so chosen that "acceptable" levels of $\mathbf{x}(t_\mathrm{f})$, $\mathbf{x}(t)$, and $\mathbf{u}(t)$ are obtained.)

Using the necessary conditions derived earlier in §3.4.5, a minimizing control $\mathbf{u}^\star$ for (3.58) subject to (3.57) must satisfy the Euler-Lagrange equations,

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \mathcal{H}_{\boldsymbol{\lambda}}(t,\mathbf{x}(t),\mathbf{u}(t),\boldsymbol{\lambda}(t)); \quad \mathbf{x}(t_0)=\mathbf{x}_0\\ \dot{\boldsymbol{\lambda}}(t) &= -\mathcal{H}_{\mathbf{x}}(t,\mathbf{x}(t),\mathbf{u}(t),\boldsymbol{\lambda}(t)); \quad \boldsymbol{\lambda}(t_\mathrm{f})=\mathsf{S}_\mathrm{f}\mathbf{x}(t_\mathrm{f})\\ \mathbf{0} &= \mathcal{H}_{\mathbf{u}}(t,\mathbf{x}(t),\mathbf{u}(t),\boldsymbol{\lambda}(t)),\end{aligned} \tag{3.59}$$

where

$$\mathcal{H}(t,\mathbf{x},\mathbf{u},\boldsymbol{\lambda}) = \tfrac{1}{2}\mathbf{u}^\mathsf{T}\mathsf{Q}(t)\mathbf{u} + \tfrac{1}{2}\mathbf{x}^\mathsf{T}\mathsf{R}(t)\mathbf{x} + \boldsymbol{\lambda}^\mathsf{T}[\mathsf{A}(t)\mathbf{x}+\mathsf{B}(t)\mathbf{u}].$$

Conversely, a control $\mathbf{u}^\star$ satisfying the Euler-Lagrange equations is a global optimal control since the differential equation is linear, the integral cost is jointly convex in $(\mathbf{x},\mathbf{u})$, and the terminal cost is convex in $\mathbf{x}$.

From (3.59), we have

$$\mathbf{u}^\star(t) = -\mathsf{Q}(t)^{-1}\mathsf{B}(t)^\mathsf{T}\boldsymbol{\lambda}^\star(t), \tag{3.60}$$

which in turn gives

$$\begin{bmatrix}\dot{\mathbf{x}}^\star(t)\\ \dot{\boldsymbol{\lambda}}^\star(t)\end{bmatrix} = \begin{bmatrix}\mathsf{A} & -\mathsf{B}\mathsf{Q}^{-1}\mathsf{B}^\mathsf{T}\\ -\mathsf{R} & -\mathsf{A}^\mathsf{T}\end{bmatrix}\begin{bmatrix}\mathbf{x}^\star(t)\\ \boldsymbol{\lambda}^\star(t)\end{bmatrix}; \quad \begin{matrix}\mathbf{x}^\star(t_0)=\mathbf{x}_0\\ \boldsymbol{\lambda}^\star(t_\mathrm{f})=\mathsf{S}_\mathrm{f}\mathbf{x}^\star(t_\mathrm{f}).\end{matrix} \tag{3.61}$$

Note that since these differential equations and the terminal boundary condition are homogeneous, their solutions $\mathbf{x}^\star(t)$ and $\boldsymbol{\lambda}^\star(t)$ are proportional to $\mathbf{x}(t_0)$.

An efficient method for solving the TPBVP (3.61) is the so-called *sweep method*. The idea is to determine the missing initial condition $\boldsymbol{\lambda}(t_0)$, so that (3.61) can be integrated forward in time as an initial value problem. For this, the coefficients of the terminal condition $\boldsymbol{\lambda}^\star(t_\mathrm{f}) = \mathsf{S}_\mathrm{f}\mathbf{x}^\star(t_\mathrm{f})$ are *swept* backward to the initial time, so that $\boldsymbol{\lambda}^\star(t_0) = \mathsf{S}(t_0)\mathbf{x}^\star(t_0)$. At intermediates times, substituting the relation $\boldsymbol{\lambda}^\star(t) = \mathsf{S}(t)\mathbf{x}^\star(t)$ into (3.61) yields the following *matrix Riccati equation*:

$$\dot{\mathsf{S}} = -\mathsf{S}\mathsf{A} - \mathsf{A}^\mathsf{T}\mathsf{S} + \mathsf{S}\mathsf{B}\mathsf{Q}^{-1}\mathsf{B}^\mathsf{T}\mathsf{S} - \mathsf{R}; \quad \mathsf{S}(t_\mathrm{f}) = \mathsf{S}_\mathrm{f}. \tag{3.62}$$

It is clear that $\mathsf{S}(t)$ is a symmetric matrix at each $t_0 \le t \le t_\mathrm{f}$ since S_f is symmetric and so is (3.62). By integrating (sweeping) (3.62) from t_f back to t_0, one gets

$$\boldsymbol{\lambda}^\star(t_0) = \mathsf{S}(t_0)\mathbf{x}^\star(t_0),$$

which may be regarded as the equivalent of the boundary terminal condition in (3.61) at an earlier time. Then, once $\boldsymbol{\lambda}^\star(t_0)$ is known, $\mathbf{x}^\star(t)$ and $\boldsymbol{\lambda}^\star(t)$ are found by forward integration of (3.61) from $\mathbf{x}^\star(t_0)$ and $\boldsymbol{\lambda}^\star(t_0)$, respectively, which finally gives $\mathbf{u}^\star(t)$, $t_0 \le t \le t_\mathrm{f}$, from (3.60).

Even more interesting, one can also use the entire trajectory $\mathsf{S}(t)$, $t_0 \le t \le t_\mathrm{f}$, to determine the continuous feedback law for optimal control as

$$\mathbf{u}^\star(t) = -\left[\mathsf{Q}(t)^{-1}\mathsf{B}(t)^\mathsf{T}\mathsf{S}(t)\right]\mathbf{x}^\star(t). \tag{3.63}$$

Remark 3.24. The foregoing approach can be readily extended to LQR problems having mixed state/control terms in the integral cost:

$$\mathcal{J}(\mathbf{u}) := \int_{t_0}^{t_\mathrm{f}} \tfrac{1}{2}\begin{bmatrix}\mathbf{u}(t)\\ \mathbf{x}(t)\end{bmatrix}^\mathsf{T}\begin{bmatrix}\mathsf{Q} & \mathsf{P}\\ \mathsf{P}^\mathsf{T} & \mathsf{R}\end{bmatrix}\begin{bmatrix}\mathbf{u}(t)\\ \mathbf{x}(t)\end{bmatrix}\,\mathrm{d}t + \tfrac{1}{2}\mathbf{x}(t_\mathrm{f})^\mathsf{T}\mathsf{S}_\mathrm{f}\mathbf{x}(t_\mathrm{f}).$$

The matrix Riccati equation (3.62) becomes

$$\dot{\mathsf{S}} = -\mathsf{S}\left(\mathsf{A}-\mathsf{B}\mathsf{Q}^{-1}\mathsf{P}^{\mathsf{T}}\right)-\left(\mathsf{A}-\mathsf{B}\mathsf{Q}^{-1}\mathsf{P}^{\mathsf{T}}\right)^{\mathsf{T}}\mathsf{S}+\mathsf{S}\mathsf{B}\mathsf{Q}^{-1}\mathsf{B}^{\mathsf{T}}\mathsf{S}+\mathsf{P}\mathsf{Q}^{-1}\mathsf{P}^{\mathsf{T}}-\mathsf{R},$$

the state/adjoint equations as

$$\begin{bmatrix} \dot{\mathbf{x}}^{\star}(t) \\ \dot{\boldsymbol{\lambda}}^{\star}(t) \end{bmatrix} = \begin{bmatrix} \mathsf{A}-\mathsf{B}\mathsf{Q}^{-1}\mathsf{P}^{\mathsf{T}} & -\mathsf{B}\mathsf{Q}^{-1}\mathsf{B}^{\mathsf{T}} \\ -\mathsf{R}+\mathsf{P}\mathsf{Q}^{-1}\mathsf{P}^{\mathsf{T}} & -\left(\mathsf{A}-\mathsf{B}\mathsf{Q}^{-1}\mathsf{P}^{\mathsf{T}}\right)^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{\star}(t) \\ \boldsymbol{\lambda}^{\star}(t) \end{bmatrix},$$

and the control is given by

$$\mathbf{u}^{\star}(t) = -\mathsf{Q}(t)^{-1}\left[\mathsf{P}(t)^{\mathsf{T}}\mathbf{x}^{\star}(t)+\mathsf{B}(t)^{\mathsf{T}}\boldsymbol{\lambda}^{\star}(t)\right] = -\mathsf{Q}(t)^{-1}\left[\mathsf{P}(t)^{\mathsf{T}}+\mathsf{B}(t)^{\mathsf{T}}\mathsf{S}(t)\right]\mathbf{x}^{\star}(t).$$

3.5 MAXIMUM PRINCIPLES

In §3.4, we described first-order conditions that every (continuous or piecewise continuous) optimal control must necessarily satisfy, provided that no path restriction is placed on the control or the state variables. In this section, we shall present more general necessary conditions of optimality for those optimal control problems having path constraints. Such conditions are known collectively as the *Pontryagin Maximum Principle (PMP)*. The announcement of the PMP in the late 1950's can properly be regarded as the birth of the mathematical theory of optimal control.

In §3.5.1, we shall describe and illustrate the PMP for autonomous problems (a complete proof is omitted herein because it is too technical). Two important extensions, one to non-autonomous problems, and the other to problems involving sets as target data (e.g., $\mathbf{x}(t_{\mathrm{f}}) \in S_{\mathrm{f}}$, where S_{f} is a specified set) shall be discussed in §3.5.2. An application of the PMP to linear time-optimal problems shall be presented in §3.5.3. Then, the case of *singular problems* shall be considered in §3.5.4. Finally, necessary conditions for problems with mixed and pure state path constraints shall be presented in §3.5.5 and §3.5.6, respectively.

3.5.1 Pontryagin Maximum Principle for Autonomous Systems

Throughout this subsection, we shall consider the problem to minimize the cost functional

$$\mathcal{J}(\mathbf{u}, t_{\mathrm{f}}) := \int_{t_0}^{t_{\mathrm{f}}} \ell(\mathbf{x}(t), \mathbf{u}(t))\, \mathrm{d}t,$$

with fixed initial time t_0 and *unspecified* final time t_{f}, subject to the *autonomous* dynamical system

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)); \quad \mathbf{x}(t_0) = \mathbf{x}_0,$$

and a fixed target state

$$\mathbf{x}(t_{\mathrm{f}}) = \mathbf{x}_{\mathrm{f}}.$$

The admissible controls shall be taken in the class of piecewise continuous functions

$$\mathbf{u} \in \mathcal{U}[t_0, T] := \{\mathbf{u} \in \hat{\mathcal{C}}[t_0, T] : \mathbf{u}(t) \in U \text{ for } t_0 \le t \le t_{\mathrm{f}}\},$$

with $t_{\mathrm{f}} \in [t_0, T]$, where T is "sufficiently" large, and the nonempty, possibly closed and nonconvex, set U denotes the control region.

Observe that since the problem is autonomous (neither $\mathbf{f}$ nor ℓ show explicit dependence on time), a translation along the t-axis does not change the properties of the controls. In other words, if the control $\mathbf{u}(t)$, $t_0 \leq t \leq t_\mathrm{f}$, transfers the phase point from $\mathbf{x}_0$ to $\mathbf{x}_\mathrm{f}$, and imparts the value $\mathcal{J}$ to the cost functional, then the control $\mathbf{u}(t+\theta)$, $t_0 - \theta \leq t \leq t_\mathrm{f} - \theta$, also transfers the phase point from $\mathbf{x}_0$ to $\mathbf{x}_\mathrm{f}$ and imparts the same value $\mathcal{J}$ to the cost functional, for any real number θ. This makes it possible to relocate the initial time t_0 from which the control is given anywhere on the time axis.

Before we can state the PMP, some notation and analysis is needed. For a given control $\mathbf{u}$ and its corresponding response $\mathbf{x}$, we define the *dynamic cost variable* c as

$$c(t) := \int_{t_0}^{t} \ell(\mathbf{x}(\tau), \mathbf{u}(\tau))\, \mathrm{d}\tau.$$

If $\mathbf{u}$ is feasible, $\mathbf{x}(t_\mathrm{f}) = \mathbf{x}_\mathrm{f}$ for some $t_\mathrm{f} \geq t_0$, and the associated cost is $\mathcal{J}(\mathbf{u}, t_\mathrm{f}) = c(t_\mathrm{f})$. Then, introducing the $(n_x + 1)$-vector $\tilde{\mathbf{x}}(t)^\mathsf{T} := (c(t), \mathbf{x}(t)^\mathsf{T})$ (extended response), and defining $\tilde{\mathbf{f}}(\mathbf{x}, \mathbf{u})^\mathsf{T} := (\ell(\mathbf{x}, \mathbf{u}), \mathbf{f}(\mathbf{x}, \mathbf{u})^\mathsf{T})$ (extended system), an equivalent formulation for the optimal control problem is as follows:

Problem 3.25 (Reformulated Optimal Control Problem). *Find an admissible control* $\mathbf{u}$ *and final time* t_f *such that the* $(n_x + 1)$*-dimensional solution of*

$$\dot{\tilde{\mathbf{x}}}(t) = \tilde{\mathbf{f}}(\mathbf{x}(t), \mathbf{u}(t)); \quad \tilde{\mathbf{x}}(t_0) = \begin{pmatrix} 0 \\ \mathbf{x}_0 \end{pmatrix},$$

terminates at

$$\begin{pmatrix} c(t_\mathrm{f}) \\ \mathbf{x}_\mathrm{f} \end{pmatrix} \quad (\mathbf{x}_\mathrm{f} \textit{ the given target state}),$$

with $c(t_\mathrm{f})$ *taking on the least possible value.*

A geometrical interpretation of Problem 3.25 is proposed in Fig. 3.7. below. If we let Π be the line passing through the point $(0, \mathbf{x}_\mathrm{f})$ and parallel to the c-axis (this line is made up of all the points $(\zeta, \mathbf{x}_\mathrm{f})$ where ζ is arbitrary), then the (extended) response corresponding to any feasible control $\mathbf{u}$ passes through a point on Π. Moreover, if $\mathbf{u}^\star$ is a (globally) optimal control, no extended response $\tilde{\mathbf{x}} := (\mathcal{J}(\mathbf{u}, t_\mathrm{f}), \mathbf{x}_\mathrm{f})$ can hit the line Π below $\tilde{\mathbf{x}}^\star := (\mathcal{J}(\mathbf{u}^\star, t_\mathrm{f}^\star), \mathbf{x}_\mathrm{f})$.

To establish the PMP, the basic idea is to perturb an optimal control, say $\mathbf{u}^\star$, by changing its value to any admissible vector $\mathbf{v}$ over any small time interval. In particular, the corresponding perturbations in the response belong to a cone $\mathcal{K}(t)$ in the $(n_x + 1)$-dimensional extended response space (namely, the *cone of attainability*). That is, if a pair $(\mathbf{u}^\star, \mathbf{x}^\star)$ is optimal, then $\mathcal{K}(t_\mathrm{f}^\star)$ does not contain any vertical downward vector, $\tilde{\mathbf{d}} = \mu(1, 0, \ldots, 0)^\mathsf{T}$, $\mu < 0$, at $\tilde{\mathbf{x}}^\star(t_\mathrm{f}^\star)$ (provided $t_\mathrm{f}^\star$ is regular). In other words, $\mathbb{R}^{n_x+1}$ can be separated into two half-spaces by means of a support hyperplane passing at the vertex $(0, \mathbf{x}_\mathrm{f})$ of $\mathcal{K}(t_\mathrm{f}^\star)$,

$$\tilde{\mathbf{d}}^\mathsf{T} \tilde{\mathbf{x}} \leq 0, \quad \forall \tilde{\mathbf{x}} \in \mathcal{K}(t_\mathrm{f}^\star).$$

it is this latter inequality that leads to the PMP:

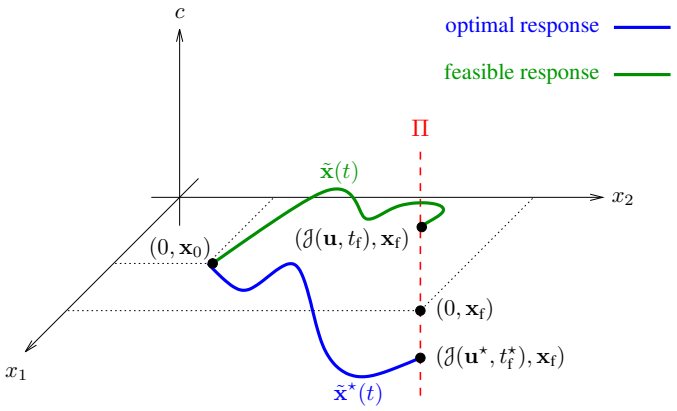


Figure 3.7. Geometric representation of Problem 3.25. (The c-axis is vertical for clarity.)

Theorem 3.26 (Pontryagin Maximum Principle for Autonomous Systems[4]). *Consider the optimal control problem*

$$\textit{minimize:}\quad \mathcal{J}(\mathbf{u}, t_\mathrm{f}) := \int_{t_0}^{t_\mathrm{f}} \ell(\mathbf{x}(t), \mathbf{u}(t))\, \mathrm{d}t \tag{3.64}$$

$$\textit{subject to:}\quad \dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)); \quad \mathbf{x}(t_0) = \mathbf{x}_0; \quad \mathbf{x}(t_\mathrm{f}) = \mathbf{x}_\mathrm{f} \tag{3.65}$$

$$\mathbf{u}(t) \in U, \tag{3.66}$$

with fixed initial time t_0 and free terminal time t_f. Let ℓ and $\mathbf{f}$ be continuous in $(\mathbf{x}, \mathbf{u})$ and have continuous first partial derivatives with respect to $\mathbf{x}$, for all $(\mathbf{x}, \mathbf{u}) \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_u}$. Suppose that $(\mathbf{u}^\star, t_\mathrm{f}^\star) \in \mathcal{C}[t_0, T]^{n_u} \times [t_0, T)$ is a minimizer for the problem, and let $\tilde{\mathbf{x}}^\star$ denote the optimal extended response. Then, there exists a $n_x + 1$-dimensional piecewise continuously differentiable vector function $\tilde{\boldsymbol{\lambda}}^\star = (\lambda_0^\star, \lambda_1^\star, \ldots, \lambda_{n_x}^\star) \neq (0, 0, \ldots, 0)$ such that

$$\dot{\tilde{\boldsymbol{\lambda}}}^\star(t) = -\,\mathcal{H}_{\tilde{\mathbf{x}}}(\mathbf{x}^\star(t), \mathbf{u}^\star(t), \tilde{\boldsymbol{\lambda}}^\star(t)), \tag{3.67}$$

with $\mathcal{H}(\mathbf{x}, \mathbf{u}, \tilde{\boldsymbol{\lambda}}) := \tilde{\boldsymbol{\lambda}}^\mathsf{T} \tilde{\mathbf{f}}(\mathbf{x}, \mathbf{u})$, and:

(i) the function $\mathcal{H}(\mathbf{x}^\star(t), \mathbf{v}, \tilde{\boldsymbol{\lambda}}^\star(t))$ attains its minimum on U at $\mathbf{v} = \mathbf{u}^\star(t)$:

$$\mathcal{H}(\mathbf{x}^\star(t), \mathbf{v}, \tilde{\boldsymbol{\lambda}}^\star(t)) \geq \mathcal{H}(\mathbf{x}^\star(t), \mathbf{u}^\star(t), \tilde{\boldsymbol{\lambda}}^\star(t)), \quad \forall \mathbf{v} \in U, \tag{3.68}$$

for every $t_0 \leq t \leq t_\mathrm{f}^\star$;

(ii) the following relations

$$\lambda_0^\star(t) = \text{const.} \geq 0 \tag{3.69}$$

$$\mathcal{H}(\mathbf{x}^\star(t), \mathbf{u}^\star(t), \tilde{\boldsymbol{\lambda}}^\star(t)) = \text{const.}, \tag{3.70}$$

[4] In the original Maximum Principle formulation [41], the condition (3.68) is in fact a maximum condition, and the sign requirement for the costate variable $\lambda_0^\star$ in (3.69) is reversed. We decided to present the PMP with (3.68) and (3.69) in order that that the resulting necessary conditions be consistent with those derived earlier in §3.4 based on the variational approach. Therefore, the PMP corresponds to a "minimum principle" herein.

are satisfied at every $t \in [t_0, t_\mathrm{f}^\star]$. *In particular, if the final time is unspecified, the following transversal condition holds:*

$$\mathcal{H}(\mathbf{x}^\star(t_\mathrm{f}^\star), \mathbf{u}^\star(t_\mathrm{f}^\star), \tilde{\boldsymbol{\lambda}}^\star(t_\mathrm{f}^\star)) = 0. \tag{3.71}$$

Proof. A complete proof of the PMP can be found, e.g., in [41, Chapter 2] or in [38, Chapter 5]. □

The PMP allows to single out, from among all controls whose response starts at $\mathbf{x}_0$ and ends at some point of Π, those satisfying all of the formulated conditions. Observe that we have a complete set of $2 \times n_x + n_u + 3$ conditions for the $n_u + 2 \times n_x + 3$ variables $(\mathbf{u}, \tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}}, t_\mathrm{f})$. In particular, the extended state $\tilde{\mathbf{x}}$ and adjoint $\tilde{\boldsymbol{\lambda}}$ trajectories are determined by $2 \times n_x + 2$ ODEs, with the corresponding $n_x + 1$ initial conditions $\tilde{\mathbf{x}}(t_0)^\mathsf{T} = (0, \mathbf{x}_0{}^\mathsf{T})$ and the n_x terminal conditions $\mathbf{x}(t_\mathrm{f}) = \mathbf{x}_\mathrm{f}$, plus the adjoint terminal condition $\lambda_0(t_\mathrm{f}) \geq 0$. We thus have either one of two possibilities:

(i) If $\lambda_0(t) > 0$, $t_0 \leq t \leq t_\mathrm{f}$, then the functions λ_i, $0 \leq i \leq n_x$, are defined up to a common multiple (since the function $\mathcal{H}$ is homogeneous with respect to $\boldsymbol{\lambda}$). This case is known as the *normal case*, and it is common practice to normalize the adjoint variables by taking $\lambda_0(t) = 1$, $t_0 \leq t \leq t_\mathrm{f}$.

(ii) If $\lambda_0(t) = 0$, $t_0 \leq t \leq t_\mathrm{f}$, the adjoint variables are determined uniquely. This case is known as the *abnormal case*, however, since the necessary conditions of optimality become independent of the cost functional.

Besides differential equations, the minimum condition (3.68) determines the control variables $\mathbf{u}$, and the transversal condition (3.70) determines t_f.

Notice that the PMP as stated in Theorem 3.26 applies to a *minimization* problem. If instead, one wishes to maximize the cost functional (3.64), the sign of the inequality (3.69) should be reversed,

$$\lambda_0^\star(t_\mathrm{f}^\star) \leq 0.$$

(But the minimum condition (3.68) should *not* be made a maximum condition for a maximization problem!)

Remark 3.27 (Link to the First-Order Necessary Conditions of §3.4). It may appear on first thought that the requirement in (3.68) could have been more succinctly embodied in the first-order conditions

$$\mathcal{H}_\mathbf{u}(\mathbf{u}^\star(t), \mathbf{x}^\star(t), \boldsymbol{\lambda}^\star(t)) = \mathbf{0},$$

properly supported by the second-order condition

$$\mathcal{H}_{\mathbf{uu}}(\mathbf{u}^\star(t), \mathbf{x}^\star(t), \boldsymbol{\lambda}^\star(t)) \succeq \mathbf{0},$$

for each $t_0 \leq t \leq t_\mathrm{f}^\star$. It turns out, however, that the requirement (3.68) is a much broader statement. First, it allows handling restrictions in the control variables, which was not the case for the first- and second-order conditions obtained with the variational approach in §3.4. In particular, the condition $\mathcal{H}_\mathbf{u} = \mathbf{0}$ does not even apply when the minimum of $\mathcal{H}$ occurs on the boundary of the control region U. Moreover, like the Weierstrass condition in the classical calculus of variations (see Theorem 2.36, p. 87), the condition (3.68) allows to

detect *strong* minima, and not merely *weak* minima.[5] Overall, the PMP can thus be thought of as the generalization to optimal control problems of the Euler equation, the Legendre condition, and the Weierstrass condition in the classical calculus of variations, taken all together. Observe also that the PMP is less restrictive that the variational approach since ℓ and $\mathbf{f}$ are not required to be continuously differentiable with respect to $\mathbf{u}$ (only continuous differentiability with respect to $\mathbf{x}$ is needed).

Example 3.28. Consider the same optimal control problem as in Example 3.22,

$$\text{minimize:} \quad \mathcal{J}(u) \;:=\; \int_0^1 \tfrac{1}{2}u(t)^2 \,\mathrm{d}t \tag{3.72}$$

$$\text{subject to:} \quad \dot{x}(t) \;=\; u(t) - x(t); \quad x(0) \;=\; 1; \quad x(1) \;=\; 0, \tag{3.73}$$

where the control u is now constrained by the condition that $-0.6 \leq u(t) \leq 0$, for $t \in [0,1]$.

The Hamiltonian function for the problem reads

$$\mathcal{H}(x, u, \lambda_0, \lambda) \;=\; \tfrac{1}{2}\lambda_0 u^2 + \lambda(u - x).$$

The optimal adjoint variables $\lambda_0^\star$ and $\lambda^\star$ must therefore satisfy the differential equations

$$\begin{aligned} \dot{\lambda}_0(t) &= -\mathcal{H}_c = 0 \\ \dot{\lambda}(t) &= -\mathcal{H}_x = \lambda(t), \end{aligned}$$

from which we get

$$\begin{aligned} \lambda_0^\star(t) &= K_0 \\ \lambda^\star(t) &= K\mathrm{e}^t. \end{aligned}$$

(We shall set $K_0 = 1$ subsequently, since the problem (3.72,3.73) is not abnormal.)

The PMP imposes that every optimal control $u^\star$ must be such that

$$u^\star(t) \in \arg\min_v \{\mathcal{H}(x^\star(t), v, \lambda_0^\star(t), \lambda^\star(t)) : -0.6 \leq v \leq 0\},$$

for each $t \in [0,1]$, from which we get

$$u^\star(t) = \begin{cases} 0 & \text{if } \lambda^\star(t) \leq 0 \\ -0.6 & \text{if } \lambda^\star(t) \geq 0.6 \\ -\lambda^\star(t) = -K\mathrm{e}^t & \text{otherwise} \end{cases}$$

Note that $K \leq 0$ implies $\lambda^\star(t) \leq 0$ at each time, and $u^\star(t) = 0$, $0 \leq t \leq 1$. However, this control yields an infeasible response ($x^\star(1) \neq 0$), and hence $K > 0$. That is, *every optimal control is a piecewise continuous function which takes the values* $-K\mathrm{e}^t$ *or* -0.6*, and has at most* 1 *corner point* (since $\lambda^\star$ is strictly decreasing in $[0,1]$):

$$u^\star(t) = u^\star_{(1)}(t) = -K\mathrm{e}^t, \quad 0 \leq t \leq t_s^\star \qquad u^\star(t) = u^\star_{(2)}(t) = -0.6, \quad t_s^\star < t \leq 1,$$

[5] By analogy to the classical calculus of variations (see §2.3, p. 66), *weak* minima correspond to "sufficiently" small perturbations in $\mathbf{u}$ that assure negligible higher-order terms both in $\|\boldsymbol{\delta}\mathbf{x}\|^2$ and $\|\boldsymbol{\delta}\mathbf{u}\|^2$; on the other hand, *strong* minima consider more general variations in $\mathbf{u}$ that assure negligible higher-order terms in $\|\boldsymbol{\delta}\mathbf{x}\|^2$ only.

where $t_s^\star$ denotes the (optimal) switching time. In particular, there must be a corner point since the control $u^\star(t) = -0.6$, $0 \leq t \leq 1$, yields an infeasible response.

- For the time interval $0 \leq t \leq t_s^\star$, we have

$$x_{(1)}^\star(t) = C_1 \mathrm{e}^{-t}\left(1 - \frac{K}{2C_1}\mathrm{e}^{2t}\right)$$
$$\mathcal{H}(u_{(1)}^\star(t), x_{(1)}^\star(t), \tilde{\boldsymbol{\lambda}}^\star(t)) = -KC_1,$$

  where C_1 is a constant of integration.

- For the time interval $t_s^\star < t \leq 1$, on the other hand, we have

$$x_{(2)}^\star(t) = C_2 \mathrm{e}^{-t} - 0.6$$
$$\mathcal{H}(u_{(2)}^\star(t), x_{(2)}^\star(t), \tilde{\boldsymbol{\lambda}}^\star(t)) = -KC_2 + \frac{(-0.6)^2}{2},$$

  where C_2 is another constant of integration.

Moreover, since the arc $x_{(1)}^\star$ starts at $t = 0$ with $x_{(1)}^\star(0) = 1$, and the arc $x_{(2)}^\star$ ends at $t = 1$ with $x_{(2)}^\star(1) = 0$, the constants of integration C_1 and C_2 are given by

$$C_1 = 1 + \frac{K}{2} \quad \text{and} \quad C_2 = 0.6\mathrm{e}.$$

The Hamiltonian function $\mathcal{H}$ being constant along an optimal solution, we have

$$\mathcal{H}(u_{(1)}^\star(t_s^\star), x_{(1)}^\star(t_s^\star), \tilde{\boldsymbol{\lambda}}^\star(t_s^\star)) = \mathcal{H}(u_{(2)}^\star(t_s^\star), x_{(2)}^\star(t_s^\star), \tilde{\boldsymbol{\lambda}}^\star(t_s^\star)),$$

from which we get

$$K = -(1 - 0.6\mathrm{e}) - \sqrt{(1 - 0.6\mathrm{e})^2 - (-0.6)^2} \approx 0.436.$$

(The other possible value of $K = -(1-0.6\mathrm{e}) + \sqrt{(1 - 0.6\mathrm{e})^2 - (-0.6)^2} \approx 0.826$ giving a switching time $t_s^\star$ not in the range $[0, 1]$.) Finally, the switching time $t_s^\star$ is deduced from the state continuity condition, $x_{(1)}^\star(t_s^\star) = x_{(2)}^\star(t_s^\star)$. Numerically, we get $t_s^\star \approx 0.320$.

The optimal trajectories $u^\star(t)$ and $x^\star(t)$, $0 \leq t \leq 1$, are shown in Fig. 3.8. below. Notice, in particular, that the optimal control is continuous. That is, the condition $\mathcal{H}(u_{(1)}^\star(t_s^\star), x_{(1)}^\star(t_s^\star), \tilde{\boldsymbol{\lambda}}^\star(t_s^\star)) = \mathcal{H}(u_{(2)}^\star(t_s^\star), x_{(2)}^\star(t_s^\star), \tilde{\boldsymbol{\lambda}}^\star(t_s^\star))$ determining K imposes that the trajectories $x_{(1)}$ and $x_{(2)}$ be tangent at $t_s^\star$. These optimal trajectories should also be compared to those obtained in Example 3.22 without restriction placed on the control variables.

3.5.2 Extensions of the Pontryagin Maximum Principle

In this subsection, we shall treat two extensions of the PMP. The first extension is for the case where the terminal condition $\mathbf{x}(t_\mathrm{f}) = \mathbf{x}_\mathrm{f}$ is replaced by the target set condition $\mathbf{x}(t_\mathrm{f}) \in X_\mathrm{f} \subset \mathbb{R}^{n_x}$. The second extension is to non-autonomous problems, and makes use of the former.

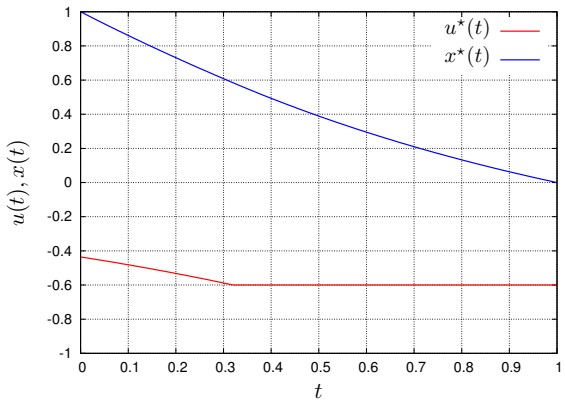


Figure 3.8. Optimal trajectories $u^\star(t)$ and $x^\star(t)$, $0 \le t \le 1$, in Example 3.28.

Regarding target set terminal conditions, we have the following theorem:

Theorem 3.29 (Transversal Conditions). *Consider the optimal control problem*

$$\textit{minimize:} \quad \mathcal{J}(\mathbf{u}, t_\mathrm{f}) := \int_{t_0}^{t_\mathrm{f}} \ell(\mathbf{x}(t), \mathbf{u}(t))\, \mathrm{d}t \tag{3.74}$$

$$\textit{subject to:} \quad \dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)); \quad \mathbf{x}(t_0) = \mathbf{x}_0; \quad \mathbf{x}(t_\mathrm{f}) \in X_\mathrm{f} \tag{3.75}$$

$$\mathbf{u}(t) \in U, \tag{3.76}$$

with fixed initial time t_0 and free terminal time t_f, and with X_f a smooth manifold of dimension $n_\mathrm{f} \le n_x$. Let ℓ and $\mathbf{f}$ be continuous in $(\mathbf{x}, \mathbf{u})$ and have continuous first partial derivatives with respect to $\mathbf{x}$, for all $(\mathbf{x}, \mathbf{u}) \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_u}$. Suppose that $(\mathbf{u}^\star, t_\mathrm{f}^\star) \in \mathcal{C}[t_0, T]^{n_u} \times [t_0, T)$ is a minimizer for the problem, and let $\tilde{\mathbf{x}}^\star$ denote the optimal extended response. Then, there exists a piecewise continuously differentiable vector function $\tilde{\boldsymbol{\lambda}}^\star = (\lambda_0^\star, \lambda_1^\star, \ldots, \lambda_{n_x}^\star) \neq (0, 0, \ldots, 0)$ solving (3.67) *and satisfying conditions (3.68–3.71) of Theorem 3.26. Moreover, $\boldsymbol{\lambda}^\star(t_\mathrm{f}^\star) := (\lambda_1^\star(t_\mathrm{f}^\star), \ldots, \lambda_{n_x}^\star(t_\mathrm{f}^\star))$ is orthogonal to the tangent plane, $\mathscr{T}(\mathbf{x}^\star(t_\mathrm{f}^\star))$, to X_f at $\mathbf{x}^\star(t_\mathrm{f}^\star)$:*

$$\boldsymbol{\lambda}^\star(t_\mathrm{f}^\star)^\mathsf{T} \mathbf{d} = 0, \quad \forall \mathbf{d} \in \mathscr{T}(\mathbf{x}^\star(t_\mathrm{f}^\star)). \tag{3.77}$$

Proof. A complete proof of the transversal conditions (3.77) can be found, e.g., in [41, Chapter 2]. □

Notice that when the set X_f degenerates into a point, the transversality condition at $t_\mathrm{f}^\star$ can be replaced by the condition that the optimal response $\mathbf{x}^\star$ pass through this point, as in Theorem 3.26.

In many practical problems, the target set X_f is specified as the intersection of $n_\psi = n_x - n_\mathrm{f}$ hypersurfaces defined by the equations

$$\psi_k(\mathbf{x}) = 0, \quad k = 1, \ldots, n_\psi.$$

Provided that the functions $\psi_1, \ldots, \psi_{n_\psi}$ are linearly independent at $\mathbf{x}^\star(t_\mathrm{f}^\star)$, i.e., the following *constraint qualification* holds:

$$\operatorname{rank}[\boldsymbol{\psi}_\mathbf{x}(\mathbf{x}^\star(t_\mathrm{f}^\star))] = n_\psi, \tag{3.78}$$

the tangent set $\mathscr{T}(\mathbf{x}^\star(t_\mathrm{f}^\star))$ is such that

$$\mathscr{T}(\mathbf{x}^\star(t_\mathrm{f}^\star)) = \{\mathbf{d} \in \mathbb{R}^{n_x} : \boldsymbol{\psi}_\mathbf{x}(\mathbf{x}^\star(t_\mathrm{f}^\star))\, \mathbf{d} = \mathbf{0}\},$$

(see, e.g., Lemma 1.48, p. 23). For the transversal condition (3.77) to be satisfied, it is thus necessary that $\boldsymbol{\lambda}^\star(t_\mathrm{f}^\star)$ be in the subspace spanned by the row vectors of $\boldsymbol{\psi}_\mathbf{x}(\mathbf{x}^\star(t_\mathrm{f}^\star))$; in other words, there must exist a vector $\boldsymbol{\nu}$ of Lagrange multipliers such that

$$\boldsymbol{\lambda}^\star(t_\mathrm{f}^\star) = \boldsymbol{\nu}^\mathsf{T} \boldsymbol{\psi}_\mathbf{x}(\mathbf{x}^\star(t_\mathrm{f}^\star)),$$

hence yielding the same terminal adjoint condition as in Theorem 3.18.

We now turn to the extension of the PMP for non-autonomous problems. We shall consider the optimal control problem in the same form as in (3.64–3.66), but for the case in which ℓ and $\mathbf{f}$ depend *explicitly* on time (the control region U is assumed independent of time). Thus, the system equations and the cost functional take the form

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) \\ \mathcal{J}(\mathbf{u}, t_\mathrm{f}) &:= \int_{t_0}^{t_\mathrm{f}} \ell(t, \mathbf{x}(t), \mathbf{u}(t))\, \mathrm{d}t. \end{aligned} \tag{3.79}$$

In order to solve this problem, we shall introduce yet another auxiliary variable, x_{n_x+1}, defined by

$$\dot{x}_{n_x+1}(t) = 1; \quad x_{n_x+1}(t_0) = t_0.$$

It is obvious that $x_{n_x+1}(t) = t$, $t_0 \le t \le t_\mathrm{f}$. Therefore, we get the (n_x+1)-dimensional system

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{f}(x_{n_x+1}(t), \mathbf{x}(t), \mathbf{u}(t)) \\ \dot{x}_{n_x+1}(t) &= 1. \end{aligned}$$

Next, we apply the autonomous version of the PMP with transversal conditions (Theorem 3.29) to find necessary conditions of optimality, in terms of the (n_x+2)-dimensional vector $(c, \mathbf{x}^\mathsf{T}, x_{n_x+1})$, where

$$\dot{c} = \ell(x_{n_x+1}(t), \mathbf{x}(t), \mathbf{u}(t)); \quad c(t_0) = 0.$$

Using the same notations as in §3.5.1 for the extended response, $\tilde{\mathbf{x}}^\mathsf{T} := (c, \mathbf{x}^\mathsf{T})$, and the extended system, $\tilde{\mathbf{f}}^\mathsf{T} := (\ell, \mathbf{f}^\mathsf{T})$, the equations giving the (n_x+2) adjoint variables $(\tilde{\boldsymbol{\lambda}}^\mathsf{T}, \lambda_{n_x+1}) := (\lambda_0, \lambda_1, \ldots, \lambda_{n_x}, \lambda_{n_x+1})$ read

$$\begin{aligned} \dot{\lambda}_0(t) &= 0 \\ \dot{\lambda}_i(t) &= -\tilde{\boldsymbol{\lambda}}(t)^\mathsf{T} \tilde{\mathbf{f}}_{x_i}(x_{n_x+1}, \mathbf{x}, \mathbf{u}), \quad i = 1, \ldots, n_x \\ \dot{\lambda}_{n_x+1}(t) &= -\tilde{\boldsymbol{\lambda}}(t)^\mathsf{T} \tilde{\mathbf{f}}_t(x_{n_x+1}, \mathbf{x}, \mathbf{u}). \end{aligned}$$

Moreover, the transversal condition at t_f requires that X_f (which is parallel to the x_{n_x+1} axis) be orthogonal to the vector $(\lambda_1, \ldots, \lambda_{n_x}, \lambda_{n_x+1})$. But since X_f is parallel to the x_{n_x+1} axis, it follows that

$$\lambda_{n_x+1}(t_\mathrm{f}) = 0.$$

Overall, a version of the PMP for non-autonomous system is as follows:

Theorem 3.30 (Pontryagin Maximum Principle for Non-Autonomous Systems). *Consider the optimal control problem*

$$\textit{minimize:}\quad \mathcal{J}(\mathbf{u}, t_\mathrm{f}) := \int_{t_0}^{t_\mathrm{f}} \ell(t, \mathbf{x}(t), \mathbf{u}(t))\, \mathrm{d}t \tag{3.80}$$

$$\textit{subject to:}\quad \dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)); \quad \mathbf{x}(t_0) = \mathbf{x}_0; \quad \mathbf{x}(t_\mathrm{f}) = \mathbf{x}_\mathrm{f} \tag{3.81}$$

$$\mathbf{u}(t) \in U, \tag{3.82}$$

with fixed initial time t_0 and free terminal time t_f. Let ℓ and $\mathbf{f}$ be continuous in $(t, \mathbf{x}, \mathbf{u})$ and have continuous first partial derivatives with respect to $(t, \mathbf{x})$, for all $(t, \mathbf{x}, \mathbf{u}) \in [t_0, T] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u}$. Suppose that $(\mathbf{u}^\star, t_\mathrm{f}^\star) \in \mathcal{C}[t_0, T]^{n_u} \times [t_0, T)$ is a minimizer for the problem, and let $\tilde{\mathbf{x}}^\star$ denote the optimal extended response. Then, there exists a $(n_x + 1)$-dimensional, piecewise continuously differentiable vector function $\tilde{\boldsymbol{\lambda}}^\star = (\lambda_0^\star, \lambda_1^\star, \ldots, \lambda_{n_x}^\star) \neq (0, 0, \ldots, 0)$ such that

$$\dot{\tilde{\boldsymbol{\lambda}}}^\star(t) = -\,\mathcal{H}_{\tilde{\mathbf{x}}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t), \tilde{\boldsymbol{\lambda}}^\star(t)), \tag{3.83}$$

with $\mathcal{H}(t, \mathbf{x}, \mathbf{u}, \tilde{\boldsymbol{\lambda}}) := \tilde{\boldsymbol{\lambda}}^\mathsf{T} \tilde{\mathbf{f}}(t, \mathbf{x}, \mathbf{u})$, and:

(i) the function $\mathcal{H}(\mathbf{x}^\star(t), \mathbf{v}, \tilde{\boldsymbol{\lambda}}^\star(t))$ attains its minimum on U at $\mathbf{v} = \mathbf{u}^\star(t)$:

$$\mathcal{H}(t, \mathbf{x}^\star(t), \mathbf{v}, \tilde{\boldsymbol{\lambda}}^\star(t)) \geq \mathcal{H}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t), \tilde{\boldsymbol{\lambda}}^\star(t)), \quad \forall \mathbf{v} \in U, \tag{3.84}$$

for every $t_0 \leq t \leq t_\mathrm{f}^\star$;

(ii) the following relations

$$\lambda_0^\star(t) = \text{const.} \geq 0 \tag{3.85}$$

$$\dot{\mathcal{H}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t), \tilde{\boldsymbol{\lambda}}^\star(t)) = \tilde{\boldsymbol{\lambda}}^\star(t)^\mathsf{T} \tilde{\mathbf{f}}_t(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t)), \tag{3.86}$$

are satisfied at any $t \in [t_0, t_\mathrm{f}^\star]$. Moreover, in the case wherein the final time is unspecified, $t_\mathrm{f}^\star$ is determined from the transversal condition

$$\mathcal{H}(t_\mathrm{f}^\star, \mathbf{x}^\star(t_\mathrm{f}^\star), \mathbf{u}^\star(t_\mathrm{f}^\star), \tilde{\boldsymbol{\lambda}}^\star(t_\mathrm{f}^\star)) = 0. \tag{3.87}$$

3.5.3 Application: Linear Time-Optimal Problems

An interesting application of the PMP is in the special case of a linear time-invariant system and a linear cost functional,

$$\text{minimize:}\quad \mathcal{J}(\mathbf{u}, t_\mathrm{f}) := \int_{t_0}^{t_\mathrm{f}} [\mathbf{a}^\mathsf{T}\mathbf{u}(t) + \mathbf{b}^\mathsf{T}\mathbf{x}(t) + c]\, \mathrm{d}t \tag{3.88}$$

$$\text{subject to:}\quad \dot{\mathbf{x}}(t) = \mathsf{F}\mathbf{x}(t) + \mathsf{G}\mathbf{u}(t); \quad \mathbf{x}(t_0) = \mathbf{x}_0; \quad \mathbf{x}(t_\mathrm{f}) = \mathbf{x}_\mathrm{f} \tag{3.89}$$

$$\mathbf{u}(t) \in U. \tag{3.90}$$

with fixed initial time t_0 and free terminal time t_f. In the case where the control region U is unbounded, no minimum exists, in general, for such problems; this is because the control

may take on infinite values, which correspond to instantaneous jumps of the state variables in the phase space. On the other hand, when U is bounded, e.g., $U := \left[\mathbf{u}^L, \mathbf{u}^U\right]$, it is reasonable to expect that the control will lie on the *boundary* of U, and that it will jump from one boundary of U to another during the time of operation of the system. The name *bang-bang control* has been coined to describe such situations wherein the controls move suddenly from one boundary point of the control region to another boundary point.

In this subsection, we shall only consider the so-called *linear time-optimal problem*, and limit our discussion to the case of a scalar control. More precisely, we consider the problem of finding a piecewise continuous control $u \in \hat{\mathcal{C}}[t_0, T]$ that brings the system from an initial state $\mathbf{x}_0 \neq \mathbf{0}$ to the origin, in minimum time:

$$\text{minimize:} \quad \mathcal{J}(\mathbf{u}, t_\mathrm{f}) := \int_{t_0}^{t_\mathrm{f}} \mathrm{d}t = t_\mathrm{f} - t_0 \tag{3.91}$$

$$\text{subject to:} \quad \dot{\mathbf{x}}(t) = \mathsf{F}\mathbf{x}(t) + \mathbf{g}u(t); \quad \mathbf{x}(t_0) = \mathbf{x}_0; \quad \mathbf{x}(t_\mathrm{f}) = \mathbf{0} \tag{3.92}$$

$$u(t) \in [u^L, u^U]. \tag{3.93}$$

The Hamiltonian function for this problem reads

$$\mathcal{H}(\mathbf{x}, u, \tilde{\boldsymbol{\lambda}}) = \lambda_0 + \boldsymbol{\lambda}^\mathsf{T}[\mathsf{F}\mathbf{x} + \mathbf{g}u].$$

It shall be assumed throughout that the problem is normal, and we take $\lambda_0(t) = 1$. Then, upon application of the PMP (Theorem 3.26), a necessary condition for $u^\star$ to be an optimal control is[6]

$$u^\star(t) = \begin{cases} u^U & \text{if } \boldsymbol{\lambda}^\star(t)^\mathsf{T}\mathbf{g} < 0 \\ u^L & \text{if } \boldsymbol{\lambda}^\star(t)^\mathsf{T}\mathbf{g} > 0, \end{cases} \tag{3.94}$$

for each $t_0 \leq t \leq t_\mathrm{f}^\star$, where $(x^\star(t), \lambda^\star(t))$ satisfy

$$\begin{aligned} \dot{\mathbf{x}}^\star(t) &= \mathsf{F}\,\mathbf{x}^\star(t) + \mathbf{g}\,u^\star(t) \\ \dot{\boldsymbol{\lambda}}^\star(t) &= -\mathsf{F}^\mathsf{T}\boldsymbol{\lambda}^\star(t), \end{aligned} \tag{3.95}$$

with boundary conditions $\mathbf{x}^\star(t_0) = \mathbf{x}_0$ and $\mathbf{x}^\star(t_\mathrm{f}^\star) = \mathbf{0}$; moreover, $t_\mathrm{f}^\star$ is obtained from the transversal condition (3.71), which with $\mathbf{x}^\star(t_\mathrm{f}^\star) = \mathbf{0}$ gives:

$$\boldsymbol{\lambda}^\star(t_\mathrm{f}^\star)^\mathsf{T}\mathbf{g}u^\star(t_\mathrm{f}^\star) = -1. \tag{3.96}$$

The quantity $\boldsymbol{\lambda}^\star(t)^\mathsf{T}\mathbf{g}$ is, for obvious reasons, called the *switching function*. If $\boldsymbol{\lambda}^\star(t)^\mathsf{T}\mathbf{g} = 0$ cannot be sustained over a finite interval of time, then the optimal control is of *bang-bang* type; in other words, $u^\star(t)$ is at u^L when the switching function is positive, and at u^U when the switching function is negative.

Example 3.31 (Bang-Bang Example). Consider the linear time-optimal problem (3.91–3.93) with

$$\dot{x}_1(t) = x_2(t), \qquad \dot{x}_2(t) = u(t), \qquad -1 \leq u(t) \leq 1.$$

[6]Note that we also have the possibility that $\boldsymbol{\lambda}(t)^\mathsf{T}\mathbf{g} = 0$ on some nonempty interval of time, which corresponds to a singular control arc; singular problems shall be discussed later in §3.5.4.

For this simple system, an optimal control $u^\star$ must satisfy

$$u^\star(t) = \begin{cases} 1 & \text{if } \lambda_2^\star(t) < 0 \\ -1 & \text{if } \lambda_2^\star(t) > 0. \end{cases}$$

The adjoint variables $\boldsymbol{\lambda}^\star$ verify the differential equations (3.95),

$$\begin{aligned} \dot{\lambda}_1^\star(t) &= 0 \\ \dot{\lambda}_2^\star(t) &= -\lambda_1^\star(t), \end{aligned}$$

which are readily solved as

$$\begin{aligned} {\lambda_1}^\star(t) &= A_1 \\ {\lambda_2}^\star(t) &= -A_1 t + A_2, \end{aligned}$$

where A_1 and A_2 are constants of integration. That is, the switching function $\boldsymbol{\lambda}^\star(t)^\mathsf{T}\mathbf{g} = -A_1 t + A_2$ is a linear function of time, and it follows that *every optimal control $u^\star(t)$, $t_0 \le t \le t_\mathrm{f}^\star$, is a piecewise constant function which takes on the values ± 1, and has at most two intervals on which it is constant.*

- For the time interval on which $u^\star(t) = 1$, we have

$$x_2^\star(t) = t + K_1, \qquad x_1^\star(t) = \frac{t^2}{2} + K_2 t + K_1 = \frac{1}{2}(t + K_2)^2 + \left(K_1 - \frac{K_2^2}{2}\right),$$

  (where K_1 and K_2 are constants of integration), from which we get

$$x_1^\star(t) = \frac{1}{2}[x_2^\star(t)]^2 + K, \tag{3.97}$$

  with $K = K_1 - \frac{1}{2}K_2^2$. Thus, the portion of the optimal response for which $u(t) = 1$ is an arc of the parabola (3.97), along which the phase points move upwards (since $\dot{x}_2 = 1 > 0$).

- analogously, for the time interval on which $u^\star(t) = -1$, we have

$$x_2^\star(t) = -t + K_1', \qquad x_1^\star(t) = \frac{t^2}{2} + K_2' t + K_1' = -\frac{1}{2}(-t + K_2')^2 + \left(K_1' + \frac{{K_2'}^2}{2}\right),$$

  from which we obtain

$$x_1^\star(t) = -\frac{1}{2}[x_2^\star(t)]^2 + K'. \tag{3.98}$$

  thus, the portion of the optimal response for which $u(t) = -1$ is an arc of the parabola (3.98), along which the phase points move downwards (since $\dot{x}_2 = -1 < 0$).

Observe that if $u^\star$ is initially equal to 1, and to -1 afterwards, the response consists of two adjoining parabolic arcs, and the second arc lies on that parabola defined by (3.98) which passes through the origin:

$$x_1^\star(t) = -\frac{1}{2}[x_2^\star(t)]^2. \tag{3.99}$$

Likewise, if $u^\star = -1$ first, and $u^\star = 1$ afterwards, the second arc lies on that parabola defined by (3.97) which passes through the origin:

$$x_1^\star(t) = \frac{1}{2}[x_2^\star(t)]^2. \tag{3.100}$$

The *switching curve* is therefore made up of the parabolas (3.99) (for $x_2 > 0$) and (3.100) (for $x_2 < 0$). By inspection, it is apparent that (i) $u^\star = -1$ above the switching curve, and (ii) $u^\star = 1$ below the switching curve. Overall, the optimal feedback law for the problem may thus be written as:

$$u^\star(t) = \begin{cases} 1 & \text{if} \quad [x_2^\star]^2 \operatorname{sign} x_2 < -2x_1^\star \quad \text{or} \quad [x_2^\star]^2 \operatorname{sign} x_2 = -2x_1^\star, \quad x_1^\star > 0 \\ -1 & \text{if} \quad [x_2^\star]^2 \operatorname{sign} x_2 > -2x_1^\star \quad \text{or} \quad [x_2^\star]^2 \operatorname{sign} x_2 = -2x_1^\star, \quad x_1^\star < 0. \end{cases}$$

The switching curve is illustrated in Fig. 3.9. below, along with typical optimal responses obtained for different initial conditions.

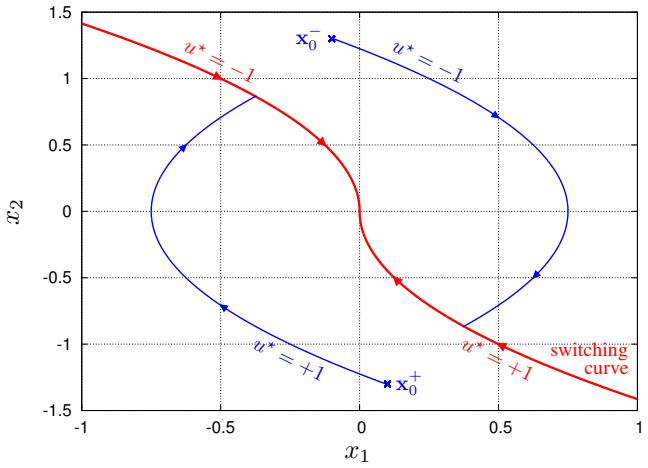


Figure 3.9. Switching curve and typical optimal responses for Example 3.31 — Red line: switching curve; blue line: typical path.

3.5.4 Singular Optimal Control Problems

In all the optimal control problems considered so far, the values $\mathbf{u}^\star(t)$ of candidate optimal controls could be explicitly determined by a minimum condition such as

$$\mathcal{H}(t, \mathbf{x}^\star(t), \mathbf{v}, \tilde{\boldsymbol{\lambda}}^\star(t)) \geq \mathcal{H}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t), \tilde{\boldsymbol{\lambda}}^\star(t)), \quad \forall \mathbf{v} \in U, \tag{3.101}$$

at each time instant $t \in [t_0, t_\mathrm{f}]$. However, it may happen for some problems that $\mathbf{u}^\star(t)$ cannot be directly determined by the foregoing condition.

To illustrate this situation, consider the following (scalar) optimal control problem:

$$\begin{aligned} \text{minimize:} \quad & \mathcal{J}(u) := \int_{t_0}^{t_\mathrm{f}} \ell^0(t,\mathbf{x}(t)) + u(t)\,\ell^1(t,\mathbf{x}(t))\,\mathrm{d}t \\ \text{subject to:} \quad & \dot{\mathbf{x}}(t) = \mathbf{f}^0(t,\mathbf{x}(t)) + u(t)\,\mathbf{f}^1(t,\mathbf{x}(t)); \quad \mathbf{x}(t_0) = \mathbf{x}_0; \quad \mathbf{x}(t_\mathrm{f}) = \mathbf{x}_\mathrm{f} \\ & u(t) \in U := [u^L, u^U], \end{aligned}$$

for $u \in \hat{\mathcal{C}}[t_0, t_\mathrm{f}]$, with fixed initial time t_0 and final time t_f. Since the Hamiltonian function is affine in the control u (i.e., contains u in at most the first power), we have

$$\mathcal{H}(t,\mathbf{x},u,\tilde{\boldsymbol{\lambda}}) = \mathcal{H}^0(t,\mathbf{x},\tilde{\boldsymbol{\lambda}}) + u(t)\,\mathcal{H}^1(t,\mathbf{x},\tilde{\boldsymbol{\lambda}}),$$

where

$$\mathcal{H}^0 := \lambda_0\ell^0 + \boldsymbol{\lambda}^\mathsf{T}\mathbf{f}^0, \quad \text{and} \quad \mathcal{H}^1 := \lambda_1\ell^1 + \boldsymbol{\lambda}^\mathsf{T}\mathbf{f}^1.$$

To minimize the Hamiltonian function at a given $t \in [t_0, t_\mathrm{f}]$, one has to take $u^\star(t) = u^U$ if $\mathcal{H}^1(t,\mathbf{x}^\star(t),\tilde{\boldsymbol{\lambda}}^\star(t)) > 0$, and $u^\star(t) = u^L$ if $\mathcal{H}^1(t,\mathbf{x}^\star(t),\tilde{\boldsymbol{\lambda}}^\star(t)) < 0$. We then have either one of two situations:

(i) If the term $\mathcal{H}^1(t,\mathbf{x}^\star(t),\tilde{\boldsymbol{\lambda}}^\star(t))$ vanishes only at isolated times, then the control u switches from u^L to u^U or vice versa each time $\mathcal{H}^1(t,\mathbf{x}^\star(t),\tilde{\boldsymbol{\lambda}}^\star(t))$ crosses zero; the control is said to be *bang-bang*. An illustration of this behavior was given earlier in Example 3.31.

(ii) On the other hand, if $\mathcal{H}^1(t,\mathbf{x}^\star(t),\tilde{\boldsymbol{\lambda}}^\star(t)) = 0$ can be sustained over some finite interval $(\theta_1,\theta_2) \subset [t_0, t_\mathrm{f}]$, then *any* value of $u \in [u^L, u^U]$ trivially meets the minimum condition (3.101). In other words, the control does not affect the Hamiltonian function on (θ_1,θ_2), and we have a *singular arc*.

For more general scalar optimal control problems of the form (3.80–3.82), singular arcs are obtained when the stationarity condition

$$\mathcal{H}_u(t,\mathbf{x}^\star(t),u,\tilde{\boldsymbol{\lambda}}^\star(t)) = 0,$$

is trivially satisfied by any admissible control on some nonempty subinterval $(\theta_1,\theta_2) \subset [t_0, t_\mathrm{f}]$, i.e., the matrix $\mathcal{H}_{uu}$ is singular. In the case of a vector control problem, the subsequent developments apply readily to each component $u_k(t)$, $k = 1,\ldots,n_u$, of $\mathbf{u}(t)$.

The following idea is used to determine the value of an optimal control along a singular arc. Since $\mathcal{H}_u = 0$ for all $t \in (\theta_1,\theta_2)$, its successive time derivatives $\frac{\mathrm{d}^q}{\mathrm{d}t^q}\mathcal{H}_u$, $q = 1,2,\ldots$, must also vanish on (θ_1,θ_2). In particular, we may find a smallest positive integer $\bar{q}$ such that

$$\begin{aligned} \frac{\mathrm{d}^{\bar{q}}}{\mathrm{d}t^{\bar{q}}}\mathcal{H}_u(t,\mathbf{x}^\star(t),\cdot,\tilde{\boldsymbol{\lambda}}^\star(t)) &= 0 \\ \frac{\partial}{\partial u}\left[\frac{\mathrm{d}^{\bar{q}}}{\mathrm{d}t^{\bar{q}}}\mathcal{H}_\mathbf{u}(t,\mathbf{x}^\star(t),\cdot,\tilde{\boldsymbol{\lambda}}^\star(t))\right] &\neq 0. \end{aligned}$$

Note that if such a smallest integer $\bar{q}$ exists, it must be *even* (see, e.g., [29] for a proof). Then, the nonnegative integer p such that $\bar{q} = 2p$ is called *the order of the singular arc*. Note also that singular arcs are not possible at all points of the $(\mathbf{x},\tilde{\boldsymbol{\lambda}})$-space. Along a singular

arc, the state and adjoint variables must lie on the so-called *singular surface* defined by the equations:

$$\begin{aligned}
\mathcal{H}_u(t, \mathbf{x}^\star(t), u^\star(t), \tilde{\boldsymbol{\lambda}}^\star(t)) &= 0 \\
\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}_u(t, \mathbf{x}^\star(t), u^\star(t), \tilde{\boldsymbol{\lambda}}^\star(t)) &= 0 \\
&\vdots \\
\frac{\mathrm{d}^{2p-1}}{\mathrm{d}t^{2p-1}}\mathcal{H}_u(t, \mathbf{x}^\star(t), u^\star(t), \tilde{\boldsymbol{\lambda}}^\star(t)) &= 0,
\end{aligned}$$

together with the additional equation $\mathcal{H}_u(t, \mathbf{x}^\star(t), u^\star(t), \tilde{\boldsymbol{\lambda}}^\star(t)) = 0$ if the final time is unspecified (see Theorem 3.30).

The analog to the Legendre-Clebsch condition (see Remark 3.21) along a singular arc reads

$$(-1)^p \frac{\partial}{\partial u}\left[\frac{\mathrm{d}^{2p}}{\mathrm{d}t^{2p}}\mathcal{H}_u(t, \mathbf{x}^\star(t), u^\star(t), \tilde{\boldsymbol{\lambda}}^\star(t))\right] \geq 0,$$

for each $t \in (\theta_1, \theta_2)$; this condition is often referred to as the *generalized Legendre-Clebsch condition*. (The inequality is reversed for a maximization problem.) Moreover, similar to non-singular optimal control problems (without inequality state constraint), both the adjoint variables $\tilde{\boldsymbol{\lambda}}$ and the Hamiltonian function $\mathcal{H}$ must be continuous, along an optimal trajectory, in a singular control problem (see §3.4.3).

In general, solutions to optimal control problems have a mixture of arcs, some singular and some nonsingular. In order to find the correct sequence of arcs, one has to postulate a particular sequence, and then check whether or not the necessary conditions of optimality are satisfied for that sequence.[7] Note, however, that finding the correct sequence of controls analytically may be very complicated and is even impossible for many problems.

In addition to the necessary conditions that the adjoint and the Hamiltonian must be continuous along an optimal trajectory, additional conditions must hold at the joining of a nonsingular arc to a singular arc, and vice versa. For first-order singular control problems, $p = 1$, the control variable u at the entry point θ_1 and the exit point θ_2 of a singular arc is either discontinuous (i.e., corner junctions are permitted), or continuously differentiable (see, e.g., [40] for a proof). In other words, an optimal control $\mathbf{u}^\star$ cannot be continuous at a junction time if its time derivative $\dot{\mathbf{u}}^\star$ is discontinuous. We present a first-order singular problem in Example 3.32 below.

Example 3.32 (First-Order Singular Optimal Control Problem). Consider the scalar optimal control problem:

$$\text{minimize:} \quad \mathcal{J}(u) := \int_0^2 \tfrac{1}{2}[x_1(t)]^2 \, \mathrm{d}t \tag{3.102}$$

$$\text{subject to:} \quad \dot{x}_1(t) = x_2(t) + u(t); \quad x_1(0) = 1; \quad x_1(2) = 0 \tag{3.103}$$

$$\dot{x}_2(t) = -u(t); \quad x_2(0) = 1; \quad x_2(2) = 0 \tag{3.104}$$

$$-10 \leq u(t) \leq 10, \quad 0 \leq t \leq 2, \tag{3.105}$$

[7] The situation is quite similar to NLP problems where one has to guess the set of active constraints, and then check whether the KKT necessary conditions are satisfied for that active set.

where the control u is taken in the set of piecewise continuous functions, $u \in \hat{\mathcal{C}}[0,2]$. This problem is linear in u, but nonlinear in x_1 through the cost functional.

The Hamiltonian function is given by

$$\mathcal{H}(\mathbf{x}, u, \tilde{\boldsymbol{\lambda}}) = \tfrac{1}{2}\lambda_0 x_1^2 + \lambda_1(x_2 + u) - \lambda_2 u = \tfrac{1}{2}\lambda_0 x_1^2 + \lambda_1 x_2 + (\lambda_1 - \lambda_2)u.$$

Assuming that $(u^\star, \mathbf{x}^\star, \boldsymbol{\lambda}^\star)$ is an optimal triple for the problem, and that the problem is normal (i.e., $\lambda_0(t) = 1, \forall t$), we have

$$u^\star(t) = \begin{cases} 10 & \text{if } \lambda_1^\star(t) < \lambda_2^\star(t) \\ -10 & \text{if } \lambda_1^\star(t) > \lambda_2^\star(t) \\ ? & \text{if } \lambda_1^\star(t) = \lambda_2^\star(t), \end{cases}$$

where

$$\begin{aligned} \dot{\lambda}_1^\star(t) &= -\mathcal{H}_{x_1} = -x_1^\star(t) \\ \dot{\lambda}_2^\star(t) &= -\mathcal{H}_{x_2} = -\lambda_1^\star(t). \end{aligned}$$

That is, singular control arcs are possible when

$$\mathcal{H}_u = \lambda_1^\star(t) - \lambda_2^\star(t) = 0,$$

over a finite interval of time. Upon successive differentiation of the foregoing condition with respect to time, we get

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}_u = \dot{\lambda}_1^\star(t) - \dot{\lambda}_2^\star(t) = -x_1^\star(t) + \lambda_1^\star(t) \\ 0 &= \frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathcal{H}_u = -\dot{x}_1^\star(t) + \dot{\lambda}_1^\star(t) = -x_2^\star(t) - u^\star(t) - x_1^\star(t). \end{aligned}$$

Singular arcs for the problem (3.102–3.105) are therefore of order $p = 1$, and we have

$$u^\star(t) = -x_1^\star(t) - x_2^\star(t).$$

Moreover, the state and adjoint variables must lie on the singular surface defined by

$$\lambda_1^\star(t) = \lambda_2^\star(t) = x_1^\star(t), \tag{3.106}$$

along singular arcs. Observe also that

$$-\frac{\partial}{\partial u}\left[\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathcal{H}_u\right] = 1 > 0,$$

so that the generalized Legendre-Clebsch condition for a minimum holds along singular arcs.

Since the problem is autonomous, $\mathcal{H}$ must be constant along an optimal solution:

$$\tfrac{1}{2}x_1^\star(t)^2 + \lambda_1^\star(t)\, x_2^\star(t) + [\lambda_1^\star(t) - \lambda_2^\star(t)]u^\star(t) = K.$$

In particular, since (3.105) holds along a singular arc, we have

$$\tfrac{1}{2}x_1^\star(t)^2 + x_1^\star(t)\, x_2^\star(t) = K,$$

which gives a 1-parameter family of hyperbolas in the (x_1, x_2)-space.

Upon application of a numerical optimization procedure, it is found that an optimal control for the problem (3.102–3.105) consists of 3 arcs:

1. $u^\star(t) = 10, 0 \le t \le t_1^\star$;
2. $u^\star(t) = -x_1^\star(t) - x_2^\star(t), t_1^\star \le t \le t_2^\star$;
3. $u^\star(t) = -10, t_2^\star \le t \le 2$.

with the following approximate values for the intermediate times: $t_1^\star \approx 0.299$, and $t_2^\star \approx 1.927$. This optimal control, together with the optimal response of the system is represented in Fig. 3.10. below. Note that this control is discontinuous at the junction points between singular and a non-singular arcs. Hence, all the necessary conditions of optimality are satisfied.

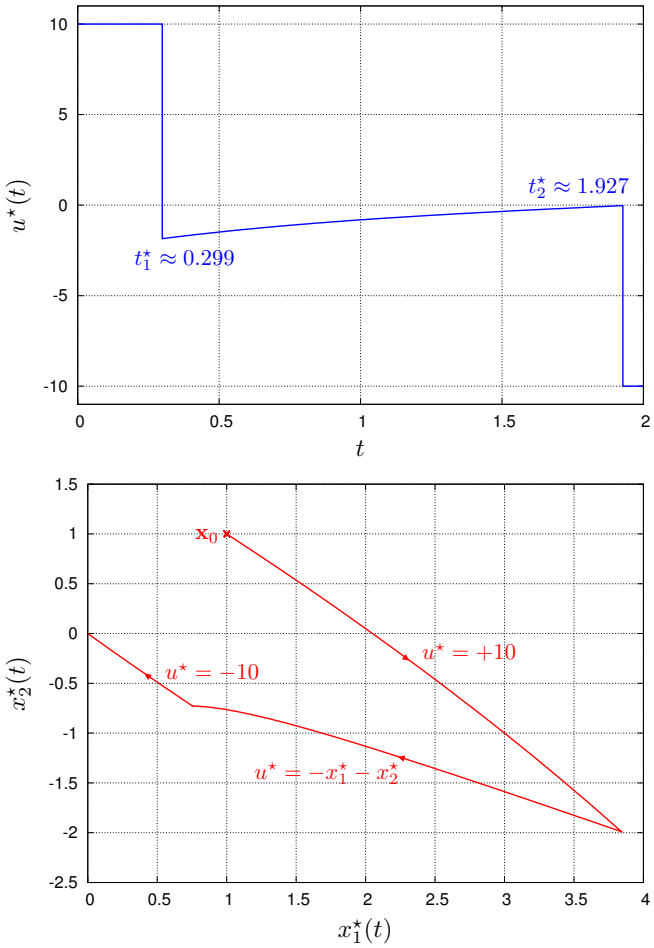


Figure 3.10. Optimal control and response for Example 3.31 – left plot: optimal control vs. time; right plot: optimal response in the phase space.

The junction phenomenon for higher-order singular control problems is a notoriously hard problem, and is still the topic of a great deal of research. In particular, the controls near high-order singular surface may exhibit a *chattering behavior*, i.e, the control has an infinite number of discontinuities in a finite time interval (see, e.g., [58] for more details).

At first sight, one may reasonably expect that a "nice" optimal control problem should have a "nice" solution, and that most if not all reasonable optimal control problems have smooth or piecewise smooth solutions. In 1961, A. T. Fuller [21] put this myth to rest by exhibiting a very simple optimal control problem whose solution chatters.

3.5.5 Optimal Control Problems with Mixed Control-State Inequality Constraints

Optimal control problems with state inequality constraints arise frequently in practical applications. These problems are notoriously hard to solve, and even the theory is not unambiguous, since there exist various forms of the necessary conditions of optimality. We refer the reader to [25] for a recent survey of the various forms of the maximum principle for problems with state inequality constraints. In this subsection, we shall consider optimal control problems with *mixed control-state inequality constraints* only. Problems with *mixed control-state inequality constraints* shall be considered later on in §3.5.6.

Consider the problem to find a piecewise continuous control $\mathbf{u}^\star \in \hat{\mathcal{C}}[t_0, T]^{n_u}$, with associated response $\mathbf{x}^\star \in \hat{\mathcal{C}}^1[t_0, T]^{n_x}$, and a terminal time $t_\mathrm{f}^\star \in [t_0, T]$, such that the following constraints are satisfied and the cost functional takes on its minimum value:

$$\text{minimize:} \quad \mathcal{J}(\mathbf{u}, t_\mathrm{f}) := \int_{t_0}^{t_\mathrm{f}} \ell(t, \mathbf{x}(t), \mathbf{u}(t))\, \mathrm{d}t \tag{3.107}$$

$$\text{subject to:} \quad \dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)); \quad \mathbf{x}(t_0) = \mathbf{x}_0; \quad \mathbf{x}(t_\mathrm{f}) = \mathbf{x}_\mathrm{f} \tag{3.108}$$

$$g_k(t, \mathbf{x}(t), \mathbf{u}(t)) \leq \mathbf{0}, \quad k = 1, \ldots, n_g. \tag{3.109}$$

In what follows, we shall always assume that the components of $\mathbf{g}$ depend *explicitly* on the control $\mathbf{u}$, and the following constraint qualification holds:

$$\operatorname{rank} \left[\begin{array}{cc} \mathbf{g}_\mathbf{u} & \operatorname{diag}(\mathbf{g}) \end{array} \right] = n_g, \tag{3.110}$$

along $(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t))$, $t_0 \leq t \leq t_\mathrm{f}^\star$. In other words, the gradients with respect to $\mathbf{u}$ of all the active constraints $\mathbf{g} \leq \mathbf{0}$ must be linearly independent.

A possible way of tackling optimal control problems with mixed inequality constraints of the form (3.109), is to form a Lagrangian function $\mathcal{L}$ by adjoining $\mathbf{g}$ to the Hamiltonian function $\mathcal{H}$ with a Lagrange multiplier vector function $\boldsymbol{\mu}$,

$$\mathcal{L}(t, \mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}, \boldsymbol{\mu}) := \mathcal{H}(t, \mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}) + \boldsymbol{\mu}^\mathsf{T} \mathbf{g}(t, \mathbf{x}, \mathbf{u}), \tag{3.111}$$

where

$$\mathcal{H}(t, \mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}) := \tilde{\boldsymbol{\lambda}}^\mathsf{T} \tilde{\mathbf{f}}(t, \mathbf{x}, \mathbf{u}) = \lambda_0 \ell(t, \mathbf{x}, \mathbf{u}) + \boldsymbol{\lambda}^\mathsf{T} \mathbf{f}(t, \mathbf{x}, \mathbf{u}). \tag{3.112}$$

The corresponding necessary conditions of optimality are stated in the following theorem:

Theorem 3.33 (Maximum Principle with Mixed Inequality Constraints). *Consider the optimal control problem (3.107–3.109), with fixed initial time t_0 and free terminal time t_f, and where ℓ, $\mathbf{f}$, and $\mathbf{g}$ are continuous and have continuous first partial derivatives with*

respect to $(t, \mathbf{x}, \mathbf{u})$ *on* $[t_0, T] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u}$. *Suppose that* $(\mathbf{u}^\star, t_\mathrm{f}^\star) \in \mathcal{C}[t_0, T]^{n_u} \times [t_0, T)$ *is a minimizer for the problem, and let* $\tilde{\mathbf{x}}^\star$ *denote the optimal (extended) response. If the constraint qualification* (3.110) *holds, then there exist a* $(n_x + 1)$*-dimensional piecewise continuously differentiable vector function* $\tilde{\boldsymbol{\lambda}}^\star(\cdot) = (\lambda_0^\star(\cdot), \boldsymbol{\lambda}^\star(\cdot))$*, and a* n_g*-dimensional piecewise continuous vector function* $\boldsymbol{\mu}^\star(\cdot)$*, such that* $(\tilde{\boldsymbol{\lambda}}^\star(t), \boldsymbol{\mu}^\star(t)) \neq \mathbf{0}$ *for every* $t \in [t_0, t_\mathrm{f}^\star]$*, and:*

(i) the function $\mathcal{H}(\mathbf{x}^\star(t), \mathbf{v}, \tilde{\boldsymbol{\lambda}}^\star(t))$ *attains its minimum on* $U(\mathbf{x}^\star(t), t)$ *at* $\mathbf{v} = \mathbf{u}^\star(t)$*, for every* $t \in [t_0, t_\mathrm{f}^\star]$*,*

$$\mathcal{H}(t, \mathbf{x}^\star(t), \mathbf{v}, \tilde{\boldsymbol{\lambda}}^\star(t)) \geq \mathcal{H}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t), \tilde{\boldsymbol{\lambda}}^\star(t)), \quad \forall \mathbf{v} \in U(\mathbf{x}^\star(t), t), \tag{3.113}$$

where $U(\mathbf{x}, t) := \{\mathbf{u} \in \mathbb{R}^{n_u} : \mathbf{g}(t, \mathbf{x}, \mathbf{u}) \leq \mathbf{0}\}$*;*

(ii) the quadruple $(\mathbf{u}^\star, \mathbf{x}^\star, \tilde{\boldsymbol{\lambda}}^\star, \boldsymbol{\mu}^\star)$ *verifies the equations*

$$\dot{\tilde{\mathbf{x}}}^\star(t) = \mathcal{L}_{\tilde{\boldsymbol{\lambda}}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t), \tilde{\boldsymbol{\lambda}}^\star(t), \boldsymbol{\mu}^\star(t)) \tag{3.114}$$

$$\dot{\tilde{\boldsymbol{\lambda}}}^\star(t) = -\mathcal{L}_{\tilde{\mathbf{x}}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t), \tilde{\boldsymbol{\lambda}}^\star(t), \boldsymbol{\mu}^\star(t)) \tag{3.115}$$

$$\mathbf{0} = \mathcal{L}_{\mathbf{u}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t), \tilde{\boldsymbol{\lambda}}^\star(t), \boldsymbol{\mu}^\star(t)), \tag{3.116}$$

at each instant t *of continuity of* $\mathbf{u}^\star$*;*

(iii) the vector function $\boldsymbol{\mu}^\star$ *is continuous at each instant* t *of continuity of* $\mathbf{u}^\star$*, and satisfies*

$$\mu_k^\star(t)\, g_k(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t)) = 0, \qquad \mu_k^\star(t) \geq 0, \tag{3.117}$$

for each $k = 1, \ldots, n_g$*;*

(iv) the relations

$$\lambda_0^\star(t) = \text{const.} \geq 0 \tag{3.118}$$

$$\mathcal{H}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t), \tilde{\boldsymbol{\lambda}}^\star(t)) = -\int_t^{t_\mathrm{f}^\star} \mathcal{L}_t(\tau, \mathbf{x}^\star(\tau), \mathbf{u}^\star(\tau), \tilde{\boldsymbol{\lambda}}^\star(\tau), \boldsymbol{\mu}^\star(\tau))\, \mathrm{d}\tau, \tag{3.119}$$

are satisfied at any $t \in [t_0, t_\mathrm{f}^\star]$*, and, in particular,* $\mathcal{H}(t_\mathrm{f}^\star, \mathbf{x}^\star(t_\mathrm{f}^\star), \mathbf{u}^\star(t_\mathrm{f}^\star), \tilde{\boldsymbol{\lambda}}^\star(t_\mathrm{f}^\star)) = 0$*.*

Proof. A proof of the theorem can be found, e.g., in [48] or [23]. See also [25] for discussions. □

Similar to problems with simple control constraints of the form $\mathbf{u}(t) \in U$, $t_0 \leq t \leq t_\mathrm{f}$, solutions to optimal control problems with mixed inequality constraints consist of several constrained and unconstrained arcs, which must be pieced together in order to satisfy all the necessary conditions. At the junction points between constrained and unconstrained arcs, the optimal control may or may not be continuous; in the latter case, we get a corner point.[8] In particular, the conditions that must hold at any corner point $\theta \in [t_0, t_\mathrm{f}^\star]$ are

$$\mathbf{x}^\star(\theta^-) = \mathbf{x}^\star(\theta^+) \tag{3.120}$$

$$\tilde{\boldsymbol{\lambda}}^\star(\theta^-) = \tilde{\boldsymbol{\lambda}}^\star(\theta^+) \tag{3.121}$$

$$\mathcal{H}(\theta^-, \mathbf{x}^\star(\theta), \mathbf{u}^\star(\theta^-), \tilde{\boldsymbol{\lambda}}^\star(\theta)) = \mathcal{H}(\theta^+, \mathbf{x}^\star(\theta), \mathbf{u}^\star(\theta^+), \tilde{\boldsymbol{\lambda}}^\star(\theta)), \tag{3.122}$$

[8] As noted in §3.4.3, corners may occur at any point of an optimal trajectory, although they are more likely to occur at junction points rather than at the middle of unconstrained arcs.

where θ^- and θ^+ denote the time just before and just after the corner, respectively; $z(\theta^-)$ and $z(\theta^+)$ denote the left and right limit values of a quantity z at θ, respectively. Since each component of $\mathcal{L}_{\mathbf{u}}$ is also continuous across θ, it follows that $\boldsymbol{\mu}(t)$ is continuous if $\mathbf{u}^\star(t)$ is itself continuous across θ. Unfortunately, there seems to be no a priori method for determining the existence of corners.

Remark 3.34 (Extension to General State Terminal Constraints). The Maximum Principle in Theorem 3.33 can be extended to the case where general terminal constraints are specified on the state variables (in lieu of the terminal state condition $\mathbf{x}(t_\mathrm{f}) = \mathbf{x}_\mathrm{f}$) as

$$\psi_k(t_\mathrm{f}, \mathbf{x}(t_\mathrm{f})) = \mathbf{0}, \quad k = 1, \ldots, n_\psi. \tag{3.123}$$

$$\kappa_k(t_\mathrm{f}, \mathbf{x}(t_\mathrm{f})) \leq \mathbf{0}, \quad k = 1, \ldots, n_\kappa, \tag{3.124}$$

and a terminal term is added to the cost functional (3.107) as

$$\mathcal{J}(\mathbf{u}, t_\mathrm{f}) := \int_{t_0}^{t_\mathrm{f}} \ell(t, \mathbf{x}(t), \mathbf{u}(t))\, \mathrm{d}t + \phi(t_\mathrm{f}, \mathbf{x}(t_\mathrm{f})), \tag{3.125}$$

where ϕ, $\boldsymbol{\psi}$, and $\boldsymbol{\kappa}$ are continuous and have continuous first partial derivatives with respect to $(t, \mathbf{x})$, for all $(t, \mathbf{x}) \in [t_0, T] \times \mathbb{R}^{n_x}$. Suppose that the terminal constraints (3.123,3.124) satisfy the constraint qualification

$$\operatorname{rank} \begin{bmatrix} \boldsymbol{\psi}_{\mathbf{x}} & \mathbf{0} \\ \boldsymbol{\kappa}_{\mathbf{x}} & \operatorname{diag}(\boldsymbol{\kappa}) \end{bmatrix} = n_\psi + n_\kappa, \tag{3.126}$$

at $(t_\mathrm{f}^\star, \mathbf{x}^\star(t_\mathrm{f}^\star))$. Then, in addition to the necessary conditions of optimality given in Theorem 3.33, there exist Lagrange multiplier vectors $\boldsymbol{\nu}^\star \in \mathbb{R}^{n_\psi}$ and $\boldsymbol{\zeta}^\star \in \mathbb{R}^{n_\kappa}$ such that the following transversal conditions hold:

$$\boldsymbol{\lambda}^\star(t_\mathrm{f}^\star) = \Phi_{\mathbf{x}}(t_\mathrm{f}^\star, \mathbf{x}^\star(t_\mathrm{f}^\star)) \tag{3.127}$$

$$\mathcal{H}(t_\mathrm{f}^\star, \mathbf{x}^\star(t_\mathrm{f}^\star), \mathbf{u}^\star(t_\mathrm{f}^\star), \boldsymbol{\lambda}^\star(t_\mathrm{f}^\star)) + \Phi_t(t_\mathrm{f}^\star, \mathbf{x}^\star(t_\mathrm{f}^\star)) = 0, \tag{3.128}$$

where $\Phi := \lambda_0^\star \phi + \boldsymbol{\nu}^{\star\mathsf{T}} \boldsymbol{\psi} + \boldsymbol{\zeta}^{\star\mathsf{T}} \boldsymbol{\kappa}$. Moreover,

$$\psi_k(t_\mathrm{f}^\star, \mathbf{x}^\star(t_\mathrm{f}^\star)) = 0, \tag{3.129}$$

for each $k = 1, \ldots, n_\psi$, and

$$\zeta_k^\star\, \kappa_k(t_\mathrm{f}^\star, \mathbf{x}^\star(t_\mathrm{f}^\star)) = 0, \qquad \zeta_k^\star \geq 0, \tag{3.130}$$

for each $k = 1, \ldots, n_\kappa$.

In practice, applying Theorem 3.33 (and Remark 3.34) requires that an assumption be made *a priori* on the sequence of (unconstrained and constrained) arcs in the optimal solution, as well as on the set of active (inequality) terminal constraints. Then, based on the postulated *structure of the optimal solution*, one shall check whether a pair $(\mathbf{u}(\cdot), \mathbf{x}(\cdot))$, along with vector functions $\tilde{\boldsymbol{\lambda}}(\cdot)$, $\boldsymbol{\mu}(\cdot)$, and Lagrange multiplier vectors $\boldsymbol{\nu}^\star$, $\boldsymbol{\zeta}^\star$, can be found such that *all* of the necessary conditions of optimality are satisfied. If this is the case, then the corresponding control is a candidate optimal control for the problem; otherwise, one needs to investigate alternative solution structures, i.e., postulate different sequences of arcs and/or sets of active terminal constraints. An illustration of these considerations in given in Example 3.35 hereafter.

Example 3.35 (Optimal Control Problem with Mixed Inequality Constraints). Consider the scalar optimal control problem:

$$\text{minimize:} \quad \mathcal{J}(u) := \int_0^1 u(t)\,\mathrm{d}t \tag{3.131}$$

$$\text{subject to:} \quad \dot{x}(t) = -u(t); \quad x(0) = -1 \tag{3.132}$$

$$u(t) \leq 0, \quad x(t) - u(t) \leq 0, \quad 0 \leq t \leq 1, \tag{3.133}$$

where the control u is taken in the set of piecewise continuous functions, $u \in \hat{\mathcal{C}}[0,1]$. Observe that both path constraints are of mixed type, and can be rewritten as

$$x(t) \leq u(t) \leq 0, \quad 0 \leq t \leq 1. \tag{3.134}$$

The objective being to minimize the integral of $u(t)$, and since $u(t)$ is lower bounded by the state $x(t)$ via (3.134), a rather natural guess for the optimal solution is to consider that the mixed state constraint $x(t) - u(t) \leq 0$ is active for each $0 \leq t \leq 1$. We shall now check whether the necessary conditions of optimality in Theorem 3.33 can be satisfied under this choice.

- Let us suppose first that the problem (3.131–3.133) is not abnormal, and take $\lambda_0(t) = 1$ throughout. That is, the Hamiltonian function for the problem reads

  $$\mathcal{H}(x,u,\lambda) = u(1-\lambda),$$

  and the Lagrangian function, obtained by adjoining the mixed inequality constraints, reads

  $$\mathcal{L}(x,u,\lambda,\boldsymbol{\mu}) = \mathcal{H}(x,u,\lambda) + \mu_1(x-u) + \mu_2 u = (1-\lambda-\mu_1+\mu_2)u + \mu_1 x.$$

- The mixed state constraint $x(t) - u(t) \leq 0$ being active for each $0 \leq t \leq 1$, we have

  $$u^\star(t) = x^\star(t),$$

  and, from (3.132),

  $$x^\star(t) = -\mathrm{e}^{-t}, \quad 0 \leq t \leq 1.$$

  $x^\star(t)$ and, hence, $u^\star(t)$ are thus negative at any time, and from the complementarity slackness condition (3.117) we get

  $$\mu_2^\star(t) = 0, \quad 0 \leq t \leq 1.$$

  In turn, the stationarity condition (3.116) yields

  $$0 = 1 - \lambda^\star(t) - \mu_1^\star(t) + \mu_2^\star(t) = 1 - \lambda^\star(t) - \mu_1^\star(t),$$

  from which we get

  $$\mu_1^\star(t) = 1 - \lambda^\star(t), \quad 0 \leq t \leq 1.$$

- From (3.115), the differential equation giving the adjoint variable $\lambda^\star$ is

  $$\dot{\lambda}^\star(t) = -\mu_1^\star(t) = \lambda^\star(t) - 1,$$

and the terminal state being unspecified, from (3.127), we get

$$\lambda^\star(1) = 0.$$

Therefore,

$$\lambda^\star(t) = 1 - \mathrm{e}^{t-1} < 1, \quad 0 \le t \le 1,$$

and,

$$\mu_1^\star(t) = \mathrm{e}^{t-1} > 0, \quad 0 \le t \le 1,$$

hence satisfying the dual condition (3.117).

- At this point, the condition (3.119) imposing that the Hamiltonian function be constant along $(u^\star, x^\star, \lambda^\star)$ is readily verified,

$$\mathcal{H}(x^\star(t), u^\star(t), \lambda^\star(t)) = u^\star(t)(1 - \lambda^\star(t)) = -\mathrm{e}^{-1}, \quad 0 \le t \le 1.$$

- Finally, the minimum condition (3.113),

$$u^\star(t) = \begin{cases} 0 & \text{if } \lambda^\star(t) > 1 \\ x^\star(t) & \text{if } \lambda^\star(t) < 1, \end{cases}$$

is satisfied by the control $u^\star(t) = x^\star(t)$, since $\lambda^\star(t) < 1$ at each $0 \le t \le 1$.

Overall, we have checked that all the necessary conditions of optimality are satisfied provided that the mixed state constraint $x(t) - u(t) \le 0$ is active at any time. Therefore, $u^\star(t) = x^\star(t) = -\mathrm{e}^{-t}$, $0 \le t \le 1$, is a candidate optimal control for the problem (3.131–3.133).

3.5.6 Optimal Control Problems with Pure State Inequality Constraints

Besides mixed state inequality constraints, it is common to require that one or several state variables remain nonnegative during the system operation, e.g.,

$$x_i(t) \ge 0, \quad t_0 \le t \le t_\mathrm{f},$$

for $i \in \{1, \ldots, n_x\}$. More generally, optimal control problems may have so-called *pure state inequality constraints* of the form

$$h_k(t, \mathbf{x}(t)) \le 0, \quad k = 1, \ldots, n_h.$$

Pure state constraints are, in principle, more difficult to deal with than mixed control-state constraints, since $\mathbf{h}$ does not explicitly depend on $\mathbf{u}$, and $\mathbf{x}$ can be controlled only indirectly via propagation through the state equations. It is therefore convenient to differentiate $\mathbf{h}$ with respect to t as many times as required until it contains a control variable. For the jth

constraint, we have

$$
\begin{aligned}
h_j^0(t,\mathbf{x},\mathbf{u}) &:= h_j(t,\mathbf{x}) \\
h_j^1(t,\mathbf{x},\mathbf{u}) &:= \frac{\mathrm{d}}{\mathrm{d}t} h_j^0(t,\mathbf{x},\mathbf{u}) = (h_j)_{\mathbf{x}}(t,\mathbf{x})\,\mathbf{f}(t,\mathbf{x},\mathbf{u}) + (h_j)_t(t,\mathbf{x}) \qquad (3.135)\\
h_j^2(t,\mathbf{x},\mathbf{u}) &:= \frac{\mathrm{d}}{\mathrm{d}t} h_j^1(t,\mathbf{x},\mathbf{u}) = (h_j^1)_{\mathbf{x}}(t,\mathbf{x},\mathbf{u})\,\mathbf{f}(t,\mathbf{x},\mathbf{u}) + (h_j^1)_t(t,\mathbf{x},\mathbf{u}) \\
&\vdots \\
h_j^p(t,\mathbf{x},\mathbf{u}) &:= \frac{\mathrm{d}}{\mathrm{d}t} h_j^{p-1}(t,\mathbf{x},\mathbf{u}) = (h_j^{p-1})_{\mathbf{x}}(t,\mathbf{x},\mathbf{u})\,\mathbf{f}(t,\mathbf{x},\mathbf{u}) + (h_j^{p-1})_t(t,\mathbf{x},\mathbf{u}). \qquad (3.136)
\end{aligned}
$$

Then, h_j is said to be of *order*[9] p_j if

$$(h_j^i)_{\mathbf{u}}(t,\mathbf{x},\mathbf{u}) = 0 \text{ for } 0 \le i \le p_j - 1, \qquad (h_j^{p_j})_{\mathbf{u}}(t,\mathbf{x},\mathbf{u}) \neq 0.$$

A number of definition are in order. With respect to the jth constraint $h_j \le 0$, a subinterval $(\theta_1,\theta_2) \subset [t_0,t_{\mathrm{f}}]$, with $\theta_1 < \theta_2$, is called an *interior interval* of a feasible response $\mathbf{x}$ if $h_j(t,\mathbf{x}(t)) > 0$ for all $t \in (\theta_1,\theta_2)$. An interval $[\theta_1,\theta_2]$, with $\theta_1 < \theta_2$, is called a *boundary interval* if $h_j(t,\mathbf{x}(t)) = 0$ for all $t \in [\theta_1,\theta_2]$. An instant θ_1 is called an *entry time* if there is an interior interval ending at $t = \theta_1$ and a boundary interval starting at θ_1; correspondingly, θ_2 is called an *exit time* if a boundary interval ends at θ_2 and an interior interval starts at θ_2. If the response $\mathbf{x}$ just touches the boundary at time θ_c, i.e., $h_j(\theta_c,\mathbf{x}(\theta_c)) = 0$, and $\mathbf{x}$ is in the interior just before and after θ_c, then θ_c is called a *contact time*. Taken together, entry, exit, and contact times are *junction times*. These definitions are illustrated on Fig. 3.11. below.

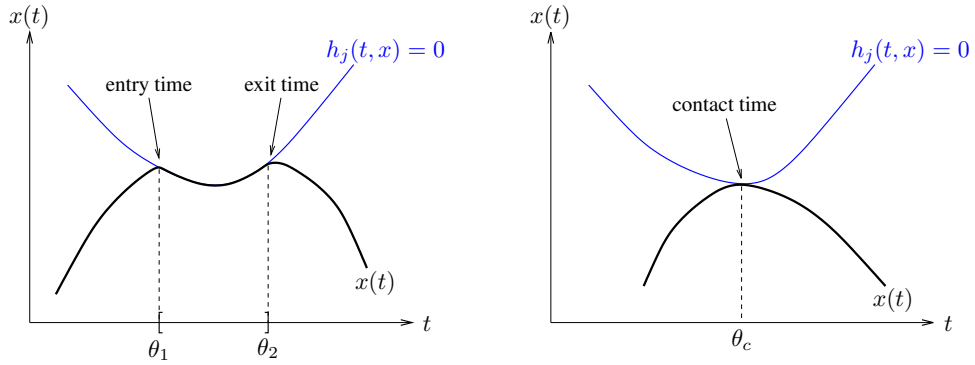


Figure 3.11. Junction types for optimal control problems with pure state inequality constraints.

[9]Notice that, unlike singular control arcs, the order p of a state inequality constraints is equal to the minimum number of time differentiations needed to have $\mathbf{u}$ appear explicitely in the expression of h^p.

We begin with the case of first-order state inequality constraints, $p_j = 1$, $i = 1, \ldots, n_h$. Let $(\mathbf{u}^\star, \mathbf{x}^\star)$ be an optimal pair for the problem

$$\text{minimize:} \quad \mathcal{J}(\mathbf{u}, t_\mathrm{f}) := \int_{t_0}^{t_\mathrm{f}} \ell(t, \mathbf{x}(t), \mathbf{u}(t))\, \mathrm{d}t \tag{3.137}$$

$$\text{subject to:} \quad \dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)); \quad \mathbf{x}(t_0) = \mathbf{x}_0; \quad \mathbf{x}(t_\mathrm{f}) = \mathbf{x}_\mathrm{f} \tag{3.138}$$

$$h_k(t, \mathbf{x}(t)) \leq \mathbf{0}, \quad k = 1, \ldots, n_h. \tag{3.139}$$

We shall always assume that the constraint qualification

$$\operatorname{rank} \left[\begin{array}{cc} \mathbf{h}_\mathbf{u}^1 & \operatorname{diag}(\mathbf{h}) \end{array} \right] = n_h, \tag{3.140}$$

holds at each $(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t))$; i.e., the gradients of h_j^1 with respect to $\mathbf{u}$ of the active constraints $h_j = 0$, $j \in \{1, \ldots, n_h\}$, must be linearly independent along an optimal trajectory.

Suppose that $[\theta_1, \theta_2]$ is a boundary interval for the jth constraint. In order to prevent h_j from being violated, we must have that $h_j^1(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t)) \leq 0$ for each $t \in [\theta_1, \theta_2]$. Hence, one can formally impose the constraint

$$h_j^1(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t)) \leq 0 \quad \text{whenever } h_j(t, \mathbf{x}^\star(t)) = 0.$$

A convenient way of associating a multiplier function $\eta_j(\cdot)$ to the former condition constraint is by imposing the complementarity slackness condition $\eta_j(t) h_j(t, \mathbf{x}^\star(t)) = 0$, which makes $\eta_j(t) = 0$ each time $h_j(t, \mathbf{x}^\star(t)) < 0$. This also motivates the following definition of the Lagrangian function:

$$\mathcal{L}^1(t, \mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}, \boldsymbol{\eta}) := \mathcal{H}(t, \mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}) + \boldsymbol{\eta}^\mathsf{T} \mathbf{h}^1(t, \mathbf{x}, \mathbf{u}). \tag{3.141}$$

Since the constraints are adjoined indirectly to form the Lagrangian (i.e., they are adjoined via their first time derivative), this approach is called the *indirect adjoining approach*; it was first suggested by Pontryagin [41].[10]

At the entry time θ_1 of a boundary interval for the jth constraint, it is necessary to require that the interior-point constraint

$$h_j(\theta_1, \mathbf{x}^\star(\theta_1)) = 0 \tag{3.142}$$

be satisfied, i.e., the phase velocity is tangential to the boundary at θ_1. These extra constraints give rise to jump conditions for the adjoint variables and the Hamiltonian function as

$$\boldsymbol{\lambda}^\star(\theta_1^-)^\mathsf{T} = \boldsymbol{\lambda}^\star(\theta_1^+)^\mathsf{T} + \pi_j(\theta_1)\, (h_j)_\mathbf{x}(\theta_1, \mathbf{x}^\star(\theta_1)) \tag{3.143}$$

$$\mathcal{H}[\theta_1^-] = \mathcal{H}[\theta_1^+] - \pi_j(\theta_1)\, (h_j)_t(\theta_1, \mathbf{x}^\star(\theta_1)), \tag{3.144}$$

where $\pi_j(\theta_1) \in \mathbb{R}$ is a Lagrange multiplier. Condition (3.144) determines the entry time θ_1, while $\pi_j(\theta_1)$ is so chosen that the interior point constraint (3.142) is satisfied; note that

[10] Another approach referred to as the *direct adjoining approach* has also been proposed for dealing with state inequality constrained optimal control problems. In contrast to the indirect approach, the Lagrangian function $\mathcal{L}$ is formed by adjoining directly the constraints (3.139) as

$$\mathcal{L}^0(t, \mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}, \boldsymbol{\eta}) := \mathcal{H}(t, \mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}) + \boldsymbol{\eta}^\mathsf{T} \mathbf{h}(t, \mathbf{x}).$$

We refer the interested reader to [25] for a broad discussion of the direct adjoining approach and for comparison and links between with the indirect approach. See also [27].

$\pi_j(\theta_1)$ influences (3.142) only indirectly by propagating through the adjoint equations via (3.143). Note also that from the tangency condition (3.142) holding at the entry point θ_1 could have been placed at the exit point θ_2 instead; we would then have that $\boldsymbol{\lambda}^\star$ and $\mathcal{H}$ are discontinuous at θ_2, and continuous at θ_1.

Overall, these considerations are formalized in the following theorem:

Theorem 3.36 (Maximum Principle with First-Order Pure Inequality Constraints). *Consider the optimal control problem (3.137–3.139), with fixed initial time t_0 and free terminal time t_f. Here, ℓ is continuous and has continuous first partial derivatives with respect to $(t, \mathbf{x}, \mathbf{u})$ on $[t_0, T] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u}$; $\mathbf{f}$, and $\mathbf{h}$ are continuous and have continuous partial derivatives with respect to $(t, \mathbf{x}, \mathbf{u})$ up to second order on $[t_0, T] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u}$. Suppose that $(\mathbf{u}^\star, t_\mathrm{f}^\star) \in \mathcal{C}[t_0, T]^{n_u} \times [t_0, T)$ is a minimizer for the problem, and let $\tilde{\mathbf{x}}^\star$ denote the optimal (extended) response. If the constraint qualification* (3.140) *holds with $p_1 = \cdots = p_{n_h} = 1$, then there exist a $(n_x + 1)$-dimensional piecewise continuous vector function $\tilde{\boldsymbol{\lambda}}^\star(\cdot) = (\lambda_0^\star(\cdot), \boldsymbol{\lambda}^\star(\cdot))$ whose continuous segments are continuously differentiable, a n_h-dimensional piecewise continuous vector function $\boldsymbol{\eta}^\star(\cdot)$, and Lagrange multiplier vectors $\boldsymbol{\pi}^\star(\theta_1) \in \mathbb{R}^{n_h}$ at each point θ_1 of discontinuity of $\tilde{\boldsymbol{\lambda}}^\star$, such that $(\tilde{\boldsymbol{\lambda}}^\star(t), \boldsymbol{\eta}^\star(t)) \neq \mathbf{0}$ for every $t \in [t_0, t_\mathrm{f}^\star]$, and:*

(i) the function $\mathcal{H}(\mathbf{x}^\star(t), \mathbf{v}, \tilde{\boldsymbol{\lambda}}^\star(t))$ attains its minimum on $U^1(\mathbf{x}^\star(t), t)$ at $\mathbf{v} = \mathbf{u}^\star(t)$, for every $t \in [t_0, t_\mathrm{f}^\star]$,

$$\mathcal{H}(t, \mathbf{x}^\star(t), \mathbf{v}, \tilde{\boldsymbol{\lambda}}^\star(t)) \geq \mathcal{H}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t), \tilde{\boldsymbol{\lambda}}^\star(t)), \quad \forall \mathbf{v} \in U^1(\mathbf{x}^\star(t), t), \tag{3.145}$$

where $U^1(\mathbf{x}, t) := \{\mathbf{u} \in \mathbb{R}^{n_u} : \mathbf{h}^1(t, \mathbf{x}, \mathbf{u}) \leq \mathbf{0}$ if $\mathbf{h}(t, \mathbf{x}) = \mathbf{0}\}$;

(ii) the quadruple $(\mathbf{u}^\star, \mathbf{x}^\star, \tilde{\boldsymbol{\lambda}}^\star, \boldsymbol{\eta}^\star)$ verifies the equations

$$\dot{\tilde{\mathbf{x}}}^\star(t) = \mathcal{L}^1_{\tilde{\boldsymbol{\lambda}}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t), \tilde{\boldsymbol{\lambda}}^\star(t), \boldsymbol{\eta}^\star(t)) \tag{3.146}$$

$$\dot{\tilde{\boldsymbol{\lambda}}}^\star(t) = -\mathcal{L}^1_{\tilde{\mathbf{x}}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t), \tilde{\boldsymbol{\lambda}}^\star(t), \boldsymbol{\eta}^\star(t)) \tag{3.147}$$

$$\mathbf{0} = \mathcal{L}^1_{\mathbf{u}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t), \tilde{\boldsymbol{\lambda}}^\star(t), \boldsymbol{\eta}^\star(t)), \tag{3.148}$$

on each interval of continuity of $\mathbf{u}^\star$ and $\tilde{\boldsymbol{\lambda}}^\star$;

(iii) the vector function $\boldsymbol{\eta}^\star$ satisfies the conditions

$$\eta_k^\star(t)\ h_k(t, \mathbf{x}^\star(t)) = 0, \qquad \eta_k^\star(t) \geq 0, \qquad \dot{\eta}_k^\star(t) \leq 0, \tag{3.149}$$

for each $k = 1, \ldots, n_h$;

(iv) at any entry/contact time θ_1, the adjoint function and the Hamiltonian function may have discontinuities of the form

$$\boldsymbol{\lambda}^\star(\theta_1^-)^\mathsf{T} = \boldsymbol{\lambda}^\star(\theta_1^+)^\mathsf{T} + \boldsymbol{\pi}^\star(\theta_1)^\mathsf{T} \mathbf{h}_{\mathbf{x}}(\theta_1, \mathbf{x}^\star(\theta_1)) \tag{3.150}$$

$$\mathcal{H}[\theta_1^-] = \mathcal{H}[\theta_1^+] - \boldsymbol{\pi}^\star(\theta_1)^\mathsf{T} \mathbf{h}_t(\theta_1, \mathbf{x}^\star(\theta_1)), \tag{3.151}$$

where the Lagrange multiplier vector $\boldsymbol{\pi}^\star(\theta_1)$ satisfies the conditions

$$\pi_k^\star(\theta_1)\ h_k(\theta_1, \mathbf{x}^\star(\theta_1)) = 0, \qquad \pi_k^\star(\theta_1) \geq 0, \qquad \pi_k^\star(\theta_1) \geq \eta_k^\star(\theta_1^+), \tag{3.152}$$

for each $k = 1, \ldots, n_h$*;*

(v) the relations

$$\lambda_0^\star(t_\mathrm{f}^\star) \geq 0 \tag{3.153}$$

$$\mathcal{H}(t_\mathrm{f}^\star, \mathbf{x}^\star(t_\mathrm{f}^\star), \mathbf{u}^\star(t_\mathrm{f}^\star), \tilde{\boldsymbol{\lambda}}^\star(t_\mathrm{f}^\star)) = 0, \tag{3.154}$$

are satisfied at the terminal time.

Proof. A proof of the theorem can be found, e.g., in [32]. See also [25] for discussions. □

Note the additional complementarity slackness condition $\dot{\eta}_k^\star(t) \leq 0$, $k = 1, \ldots, n_h$, in (3.149), which imposes that the multiplier function $\eta_k(\cdot)$ be nondecreasing on boundary intervals of h_k, and can only jump upwards in the case it is discontinuous. This condition is in fact absent in early papers on inequality state constraints, yet its omission may may lead to spurious extremals as shown by [53]. Also omitted in the literature is the necessary condition $\pi_k^\star(\theta_1) \geq \eta_k^\star(\theta_1^+)$, $k = 1, \ldots, n_h$, in (3.152) – see related discussion in [25].

Remark 3.37 (Mixed Sets of Pure and Mixed State Inequality Constraints). Apart from the extension of Theorem 3.36 to problems with general state constraints for which the reader is referred to Remark 3.34 above, many optimal control problems of interest contain mixed sets of pure and mixed inequality constraints:

$$\text{minimize:} \quad \mathcal{J}(\mathbf{u}, t_\mathrm{f}) := \int_{t_0}^{t_\mathrm{f}} \ell(t, \mathbf{x}(t), \mathbf{u}(t))\, \mathrm{d}t \tag{3.155}$$

$$\text{subject to:} \quad \dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)); \quad \mathbf{x}(t_0) = \mathbf{x}_0; \quad \mathbf{x}(t_\mathrm{f}) = \mathbf{x}_\mathrm{f} \tag{3.156}$$

$$g_k(t, \mathbf{x}(t), \mathbf{u}(t)) \leq \mathbf{0}, \quad k = 1, \ldots, n_g \tag{3.157}$$

$$h_k(t, \mathbf{x}(t)) \leq \mathbf{0}, \quad k = 1, \ldots, n_h. \tag{3.158}$$

As a recipe for isolating candidate optimal controls $(\mathbf{u}^\star, \mathbf{x}^\star)$ for such problems, one can adjoin the mixed inequality constraints to the Lagrangian function as

$$\mathcal{L}(t, \mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}, \boldsymbol{\mu}) := \mathcal{H}(t, \mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}) + \boldsymbol{\mu}^\mathsf{T} \mathbf{g}(t, \mathbf{x}, \mathbf{u}) + \boldsymbol{\eta}^\mathsf{T} \mathbf{h}^1(t, \mathbf{x}), \tag{3.159}$$

and restrict the control region U^1 as

$$U^1(\mathbf{x}, t) := \{\mathbf{u} \in \mathbb{R}^{n_u} : \mathbf{g}(t, \mathbf{x}, \mathbf{u}) \leq \mathbf{0} \text{ and } \mathbf{h}^1(t, \mathbf{x}, \mathbf{u}) \leq \mathbf{0} \text{ if } \mathbf{h}(t, \mathbf{x}) = \mathbf{0}\}.$$

If the *strengthened constraint qualification*

$$\operatorname{rank} \begin{bmatrix} \mathbf{g}_\mathbf{u} & \operatorname{diag}(\mathbf{g}) & \mathbf{0} \\ \mathbf{h}_\mathbf{u}^1 & \mathbf{0} & \operatorname{diag}(\mathbf{h}) \end{bmatrix} = n_g + n_h \tag{3.160}$$

holds, then the necessary conditions of optimality are those given in Theorem 3.36, as well as the additional conditions

$$\mu_k^\star(t)\, g_k(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t)) = 0, \qquad \mu_k^\star(t) \geq 0, \tag{3.161}$$

for the n_g-dimensional piecewise continuous vector function $\boldsymbol{\mu}^\star(\cdot)$ associated with the mixed inequality constraints (3.157). However, the reader should be aware that a general theorem addressing the problem (3.155–3.158), in the context of the indirect adjoining

approach, was still unavailable until recently in the literature [25] (only various subsets of the conditions stated above have been proved).

A simple optimal control problem with mixed and first-order pure state constraints is treated subsequently in Example 3.38.

Example 3.38 (Optimal Control Problem with Both Mixed and First-Order Pure Inequality Constraints). Consider the following scalar optimal control problem, with $\varrho \geq 0$:

$$\text{minimize:} \quad \mathcal{J}(u) := \int_0^3 \mathrm{e}^{-\varrho t} u(t)\, \mathrm{d}t \tag{3.162}$$

$$\text{subject to:} \quad \dot{x}(t) = u(t); \quad x(0) = 0 \tag{3.163}$$

$$0 \leq u(t) \leq 3, \quad 0 \leq t \leq 3 \tag{3.164}$$

$$1 - x(t) - (t-2)^2 \leq 0, \quad 0 \leq t \leq 3, \tag{3.165}$$

with $\varrho \geq 0$, where the control u is taken in the set of piecewise continuous functions, $u \in \mathcal{C}[0,3]$.

By inspection, it can be argued that a candidate optimal control $u^\star$ and its optimal response $x^\star$ for the problem (3.162–3.162) are as follows:

$$u^\star(t) = \begin{cases} 0, & 0 \leq t \leq 1^- \\ -2(t-2), & 1^+ \leq t \leq 2^- \\ 0, & 2^+ \leq t \leq 3, \end{cases} \qquad x^\star(t) = \begin{cases} 0, & 0 \leq t \leq 1^- \\ 1-(t-2)^2, & 1^+ \leq t \leq 2^- \\ 1, & 2^+ \leq t \leq 3. \end{cases}$$

The control $u^\star$ and its response $x^\star$ are shown in Fig. 3.12. below. In the first arc, $u^\star(t)$ is at its lower bound so that the integrand of the cost functional takes on its least value. At time $t = 1$, the pure state constraint $h := 1 - x - (t-2)^2$ becomes active, and $u^\star(t)$ must be increased so that h does not become violated; minimizing the integrand of the cost functional in the second arc then consists in taking $u^\star(t)$ so that $h(t, x^\star(t)) = 0$. Finally, $u^\star(t)$ is again at its lower bound in the third arc, since the state constraint has become inactive at $t = 2$.

Letting $g_1 := u - 3$ and $g_2 := -u$, and noting that the state constraint h is first-order with

$$h^1(t,x,u) = \dot{x} - 2(t-2) = u - 2(t-2),$$

the strengthened constraint qualification (3.160) reads

$$\operatorname{rank} \begin{bmatrix} 1 & u^\star(t) - 3 & 0 & 0 \\ -1 & 0 & -u^\star(t) & 0 \\ 1 & 0 & 0 & u^\star(t) - 2(t-2) \end{bmatrix} = 3 \tag{3.166}$$

It is readily checked that this rank condition holds for the pair $(u^\star, x^\star)$, along each arc. Hence, it makes sense to check whether $(u^\star, x^\star)$ satisfies the necessary conditions of optimality presented in Theorem 3.36 and Remark 3.37.

- Let us suppose first that the problem (3.131–3.133) is not abnormal, and take $\lambda_0(t) = 1$ throughout. The Hamiltonian function for the problem reads

$$\mathcal{H}(x,u,\lambda) = u(\mathrm{e}^{-\varrho t} + \lambda).$$

Moreover, the state constraint h being first-order, the Lagrangian function is obtained by adjoining both h^1 and the control bounds to the Hamiltonian function as

$$\mathcal{L}^1(x,u,\lambda,\boldsymbol{\mu},\eta) = \mathcal{H}(x,u,\lambda) - \mu_1(u-3) - \mu_2 u + \eta(u-2(t-2)).$$

- From (3.147), we have
$$\dot{\lambda}^\star(t) = -\mathcal{L}^1_x = 0,$$
with $\lambda^\star(3) = 0$ since the terminal state $x^\star(3)$ is free. Because $t = 1$ is an entry time for the state constraint h, the condition (3.150) yield
$$\lambda^\star(1^-) = \lambda^\star(1^+) + \pi^\star(1)\, h_x(1, x^\star(1)) = -\pi^\star(1)$$
where $\pi^\star(1)$ is obtained from (3.151) as
$$\pi^\star(1) = \mathrm{e}^{-\varrho}.$$
Notice, in particular, that $\pi^\star(1) \geq 0$, as imposed by (3.152). Overall, $\lambda^\star$ is thus given by
$$\lambda^\star(t) = \begin{cases} -\mathrm{e}^{-\varrho}, & 0 \leq t \leq 1^- \\ 0, & 1^+ \leq t \leq 3. \end{cases}$$

- The mixed state constraint $u^\star(t) - 3 \leq 0$ remaining inactive at any time, (3.161) gives
$$\mu_1^\star(t) = 0, \quad 0 \leq t \leq 3.$$

- From the stationarity condition (3.148), we get
$$0 = \mathcal{L}^1_u = \mathrm{e}^{-\varrho t} + \lambda^\star(t) - \mu_2^\star(t) - \eta(t),$$
which, together with (3.149) and (3.161), yields:
$$\mu_2^\star(t) = \begin{cases} \mathrm{e}^{-\varrho t} - \mathrm{e}^{-\varrho}, & 0 \leq t \leq 1^- \\ 0, & 1^+ \leq t \leq 2^- \\ \mathrm{e}^{-\varrho t}, & 2^+ \leq t \leq 3, \end{cases} \qquad \eta^\star(t) = \begin{cases} 0, & 0 \leq t \leq 1^- \\ \mathrm{e}^{-\varrho t}, & 1^+ \leq t \leq 2^- \\ 0, & 2^+ \leq t \leq 3. \end{cases}$$
Observe that the non-negativity requirements $\eta(t) \geq 0$ and $\mu_2(t) \geq 0$ are satisfied at any time, as well as the necessary conditions $\dot{\eta}(t) \leq 0$ and $\eta(1^+) \leq \pi(1)$.

- Finally, since $\lambda^\star(t) + \mathrm{e}^{-\varrho t} > 0$ for each $t > 0$, and since $0 < u^\star(t) < 3$ whenever $h(t, x^\star(t)) = 0$, the control $u^\star(t)$ achieves the least possible value of $\mathcal{H}(t, x^\star(t), \cdot, \lambda^\star)$ on the set
$$U^1(t,x) = \begin{cases} \{u \in \mathbb{R} : 0 \leq u \leq 3\}, & 0 \leq t \leq 1^- \text{ or } 2^+ \leq t \leq 3 \\ \{u \in \mathbb{R} : -2(t-2) \leq u \leq 3\}, & 1^+ \leq t \leq 2^-. \end{cases}$$
That is, the minimum condition (3.145) is satisfied.

Overall, we have checked that all the necessary conditions of optimality are satisfied for the pair $(u^\star, x^\star)$. Thus, $u^\star$ is a candidate optimal control for the problem (3.131–3.133).

Note that one obtains a contact point at $t = 2$ by considering the same example with $\varrho < 0$.

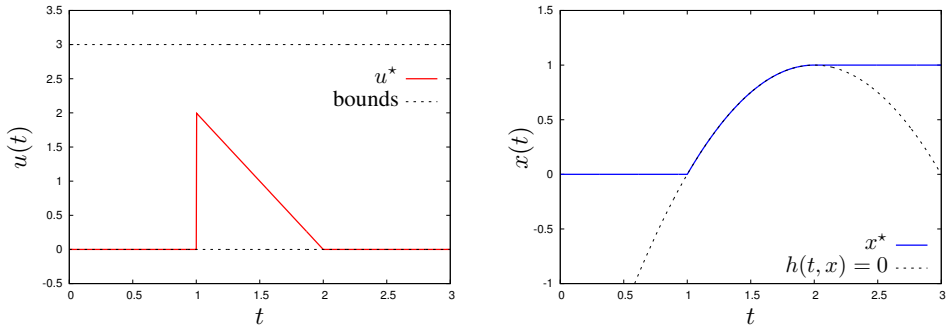


Figure 3.12. Optimal control and response for Example 3.38.

We now turn to optimal control problems having pure state inequality constraints of order $p \geq 2$. For simplicity, we shall assume that there is only one state constraint in the problem (3.137–3.139), i.e., $n_h = 1$. Following the same approach as for first-order state constraints (indirect adjoining approach), the Lagrangian function is now given by

$$\mathcal{L}^p(t, \mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}, \eta) := \mathcal{H}(t, \mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}) + \eta h^p(t, \mathbf{x}, \mathbf{u}), \tag{3.167}$$

with h^p defined in (3.136), and the control region U^p is now defined as

$$U^p(\mathbf{x}, t) := \{\mathbf{u} \in \mathbb{R}^{n_u} : h^p(t, \mathbf{x}, \mathbf{u}) \leq \mathbf{0} \text{ if } h(t, \mathbf{x}) = \mathbf{0}\}.$$

Further, for a pth order constraint, the following interior-point constraints

$$\begin{aligned} h(\theta_1, \mathbf{x}(\theta_1)) &= 0 \\ h^1(\theta_1, \mathbf{x}(\theta_1)) &= 0 \\ &\vdots \\ h^{p-1}(\theta_1, \mathbf{x}(\theta_1)) &= 0, \end{aligned}$$

must be satisfied at the entry time θ_1 of a boundary interval, and p Lagrange multipliers $\pi^1, \ldots, \pi^p$ are thus associated to these constraints subsequently.

Then, assuming that the functions $\mathbf{f}$ and h are continuously differentiable with respect to all their arguments up to order $p-1$ and p, respectively, it can be shown that the necessary conditions of optimality given in Theorem 3.36 are modified as follows (see [25]):

- In (i) and (ii), the Lagrangian function $\mathcal{L}^1$ and the control region U^1 are substituted by $\mathcal{L}^p$ and U^p, respectively;
- In (iii), the condition (3.149) is replaced by

$$\eta^\star(t)\, h(t, \mathbf{x}^\star(t)) = 0; \qquad (-1)^i (\eta^\star)^{(i)}(t) \geq 0, \quad i = 0, \ldots, p; \tag{3.168}$$

- (iv) is changed to the requirement that at any entry time θ_1, the adjoint function and the Hamiltonian function may have discontinuities of the form

$$\boldsymbol{\lambda}^\star(\theta_1^-)^\mathsf{T} = \boldsymbol{\lambda}^\star(\theta_1^+)^\mathsf{T} + \sum_{i=1}^{p} \pi^{i\star}(\theta_1)\, h_{\mathbf{x}}^{i-1}(\theta_1, \mathbf{x}^\star(\theta_1)) \tag{3.169}$$

$$\mathcal{H}[\theta_1^-] = \mathcal{H}[\theta_1^+] - \sum_{i=1}^{p} \pi^{i\star}(\theta_1)\, h_t^{i-1}(\theta_1, \mathbf{x}^\star(\theta_1)), \tag{3.170}$$

where the Lagrange multipliers $\pi^{1\star}(\theta_1), \ldots, \pi^{p\star}(\theta_1)$ satisfy the conditions

$$\pi^{i\star}(\theta_1)\, h(\theta_1, \mathbf{x}^\star(\theta_1)) = 0, \qquad \pi^{i\star}(\theta_1) \geq 0, \quad i = 1, \ldots, p; \tag{3.171}$$

Moreover, at any contact time θ_c, (3.169), (3.170), and (3.171) hold with $\pi^i(\theta_c) = 0$, for $i \geq 2$, and we have the additional conditions

$$\pi^{i\star}(\theta_c) \left\{ \begin{array}{c} \geq \\ = \end{array} \right\} (-1)^{p-i} (\eta^\star)^{(p-k)}(\theta_c^+), \quad \text{for} \left\{ \begin{array}{l} i = 1, \\ i = 2, \ldots, p; \end{array} \right. \tag{3.172}$$

- (v) remains unchanged.

Note that all of the above conditions can be generalized readily to problems having multiple state inequality constraints $h_1(t, \mathbf{x}(t)) \leq 0, \ldots, h_{n_h}(t, \mathbf{x}(t)) \leq 0$, possibly of different orders $p_1, \ldots, p_{n_h}$.[11] In particular, these conditions remain valid in the first-order case $p_1 = \cdots = p_{n_h} = 1$, and thus encompass those given in Theorem 3.36.

Finally, we close the discussion on high-order inequality state constraints by reemphasizing the fact that such constraints give rise to highly complex, possibly ill-behaved, control problems. Similar to high-order singular control problems (e.g., the Fuller problem, see §3.5.4), an optimal control may exhibit a chattering behavior near high-order boundary arcs (either at the entry or at the exit point), i.e., the costate variables may have countably many jumps. An example of this behavior can be found in [43], for a problem having a third-order inequality state constraint.

3.6 NUMERICAL METHODS FOR OPTIMAL CONTROL PROBLEMS

Unless the system equations, along with the cost functional and the constraints, of the problem at hand are rather simple, numerical methods must be employed to solve optimal control problems. With the development of economical, high speed computers over the last few decades, it has become possible to solve complicated problems in a reasonable amount of time.

Presenting a survey of numerical methods in the field of optimal control is a daunting task. Perhaps the most difficult aspect is restricting the scope of the survey to permit a meaningful discussion within a few pages only. In this objective, we shall focus on two types of numerical methods, namely, *direct solution methods* (§3.6.3) and *indirect solution methods* (§3.6.2). The distinction between direct and indirect methods can be understood as follows. A direct method attempts to find a minimum to the objective function in the feasible set, by constructing a sequence of points converging to that minimum. In contrast, an indirect method attempts to find a minimum point 'indirectly', by solving the necessary conditions of optimality. For this reason, indirect methods are often referred to as *PMP-based methods* or *variational methods* in the literature. Other approaches not discussed herein include *dynamic programming methods* [18, 37] and *stochastic optimization methods* [5].

In many numerical methods, one needs to calculate the values of functionals subject to the differential equations. Moreover, since these functionals depend on parameters (e.g., as a result of the parameterization of the control trajectories), there is much interest in evaluating their gradients with respect to the parameters. Before presenting the numerical

[11] A complex example involving a state third-order inequality constraint, two first-order inequality constraints, and a control constraint can be found in [13].

methods for optimal control, we shall therefore give a close look to the evaluation of parameter-dependent functionals and their gradients in §3.6.1.

3.6.1 Evaluation of Parameter-Dependent Functionals and their Gradients

In this subsection, our focus is on a Mayer type functional $\mathcal{F}$ defined as

$$\mathcal{F}(\mathbf{p}) := \phi(\mathbf{x}(t_\mathrm{f}), \mathbf{p}), \tag{3.173}$$

where $\mathbf{p} \in P \subset \mathbb{R}^{n_p}$ is a vector of time-invariant parameters, and the state $\mathbf{x}(t) \in \mathbb{R}^{n_x}$ is described by a set of parametric ODEs

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}, \mathbf{p}), \quad t_0 \le t \le t_\mathrm{f}; \qquad \mathbf{x}(t_0) = \mathbf{h}(\mathbf{p}). \tag{3.174}$$

In what follows, ℓ, $\mathbf{f}$ and $\mathbf{h}$ are always assumed continuous in $(t, \mathbf{x}, \mathbf{p})$ with continuous first partial derivatives with respect to $(\mathbf{x}, \mathbf{p})$, for $(t, \mathbf{x}, \mathbf{p}) \in [t_0, t_\mathrm{f}] \times \mathbb{R}^{n_x} \times P$. Remind that both Lagrange and Bolza type functionals can be converted into the Mayer form by adding an extra state variable and differential equation to (3.174) that correspond to the integrand of the integral term, and then considering the value of this extra state at final time (see §3.2.3 on p. 108).

Assuming that a unique solution $\mathbf{x}(t; \overline{\mathbf{p}})$ to the system (3.174) exists for a given $\overline{\mathbf{p}} \in P$, we wish to calculate the value $\mathcal{F}(\overline{\mathbf{p}})$ as well as its gradient $\boldsymbol{\nabla}_\mathbf{p}\mathcal{F}(\overline{\mathbf{p}})$. Obtaining $\mathcal{F}(\overline{\mathbf{p}})$ requires that the IVP (3.174) be numerically integrated, and a brief overview of numerical methods for IVPs in ODEs is thus given in §3.6.1.1. On the other hand, the computation of $\boldsymbol{\nabla}_\mathbf{p}\mathcal{F}(\overline{\mathbf{p}})$ is less straightforward, and we shall present three methods for doing this: the finite differences approach (§3.6.1.2), the sensitivity approach (§3.6.1.3), and the adjoint approach (§3.6.1.4).

3.6.1.1 Initial Value Problems

The problem of evaluating the functional $\mathcal{F}$ for given values $\overline{\mathbf{p}} \in P$ of the parameters consists of computing the value of $\mathbf{x}(t_\mathrm{f})$ that satisfies

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}, \overline{\mathbf{p}}),$$

for each $t_0 \le t \le t_\mathrm{f}$, from the initial value $\mathbf{x}(t_0) = \mathbf{h}(\overline{\mathbf{p}})$. Numerical methods for solving the foregoing IVP are relatively mature in comparison to other fields in numerical optimization. Any numerical methods for ODEs generate an approximate solution step-by-step in discrete increments across the interval of integration, in effect producing a discrete sample of approximate values $\mathbf{x}^i$ of the solution function $\mathbf{x}(t)$. Most schemes can be classified as either *one-step* or *multi-step* methods.[12]

One-Step Methods. A popular family of one-step methods is the Runge-Kutta (RK) schemes. Given an estimate $\mathbf{x}^i$ of the states at time t_i, an new estimate $\mathbf{x}^{i+1}$ at time $t_{i+1} := t_i + h_i$ is obtained as

$$\mathbf{x}^{i+1} = \mathbf{x}^i + \sum_{j=1}^{K} \omega_j \mathbf{f}_{ij},$$

where

$$\mathbf{f}_{ij} := \mathbf{f}\left(t_i + h_i\tau_i, \mathbf{x}^i + h_i \textstyle\sum_{k=1}^{K} \alpha_{jk}\mathbf{f}_{ik}, \overline{\mathbf{p}}\right), \quad 1 \le j \le K,$$

[12] An excellent website illustrating numerical methods for solving IVPs in ODEs can be found at `http://www.cse.uiuc.edu/eot/modules/ode/`.

with $0 \leq \tau_1 \leq \cdots \leq \tau_K \leq 1$, and $K \geq 1$ denotes the number of *stages* in the scheme. RK schemes differ in the choice of the parameters ω_i, τ_i, and α_{ij}, which are most conveniently represented in the so-called *Butcher diagram*:

$$\begin{array}{c|ccc} \omega_1 & \alpha_{11} & \cdots & \alpha_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_K & \alpha_{K1} & \cdots & \alpha_{KK} \\ \hline & \beta_1 & \cdots & \beta_K \end{array}$$

RK schemes are said to be *explicit* if the Butcher diagram is such that $\alpha_{jk} = 0$ for $j \leq k$), and *implicit* otherwise. Three common examples of RK schemes are the following:

Euler's Explicit: $K = 1$

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

Classical Runge-Kutta Explicit: $K = 4$

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

Trapezoidal Implicit: $K = 2$

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

An obvious appeal of an explicit scheme is that the computation of each integration step can be performed without iteration; that is, given the value of $\mathbf{x}^i$ at time t_i, the value of $\mathbf{x}^{i+1}$ at the next time t_{i+1} follows directly from available values of $\mathbf{f}$. In contrast, for an implicit scheme, the unknown value $\mathbf{x}^{i+1}$ appears nonlinearly, e.g., the Hermite-Simpson implicit method requires

$$\mathbf{F}^i := \mathbf{x}^{i+1} - \mathbf{x}^i - \frac{h_i}{2}\left[\mathbf{f}(t_{i+1}, \mathbf{x}^{i+1}, \overline{\mathbf{p}}) + \mathbf{f}(t_i, \mathbf{x}^i, \overline{\mathbf{p}})\right] = \mathbf{0}. \tag{3.175}$$

Computing $\mathbf{x}^{i+1}$ from given values of t_i, t_{i+1} and $\mathbf{x}^i$ thus requires solving the nonlinear expression (3.175) to drive the defect $\mathbf{F}_i$ to zero. The iterations required to solve this equation are called the *corrector iterations*. An initial guess to begin the iteration is usually provided by a so-called *predictor step*. There is considerable latitude in the choice of predictor and corrector schemes. For some well-behaved differential equations, a single predictor/corrector step is required. On the other hand, it may be necessary to perform multiple corrector iterations when the differential equations are stiff;[13] this is generally done based on Newton's method (see §1.8.2, p. 34).

Linear Multi-step Methods. The general form of a K-step linear multi-step method is given by

$$\mathbf{x}^{i+1} = \sum_{j=1}^{K} \alpha_j \mathbf{x}^{i-j+1} + h \sum_{j=0}^{K} \beta_j \mathbf{f}^{i-j+1}, \tag{3.176}$$

[13] An ODE whose solutions decay rapidly towards a common, slowly-varying solution is said to be *stiff*. Explicit methods are generally inefficient for solving stiff ODEs because their stability region is relatively small, which forces the step size to be much smaller than that required to achieve the desired accuracy. Implicit methods require more work per step, but their significantly larger stability regions permit much larger steps to be taken, so they are often much more efficient than explicit methods of comparable accuracy for solving stiff ODEs. (A numerical method is said to be *stable* if small perturbations do not cause the resulting numerical solutions to diverge without bound.)

where α_j and β_j are specified constants, $\mathbf{x}^i$ is the approximate solution at time t_i, and $\mathbf{f}^i := \mathbf{f}(t_i, \mathbf{x}^i, \overline{\mathbf{p}})$. If $\beta_N = 0$, the method is *explicit*, otherwise it is *implicit*. Note that the K-past integration steps are assumed to be equally spaced.

The most popular linear multi-step methods are based on polynomial interpolation, and even methods which are not based on interpolation use interpolation for such purposes as changing the step size. These methods come in families. Particularly popular for non-stiff problems is the *Adams family*, and for stiff problems, the *backward differentiation formula (BDF) family*.

- In a K-step Adams-Bashforth method, the solution is advanced at each step by integrating the interpolant of the derivative values at K previous solution points. Specifically, for approximate solution points $(t_{i-K+1}, \mathbf{x}^{i-K+1}), \ldots, (t_i, \mathbf{x}^i)$, the approximate solution value $\mathbf{x}^{i+1}$ at time $t_{i+1} = t_i + h$ is given by

$$\mathbf{x}(t_{i+1}) = \mathbf{x}(t_i) + \int_{t_{i+1}}^{t_i} \mathbf{F}(t)\, \mathrm{d}t,$$

  where $\mathbf{F}(t)$ is the unique polynomial of degree $K-1$ interpolating $\mathbf{f}(t_{i-K+1}, \mathbf{x}^{i-K+1}, \overline{\mathbf{p}}), \ldots, (t_i, \mathbf{x}^i, \overline{\mathbf{p}})$. That is, we have $\alpha_1 = 1$, and $\alpha_j = 0$ for $j > 1$, in the general form (3.176). A *K-step Adams-Moulton method* is derived similarly to a Adams-Bashforth method, except that it interpolates $\mathbf{f}$ at the unknown value t_{i+1} as well.

- In a K-step BDF method, the solution is advanced at each step by interpolating K previous solution points along with the (as yet unknown) new solution point, differentiating that interpolant, and requiring the derivative to match the ODE at the new point. Specifically, for approximate solution points $(t_{i-K+1}, \mathbf{x}^{i-K+1}), \ldots, (t_i, \mathbf{x}^i)$, the approximate solution value $\mathbf{x}^{i+1}$ at time $t_{i+1} = t_i + h$ is determined by solving the implicit equation

$$\dot{\mathbf{X}}(t_{i+1}) = \mathbf{f}(t_{i+1}, \mathbf{x}^{i+1}, \overline{\mathbf{p}})$$

  for $\mathbf{x}^{i+1}$, where $\mathbf{X}(t)$ is the unique polynomial of degree K that interpolates $(t_{i-K+1}, \mathbf{x}^{i-K+1}), \ldots, (t_i, \mathbf{x}^i), (t_{i+1}, \mathbf{x}^{i+1})$. Hence, we have $\beta_0 \neq 0$, and $\beta_j = 0$ for $j > 1$, in the general form (3.176). Note that the simplest member of this family is the implicit Euler method (i.e., $\alpha_1 = 1$ and $\beta_0 = 1$):

$$\mathbf{x}^{i+1} = \mathbf{x}^i + h\mathbf{f}^{i+1}.$$

  BDF methods have relatively large stability regions, so they are particularly suitable for solving stiff ODEs.

For a K-step method, the method is applied for $i \geq K-1$, and K initial values $\mathbf{x}^0, \ldots, \mathbf{x}^{K-1}$ are needed to start it up. A usual strategy is at a starting point is to gradually increase the method's number of steps, starting from $K = 1$. Another approach consists of using an appropriate RK method.

Differential-Algebraic Equations. Up to this point, the prototypical IVP (3.174) refers to an *explicit ODE system*,

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{p}). \tag{3.177}$$

However, a more general formulation for an ODE system is the so-called *implicit* form,

$$\mathbf{F}(t, \mathbf{x}(t), \dot{\mathbf{x}}(t), \mathbf{p}) = \mathbf{0},$$

where the Jacobian matrix $\mathbf{F}_{\dot{\mathbf{x}}}$ is assumed to be nonsingular for all argument values in an appropriate domain. In principle, it is often possible to solve for $\dot{\mathbf{x}}$ in terms of t, $\mathbf{x}$ and $\mathbf{p}$, obtaining the explicit form (3.177). However, this transformation may not always be easy or cheap to realize. Also, in general, there may be additional questions of existence and uniqueness of the solutions.

Another extension of explicit ODEs is in systems of the form

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{y}(t), \mathbf{p}) \tag{3.178}$$

$$\mathbf{0} = \mathbf{g}(t, \mathbf{x}(t), \mathbf{y}(t), \mathbf{p}), \tag{3.179}$$

where the ODEs (3.178) now depend on additional variables $\mathbf{y}(t)$, and the pair $(\mathbf{x}(t), \mathbf{y}(t))$ is forced to satisfy the algebraic constraints (3.179). Such systems are called *differential algebraic equations (DAEs) in semi-explicit form*. More generally, DAEs can be specified in *fully-implicit form*,

$$\mathbf{F}(t, \mathbf{z}(t), \dot{\mathbf{z}}(t), \mathbf{p}) = \mathbf{0}, \tag{3.180}$$

with the new variable $\mathbf{z}^\mathsf{T} := (\mathbf{x}^\mathsf{T}, \mathbf{y}^\mathsf{T})$, and where the Jacobian matrix $\mathbf{F}_{\dot{\mathbf{z}}}$ is now singular.

Note that the general theory for DAEs is much more recent and less developed than for ODEs, and it is still subject to intense research activity [11]. Since a DAE involves a mixture of differential and algebraic equations, one may hope that applying analytical time differentiations to a given system and eliminating, as needed, repeatedly if necessary, will yield an explicit ODE system for all the unknown variables. This turns out to be the case in most situations (unless the problem is singular). In particular, the number of differentiations needed for this transformation is called the *index* of the DAE system.[14] According to this definition, ODEs have index 0. An index-2 DAE system is illustrated in Example 3.39 below.

Example 3.39 (Index-2 DAE System). Consider the DAE system

$$\begin{aligned} z_1(t) - q(t) &= 0 \\ \dot{z}_1(t) - z_2(t) &= 0. \end{aligned}$$

where $q(t)$ is a given smooth function. Differentiating the first equation gives

$$z_2(t) = \dot{z}_1(t) = \dot{q}(t),$$

and, by differentiating the resulting equation, we then get

$$\dot{z}_2(t) = \ddot{z}_1(t) = \ddot{q}(t).$$

Hence, the index is 2 since two rounds of differentiation were needed.

[14] Formally, the index of a DAE system of the form (3.180) is defined as the *minimum number of times that part or all of the equations must be differentiated in order to obtain an ODE system.*

The index of a DAE system is closely connected to the question of initial conditions specification. While n initial conditions must be given to fully specify the solution of an ODE system of size n, a DAE system of size n will in general have m degrees of freedom, with m being anywhere between 0 and n. Which m pieces of information are needed to determine the DAE solution may be a difficult issue, or at least not immediately obvious. In other words, one must specify *consistent* initial conditions in the sense that the constraints of the system must be satisfied. To illustrate it, consider the DAE system in Example 3.39. The variable z_1 at initial time must satisfy the obvious constraint $z_1(t) = q(t)$, but there is also a *hidden constraints* $z_2(t) = \dot{q}(t)$ which the solution must satisfy at any time. Therefore, the only possible consistent initial conditions are $z_1(t_0) = q(t_0)$, $z_2(t_0) = \dot{q}(t_0)$, i.e., the DAE system has zero degree of freedom!

The special case of semi-explicit DAEs (3.178,3.179) is encountered in many practical problems. It is readily shown that a sufficient condition for semi-explicit DAEs to have index 1, is that the Jacobian matrix $\mathbf{g_y}$ be nonsingular. In the index-1 case, one can then distinguish between the *differential variables* $\mathbf{x}(t)$ and the *algebraic variables* $\mathbf{y}(t)$.

Remark 3.40 (Link between DAEs and the Euler-Lagrange Equations). The Euler-Lagrange equations (3.13–3.15), which are part of the necessary conditions of optimality for optimal control problems, are DAEs in semi-explicit form. Provided that the Hamiltonian function is nonsingular (i.e., $\mathcal{H}_{\mathbf{uu}}$ is not trivially equal to zero, see §3.5.4), these equations have index 1, and the differential variables are thus clearly identified to be the states and adjoints, whereas the algebraic variables correspond to the controls. But if the Hamiltonian function is now singular, the Euler-Lagrange equations have high index (≥ 2), which implies that the problem contains hidden constraints. These extra constraints correspond precisely to the equations defining the singular surface, i.e.,

$$\mathcal{H}_u = 0, \quad \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}_u = 0, \quad \ldots, \quad \frac{\mathrm{d}^{2p}}{\mathrm{d}t^{2p}}\mathcal{H}_u = 0,$$

where p denotes the order of singularity. Obviously, there exists strong connections between high-index Euler-Lagrange equations and singularity optimal control problems. The situation is similar for optimal control problems with high-order state inequality constraints (see §3.5.6).

Numerical methods for solving DAEs are mostly limited to index-1 systems. Fortunately, this is the case for many practical systems. For higher-index systems to be handled, it is necessary to first transform the DAEs into index-1 form (e.g., by applying successive differentiation), before a solution can be computed. The first general technique for solving fully implicit index-1 DAEs was proposed by Gear in 1971, It utilizes BDF methods similar to those used for ODEs, i.e., the derivative $\dot{\mathbf{z}}$ is replaced by the derivative of the polynomial interpolating the solution computed over the preceding K-steps along with the (as yet unknown) new solution point. The simplest example of a BDF method is the implicit Euler method that replaces (3.180) with

$$\mathbf{F}\left(t_{i+1}, \mathbf{z}^{i+1}, \frac{\mathbf{z}^{i+1} - \mathbf{z}^{i}}{h}, \overline{\mathbf{p}}\right) = \mathbf{0}.$$

The resulting nonlinear system in the variable $\mathbf{z}^{i+1}$ is usually solved by some form of Newton's method at each time step.

Implementation and Software. Regardless of whether a one-step or multi-step method is utilized, a successful implementation must address the accuracy of the solution. How

well does the discrete solution $\mathbf{x}^i$ for $i = 0, 1, \ldots$, produced by the integration scheme agree with the true solution $\mathbf{x}(t)$? All well-implemented integration schemes incorporate some mechanism for adjusting the integration step-size and/or the order of the method to control the integration error.[15]

A variety of excellent and widely used software for IVPs is readily available. Major differences between codes lie in the numerical scheme used, whether index-1 DAEs can be accommodated, and whether sparse systems can be handled efficiently (especially useful for large-scale systems). A number of free numerical integration codes is given in Tab. 3.1.

Table 3.1. Popular codes doing numerical integration of ODEs/DAEs.

Solver	Website	Lic.	Characteristics
DASSL	`http://www.engineering.ucsb.edu/~cse/software.html`	free	BDF schemes, ODEs and index-1 DAEs, dense or banded Jacobian
DASPK2.0	`http://www.engineering.ucsb.edu/~cse/software.html`	free	same as DASSL, designed for sparse, large-scale systems
CVODE	`http://acts.nersc.gov/sundials/`	free	Adams-Moulton and BDF schemes, ODEs only, designed for dense or sparse, large-scale systems
IDA	`http://acts.nersc.gov/sundials/`	free	BDF schemes, ODE and index-1 DAE problems, consistent initialization

The codes listed in Tab. 3.1. are stand-alone (either in C or fortran77). Moreover, various integration schemes are available in the MATLAB® environment, both for ODEs (Runge-Kutta, Adams, BDF schemes) and index-1 DAEs (BDF schemes).

3.6.1.2 Gradients via Finite Differences We now turn to the problem of calculating the gradient $\boldsymbol{\nabla}_{\mathbf{p}}\mathcal{F}(\overline{\mathbf{p}})$ of the functional (3.173), subject to the IVP in ODEs (3.174). The easiest way of getting an estimate of $\boldsymbol{\nabla}_{\mathbf{p}}\mathcal{F}(\overline{\mathbf{p}})$ is to consider a forward difference approximation so that:

$$\nabla_{p_j}\mathcal{F}(\overline{\mathbf{p}}) \approx \frac{\mathcal{F}(\overline{p}_1, \ldots, \overline{p}_j + \delta p_j, \ldots, \overline{p}_{n_p}) - \mathcal{F}(\overline{\mathbf{p}})}{\delta p_j}, \tag{3.181}$$

for each $j = 1, \ldots, n_p$. In practice, the variations δp_j can be chosen as

$$\delta p_j = \epsilon^a + \overline{p}_j \epsilon^r,$$

where ϵ^a and ϵ^r are small absolute and relative perturbations parameters, respectively; often, ϵ^a and ϵ^r are chosen as the square-roots of the absolute and relative tolerances used in the numerical integration code.

A practical procedure for calculating both the value and the gradient of $\mathcal{F}$ is as follows:

Initial Step

Integrate the ODEs (3.174) once with the actual parameter values $\overline{\mathbf{p}}$;

Calculate the value of $\mathcal{F}(\overline{\mathbf{p}})$.

Loop: $j = 1, \ldots, n_p$

[15] It should be stressed that error control mechanism in numerical integration schemes is inherently local, i.e., it can only guarantee the accuracy of the solution from one step to the next. That is, if the numerical scheme takes many integration steps, the error accumulates and there is no guarantee that the discrete solution will be a good approximation or even close to the desired solution in the end.

Set $p_i := \overline{p}_i$, $i \neq j$; $p_j := \overline{p}_j + \delta p_j$;
Integrate the ODEs (3.174) once with the perturbed parameter values $\mathbf{p}$;
Calculate the value of $\mathcal{F}(\mathbf{p})$, and

$$\nabla_{p_j}\mathcal{F}(\overline{\mathbf{p}}) \approx \frac{\mathcal{F}(\mathbf{p}) - \mathcal{F}(\overline{\mathbf{p}})}{\delta p_j}.$$

End Loop

Observe that a gradient calculation with the forward difference approach requires $n_p + 1$ integrations of the ODEs (3.174), which can be computationally expensive when n_p is large. On the other hand, however, this approach does not bring any additional complexity other than integrating the ODEs for the system, and applies readily to the case of (index-1) DAEs. A more accurate approximation (2nd-order approximation) can be obtained by using centered finite differences as

$$\nabla_{p_j}\mathcal{F}(\overline{\mathbf{p}}) \approx \frac{\mathcal{F}(\overline{p}_1, \ldots, \overline{p}_j + \delta p_j, \ldots, \overline{p}_{n_p}) - \mathcal{F}(\overline{p}_1, \ldots, \overline{p}_j - \delta p_j, \ldots, \overline{p}_{n_p})}{2\delta p_j}, \quad (3.182)$$

although this requires 2 functional evaluations per parameter, i.e., an overall $2 \times n_p$ ODEs integrations for a single gradient evaluation.

A major limitation of the finite differences approach lies in its accuracy. It is easy to see why the difference formulas (3.181) do not provide accurate values. If δp_j is small, then cancellation error reduces the number of significant figures in the gradient estimate, especially when the function values are obtained with limited accuracy from a numerical integration code. On the other hand, if δp_j is not small, then truncation errors (i.e., higher-order terms) become significant. Even if δp_j is optimally chosen, it is well known that $\boldsymbol{\nabla}_{\mathbf{p}}\mathcal{F}(\overline{\mathbf{p}})$ will be accurate to only about $\frac{1}{2}$ of the significant digits of $\mathcal{F}(\overline{\mathbf{p}})$ (or $\frac{2}{3}$ if the centered formula (3.182) is used). This motivates the forward and adjoint (reverse) sensitivity approaches of gradient calculation presented subsequently.

3.6.1.3 Gradients via Forward Sensitivity Analysis Consider the IVP in ODEs (3.174) for given parameter values $\overline{\mathbf{p}} \in P$,

$$\dot{\mathbf{x}}(t; \overline{\mathbf{p}}) = \mathbf{f}(t, \mathbf{x}(t; \overline{\mathbf{p}}), \overline{\mathbf{p}}); \qquad \mathbf{x}(t_0; \overline{\mathbf{p}}) = \mathbf{h}(\overline{\mathbf{p}}). \quad (3.183)$$

and suppose that (3.183) has a unique solution $\mathbf{x}(t; \overline{\mathbf{p}})$, $t_0 \leq t \leq t_\mathrm{f}$. The functions $\mathbf{f}$ and $\mathbf{h}$ being continuously differentiable with respect to $(\mathbf{x}, \mathbf{p})$ and $\mathbf{p}$, respectively, the solution $\mathbf{x}(t; \mathbf{p})$ of (3.174) is itself continuously differentiable with respect to $\mathbf{p}$ in a neighborhood of $\overline{\mathbf{p}}$, at each $t \in [t_0, t_\mathrm{f}]$ (see Appendix A.5.3). In particular, the first-order state sensitivity functions $\mathbf{x}_{p_j}(t; \overline{\mathbf{p}})$, $j = 1, \ldots, n_p$, are given by (A.15), or equivalently,

$$\dot{\mathbf{x}}_{p_j}(t; \overline{\mathbf{p}}) = \mathbf{f}_{\mathbf{x}}(t, \mathbf{x}(t; \overline{\mathbf{p}}), \overline{\mathbf{p}})\, \mathbf{x}_{p_j}(t; \overline{\mathbf{p}}) + \mathbf{f}_{p_j}(t, \mathbf{x}(t; \overline{\mathbf{p}}), \overline{\mathbf{p}}); \qquad \mathbf{x}_{p_j}(t_0; \overline{\mathbf{p}}) = \mathbf{h}_{p_j}(\overline{\mathbf{p}}). \quad (3.184)$$

The foregoing equations are called the *sensitivity equations with respect to parameter* p_j; in general, they are linear, non-homogeneous differential equations, and become homogeneous in the case where $\mathbf{f}_{p_j} = \mathbf{0}$.

Once the state sensitivity functions are known at $t = t_\mathrm{f}$, and since ϕ is continuously differentiable with respect to $\mathbf{x}$ and $\mathbf{p}$, the gradient $\boldsymbol{\nabla}_{\mathbf{p}}\mathcal{F}(\overline{\mathbf{p}})$ of the functional (3.173) at $\overline{\mathbf{p}}$ can be calculated as

$$\nabla_{p_j}\mathcal{F}(\overline{\mathbf{p}}) = \phi_{\mathbf{x}}(\mathbf{x}(t_\mathrm{f}; \overline{\mathbf{p}}), \overline{\mathbf{p}})^{\mathsf{T}} \mathbf{x}_{p_j}(t_\mathrm{f}; \overline{\mathbf{p}}) + \phi_{p_j}(\mathbf{x}(t_\mathrm{f}; \overline{\mathbf{p}}), \overline{\mathbf{p}}), \quad (3.185)$$

for each $j = 1, \ldots, n_p$.

A practical procedure calculating both the value and the gradient of $\mathcal{F}$ at $\overline{\mathbf{p}}$ is as follows:

State and Sensitivity Numerical Integration: $t_0 \rightarrow t_\mathrm{f}$

$$\begin{aligned}
\dot{\mathbf{x}}(t) &= \mathbf{f}(t, \mathbf{x}(t), \overline{\mathbf{p}}); & \mathbf{x}(t_0) &= \mathbf{h}(\overline{\mathbf{p}}) \\
\dot{\mathbf{x}}_{p_1}(t) &= \mathbf{f}_{\mathbf{x}}(t, \mathbf{x}(t), \overline{\mathbf{p}})\, \mathbf{x}_{p_1}(t) + \mathbf{f}_{p_1}(t, \mathbf{x}(t), \overline{\mathbf{p}}); & \mathbf{x}_{p_1}(t_0) &= \mathbf{h}_{p_1}(\overline{\mathbf{p}}) \\
&\vdots & &\vdots \\
\dot{\mathbf{x}}_{p_{n_p}}(t) &= \mathbf{f}_{\mathbf{x}}(t, \mathbf{x}(t), \overline{\mathbf{p}})\, \mathbf{x}_{p_{n_p}}(t) + \mathbf{f}_{p_{n_p}}(t, \mathbf{x}(t), \overline{\mathbf{p}}); & \mathbf{x}_{p_{n_p}}(t_0) &= \mathbf{h}_{p_{n_p}}(\overline{\mathbf{p}})
\end{aligned}$$

Function and Gradient Evaluation:

$$\begin{aligned}
\mathcal{F}(\mathbf{p}) &= \phi(\mathbf{x}(t_\mathrm{f}), \overline{\mathbf{p}}) \\
\nabla_{p_1}\mathcal{F}(\overline{\mathbf{p}}) &= \phi_{\mathbf{x}}(\mathbf{x}(t_\mathrm{f}), \overline{\mathbf{p}})^{\mathsf{T}} \mathbf{x}_{p_1}(t_\mathrm{f}) + \phi_{p_1}(\mathbf{x}(t_\mathrm{f}), \overline{\mathbf{p}}) \\
&\vdots \\
\nabla_{p_{n_p}}\mathcal{F}(\overline{\mathbf{p}}) &= \phi_{\mathbf{x}}(\mathbf{x}(t_\mathrm{f}), \overline{\mathbf{p}})^{\mathsf{T}} \mathbf{x}_{p_{n_p}}(t_\mathrm{f}) + \phi_{p_{n_p}}(\mathbf{x}(t_\mathrm{f}), \overline{\mathbf{p}})
\end{aligned}$$

Observe that the state and state sensitivity equations are solved simultaneously, so that a local error control can be performed on both the state and state sensitivity variables. However, the size of the state/sensitivity system grows proportionally to the number of states and parameters as $(n_x + 1) \times n_p$, which can lead to computationally intractable problems if both n_x and n_p are large. In response to this, effective methods have been developed in recent years that take advantage of the special structure of the problem. These methods are usually based on implicit multi-step integration schemes (see §3.6.1.1, 163), and exploit the fact that the sensitivity equations (3.184) are linear and all share the same Jacobian matrix with the original system (3.183). Three well-established methods, which differ in the way the corrector formula is solved while sharing the same predictor step, are the following:

- *Staggered direct methods* [31]: At each time step, the states are computed first by the nonlinear corrector step, and the state sensitivities are then obtained by solving a linear system. This method is sometimes considered to be inefficient because it requires that the Jacobian matrix be evaluated and factored[16] at every time step.
- *Simultaneous corrector methods* [39]: The state and sensitivity variables are computed simultaneously by the nonlinear corrector step. This method is more efficient because it evaluates and factors the Jacobian matrix only when necessary.
- *Staggered corrector method* [20]: This method is similar to the staggered direct method, except that it uses the factorization of the Jacobian matrix at some past step to solve the linear sensitivity system. This saves on the number of factorizations of the Jacobian matrix, which can be the most expensive part of the computations for systems having many state variables but relatively few parameters.

[16] Here, we refer to the LU factorization of the Jacobian matrix, which is used in the corrector step for calculating the inverse of the Jacobian matrix

Note that all these methods apply equally well for fully implicit, index-1 DAE systems of the form (3.180). In this case the sensitivity DAEs read:

$$\mathbf{F}_{\mathbf{x}}\, \mathbf{x}_{p_j} + \mathbf{F}_{\dot{\mathbf{x}}}\, \dot{\mathbf{x}}_{p_j} + \mathbf{F}_{p_j} = \mathbf{0}, \tag{3.186}$$

for each $j = 1, \ldots, n_p$. Obviously, consistent initial state sensitivities must be specified to the sensitivity DAEs; they are obtained upon direct differentiation of the initial conditions of the original DAEs.

A variety of excellent and widely used codes is available for forward sensitivity analysis of IVPs in ODEs and DAEs, such as those listed in Tab. 3.2. hereafter. Note that the version 7.4 of MATLAB® does not have a function doing forward sensitivity analysis.

Table 3.2. Popular codes doing forward sensitivity analysis of ODEs/DAEs.

Solver	Website	Lic.	Characteristics
DSL48S	`http://yoric.mit.edu/dsl48s.html`	free for acad.	based on DASSL[a], ODEs and index-1 DAEs, sparse Jacobian
DASPK3.0	`http://www.engineering.ucsb.edu/~cse/software.html`	free for acad.	based on DASSL[a], ODEs, index-1 DAEs and Hessenberg index-2 DAEs
CVODES	`http://acts.nersc.gov/sundials/`	free	based on CVODE[a], ODEs only

[a] See Tab. 3.1.

3.6.1.4 Gradients via Adjoint Sensitivity Analysis Some problems require the gradient of a functional with respect to a large number of parameters. For these problems, particularly if the number of state variables is also large, both the finite differences approach (§3.6.1.2) and the forward sensitivity approach (§3.6.1.3) are intractable. This motivates the third class of methods, namely *adjoint sensitivity analysis* (also called *reverse sensitivity analysis*), for computing the gradient of a functional [19].

Consider the functional (3.173), subject to the IVP in ODEs (3.174), at a point $\overline{\mathbf{p}} \in P$. Analogous to §3.6.1.3, we shall suppose that (3.183) has a unique solution $\mathbf{x}(t; \overline{\mathbf{p}})$, $t_0 \leq t \leq t_\mathrm{f}$. Then, adjoining the differential equations to the functional using smooth multiplier functions $\boldsymbol{\lambda} \in \mathcal{C}^1[t_0, t_\mathrm{f}]^{n_x}$, we form the augmented functional

$$\tilde{\mathcal{F}}(\overline{\mathbf{p}}) := \phi(\mathbf{x}(t_\mathrm{f}; \overline{\mathbf{p}}), \overline{\mathbf{p}}) + \int_{t_0}^{t_\mathrm{f}} \boldsymbol{\lambda}(t)^\mathsf{T} \left[\mathbf{f}(t, \mathbf{x}(t; \overline{\mathbf{p}}), \overline{\mathbf{p}}) - \dot{\mathbf{x}}(t, \overline{\mathbf{p}})\right] \, \mathrm{d}t.$$

Since $\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}, \overline{\mathbf{p}})$ at each $t \in [t_0, t_\mathrm{f}]$, the gradient $\boldsymbol{\nabla}_{p_j} \mathcal{F}(\overline{\mathbf{p}})$, $i = 1, \ldots, n_p$, is obtained by applying the chain rule of differentiation:

$$\begin{aligned} \nabla_{p_j} \mathcal{F}(\overline{\mathbf{p}}) &= \nabla_{p_j} \tilde{\mathcal{F}}(\overline{\mathbf{p}}) && (3.187)\\ &= \phi_{p_j}(\mathbf{x}(t_\mathrm{f}), \overline{\mathbf{p}}) + \phi_{\mathbf{x}}(\mathbf{x}(t_\mathrm{f}), \overline{\mathbf{p}})^\mathsf{T} \mathbf{x}_{p_j}(t_\mathrm{f}) && (3.188)\\ &\quad + \int_{t_0}^{t_\mathrm{f}} \boldsymbol{\lambda}(t)^\mathsf{T} \left[\mathbf{f}_{p_j}(t, \mathbf{x}(t), \overline{\mathbf{p}}) + \mathbf{f}_{\mathbf{x}}(t, \mathbf{x}(t), \overline{\mathbf{p}})\, \mathbf{x}_{p_j}(t) - \dot{\mathbf{x}}_{p_j}(t)\right] \, \mathrm{d}t. \end{aligned}$$

By integration by parts, we have

$$\int_{t_0}^{t_\mathrm{f}} \boldsymbol{\lambda}(t)^\mathsf{T} \dot{\mathbf{x}}_{p_j}(t) \, \mathrm{d}t = \left[\boldsymbol{\lambda}(t)^\mathsf{T} \mathbf{x}_{p_j}(t)\right]_{t_0}^{t_\mathrm{f}} - \int_{t_0}^{t_\mathrm{f}} \dot{\boldsymbol{\lambda}}(t)^\mathsf{T} \mathbf{x}_{p_j}(t) \, \mathrm{d}t.$$

Thus, (3.188) becomes

$$\nabla_{p_j}\mathcal{F}(\overline{\mathbf{p}}) = \phi_{p_j}(\mathbf{x}(t_\mathrm{f}),\overline{\mathbf{p}}) + \boldsymbol{\lambda}(t_0)^\mathsf{T}\mathbf{h}_{p_j}(\overline{\mathbf{p}}) + [\phi_{\mathbf{x}}(\mathbf{x}(t_\mathrm{f}),\overline{\mathbf{p}}) - \boldsymbol{\lambda}(t_\mathrm{f})]^\mathsf{T}\mathbf{x}_{p_j}(t_\mathrm{f}) + \int_{t_0}^{t_\mathrm{f}} \left[\boldsymbol{\lambda}(t)^\mathsf{T}\mathbf{f}_{p_j}(t,\mathbf{x}(t),\overline{\mathbf{p}}) + \left[\mathbf{f}_{\mathbf{x}}(t,\mathbf{x}(t),\overline{\mathbf{p}})^\mathsf{T}\boldsymbol{\lambda}(t) + \dot{\boldsymbol{\lambda}}(t)\right]^\mathsf{T}\mathbf{x}_{p_j}(t)\right]\,\mathrm{d}t.$$

The foregoing expression being verified for any smooth function $\boldsymbol{\lambda}(\cdot)$, it is convenient to choose $\boldsymbol{\lambda}(\cdot)$ so as to eliminate the terms depending on the sensitivity variables $\mathbf{x}_{p_j}$:

$$\dot{\boldsymbol{\lambda}}^\star(t) = -\mathbf{f}_{\mathbf{x}}(t,\mathbf{x}(t),\overline{\mathbf{p}})^\mathsf{T}\boldsymbol{\lambda}^\star(t); \qquad \boldsymbol{\lambda}^\star(t_\mathrm{f}) = \phi_{\mathbf{x}}(\mathbf{x}(t_\mathrm{f}),\overline{\mathbf{p}}),$$

from which we obtain

$$\nabla_{p_j}\mathcal{F}(\overline{\mathbf{p}}) = \phi_{p_j}(\mathbf{x}(t_\mathrm{f}),\overline{\mathbf{p}}) + \boldsymbol{\lambda}^\star(t_0)^\mathsf{T}\mathbf{h}_{p_j}(\overline{\mathbf{p}}) + \int_{t_0}^{t_\mathrm{f}} \boldsymbol{\lambda}^\star(t)^\mathsf{T}\mathbf{f}_{p_j}(t,\mathbf{x}(t),\overline{\mathbf{p}})\,\mathrm{d}t. \quad (3.189)$$

for each $j = 1,\ldots,n_p$. Interestingly, the expression of $\nabla_{p_j}\mathcal{F}(\overline{\mathbf{p}})$ can be seen as the sum of 3 contributions: (i) the direct influence of the parameter p_j on the cost functional; (ii) the influence of p_j through the initial conditions; and (iii) the influence of p_j through the system dynamics.

A practical procedure for calculating both the value and the gradient of $\mathcal{F}$ at $\overline{\mathbf{p}}$ is as follows:

State Numerical Integration: $t_0 \rightarrow t_\mathrm{f}$

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t,\mathbf{x}(t),\overline{\mathbf{p}}); \qquad \mathbf{x}(t_0) = \mathbf{h}(\overline{\mathbf{p}});$$

Store the state values $\mathbf{x}(t)$ at mesh points, $t_0 < t_1 < t_2 < \cdots < t_M = t_\mathrm{f}$.

Adjoint Numerical Integration: $t_\mathrm{f} \rightarrow t_0$

$$\begin{aligned}
\dot{\boldsymbol{\lambda}}(t) &= -\mathbf{f}_{\mathbf{x}}(t,\mathbf{x}(t),\overline{\mathbf{p}})^\mathsf{T}\boldsymbol{\lambda}(t); & \boldsymbol{\lambda}(t_\mathrm{f}) &= \phi_{\mathbf{x}}(\mathbf{x}(t_\mathrm{f}),\overline{\mathbf{p}})\\
\dot{q}_1(t) &= -\mathbf{f}_{p_1}(t,\mathbf{x}(t),\overline{\mathbf{p}})^\mathsf{T}\boldsymbol{\lambda}(t); & q_1(t_\mathrm{f}) &= 0\\
&\vdots\\
\dot{q}_{n_p}(t) &= -\mathbf{f}_{p_{n_p}}(t,\mathbf{x}(t),\overline{\mathbf{p}})^\mathsf{T}\boldsymbol{\lambda}(t); & q_{n_p}(t_\mathrm{f}) &= 0;
\end{aligned}$$

Evaluate the right-hand-sides of the adjoint equations by interpolating the state values $\mathbf{x}(t)$, $t_k \le t \le t_{k+1}$, between mesh points $\mathbf{x}(t_k)$ and $\mathbf{x}(t_{k+1})$.

Function and Gradient Evaluation:

$$\begin{aligned}
\mathcal{F}(\mathbf{p}) &= \phi(\mathbf{x}(t_\mathrm{f}),\overline{\mathbf{p}})\\
\nabla_{p_1}\mathcal{F}(\overline{\mathbf{p}}) &= \phi_{p_1}(\mathbf{x}(t_\mathrm{f}),\overline{\mathbf{p}}) + \boldsymbol{\lambda}^\star(t_0)^\mathsf{T}\mathbf{h}_{p_1}(\overline{\mathbf{p}}) + q_1(t_0)\\
&\vdots\\
\nabla_{p_{n_p}}\mathcal{F}(\overline{\mathbf{p}}) &= \phi_{p_{n_p}}(\mathbf{x}(t_\mathrm{f}),\overline{\mathbf{p}}) + \boldsymbol{\lambda}^\star(t_0)^\mathsf{T}\mathbf{h}_{p_{n_p}}(\overline{\mathbf{p}}) + q_{n_p}(t_0)
\end{aligned}$$

In number of remarks are in order. The state and adjoint equations are integrated forward and backward in time, respectively, from their natural initial/terminal conditions. This way, the adjoint equations are stable provided that the state equations are themselves stable [15]. However, integrating the adjoint equations backward in time requires that the state values be "reconstructed" at each time instant. This can be done by *storing* all the necessary information about the state variables at each time step during the forward integration, and then *interpolating* these values during the backward integration; possible interpolation schemes are linear interpolation (requires $\mathbf{x}(t)$ at each time step only), or cubic Hermite interpolation (requires both $\mathbf{x}(t)$ and $\dot{\mathbf{x}}(t)$ at each time step).

It is convenient to introduce quadrature variables q_i, $i = 1, \ldots, n_p$, and appending the corresponding equations to the adjoint system, for calculating the integral term in (3.189). Most numerical solver allow dealing with quadrature variables very efficiently since they do not enter into the Jacobian matrix.

In terms of computational efficiency, the cost of the forward sensitivity method (§3.6.1.3) is roughly proportional to the number n_p of sensitivity parameters, and is insensitive to the number of functionals (e.g., $\mathcal{F}_1, \ldots, \mathcal{F}_{n_\mathcal{F}}$). For the adjoint sensitivity method, on the other hand, the computational cost is proportional to the number $n_\mathcal{F}$ of functionals and is insensitive to the number n_p of parameters. The adjoint sensitivity method is therefore advantageous over the forward sensitivity method when the number of sensitivity parameters is large, and the number of functionals is small. Observe also that the adjoint sensitivity method has a disadvantage that it can only compute the gradient of a specified functional;,unlike the forward sensitivity approach which provide the state sensitivities at any time, the intermediate results of the adjoint variables cannot be exploited.

Remark 3.41 (Extension to Functionals Defined at Multiple Time Instants). A useful extension of the adjoint sensitivity method described previously is in the situation where the functional depends on the state values at multiple (fixed) time instants,[17]

$$\mathcal{F}'(\mathbf{p}) := \sum_{k=1}^{N} \phi^k(\mathbf{x}(t_k), \mathbf{p}), \tag{3.190}$$

where $t_0 < t_1 < t_2 < \cdots < t_N = t_\mathrm{f}$, subject to the IVP in ODEs (3.174). It can be shown that the gradient $\boldsymbol{\nabla}_{p_j} \mathcal{F}'(\overline{\mathbf{p}})$, $i = 1, \ldots, n_p$, at $\overline{\mathbf{p}} \in P$ is given by

$$\nabla_{p_j} \mathcal{F}(\overline{\mathbf{p}}) = \sum_{k=1}^{N} \left[\phi^k_{p_j}(\mathbf{x}(t_k), \overline{\mathbf{p}}) + \int_{t_{k-1}}^{t_k} \boldsymbol{\lambda}^\star(t)^\mathsf{T} \mathbf{f}_{p_j}(t, \mathbf{x}(t), \overline{\mathbf{p}}) \, \mathrm{d}t \right] + \boldsymbol{\lambda}^\star(t_0)^\mathsf{T} \mathbf{h}_{p_j}(\overline{\mathbf{p}}),$$

where the adjoint variables $\boldsymbol{\lambda}^\star$ are the solutions to the ODEs

$$\dot{\boldsymbol{\lambda}}^\star(t) = -\mathbf{f}_\mathbf{x}(t, \mathbf{x}(t), \overline{\mathbf{p}})^\mathsf{T} \boldsymbol{\lambda}^\star(t),$$

from the terminal condition

$$\boldsymbol{\lambda}^\star(t_N) = \phi^N_\mathbf{x}(\mathbf{x}(t_N), \overline{\mathbf{p}}),$$

and satisfying the jump conditions

$$\boldsymbol{\lambda}^\star(t_k^-) = \boldsymbol{\lambda}^\star(t_k^+) + \phi^k_\mathbf{x}(\mathbf{x}(t_k), \overline{\mathbf{p}}).$$

[17] Such functionals arise frequently, e.g., in parameter identification problems where the objective is to calculate the parameter values minimizing the gap between a set of measurements and the model prediction, at given time instants.

The extension of the adjoint method to the case of index-1, fully implicit DAE systems has also been proposed recently (see [15] for details).

Although they are less straightforward to implement than forward sensitivity methods, mainly due to the need to store the state profile, adjoint methods are not difficult to automate. Several codes are available for adjoint sensitivity analysis of IVPs in ODEs and DAEs. We list some of these codes in Tab. 3.2. below. Note that the version 7.4 of MATLAB® does not have a function doing adjoint sensitivity analysis.

Table 3.3. Popular codes doing adjoint sensitivity analysis of ODEs/DAEs.

Solver	Website	Lic.	Characteristics
DASPKadjoint	`http://www.engineering.ucsb.edu/~cse/software.html`	free for acad.	based on DASPK3.0[a] (still under development)
CVODES	`http://acts.nersc.gov/sundials/`	free	based on CVODE[b], ODEs only

[a] See Tab. 3.2.
[b] See Tab. 3.1.

3.6.2 Indirect Methods

Having presented numerical methods for calculating the values and gradients of general functionals, we are now ready to investigate numerical methods for solving optimal control problems. Our focus in this subsection is on indirect methods. Essentially, indirect methods attempt to solve optimal control problems by seeking a solution to the (closed system of) necessary conditions of optimality (NCOs), such as those presented earlier in §3.4 and §3.5.

Many indirect methods use iterative procedures based on successive linearization to find a solution to the system of NCOs. A nominal solution is chosen that satisfies part of the NCOs, then this nominal solution is modified by successive linearization so as to meet the remaining NCOs. Popular indirect methods for optimal control include *quasi-linearization methods*, gradient methods such as *control vector iteration*, and *indirect shooting methods* (see [12]). Only the latter class of methods shall be considered herein. We shall first present the indirect shooting method for optimal control problems having terminal equality constraints only (§3.6.2.1), and then discuss its extensions to encompass problems with terminal and/or path inequality constraints (§3.6.2.2).

3.6.2.1 Indirect Shooting Methods

To set forth the basic ideas of indirect shooting methods, we consider first the relatively simple class of problems treated in §3.4.5: Find $\mathbf{u}^\star \in \mathcal{C}^1[t_0, T]^{n_u}$ and $t_\mathrm{f}^\star \in [t_0, T)$ to

$$\text{minimize:} \quad \mathcal{J}(\mathbf{u}, t_\mathrm{f}) := \int_{t_0}^{t_\mathrm{f}} \ell(t, \mathbf{x}(t), \mathbf{u}(t))\, \mathrm{d}t + \phi(t_\mathrm{f}, \mathbf{x}(t_\mathrm{f})) \tag{3.191}$$

$$\text{subject to:} \quad \dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)); \quad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{3.192}$$

$$\psi_k(t_\mathrm{f}, \mathbf{x}(t_\mathrm{f})) = 0, \quad k = 1, \ldots, n_\psi. \tag{3.193}$$

Provided that the problem (3.191–3.193) is normal and $(\mathbf{u}^\star, t_\mathrm{f}^\star)$ is an optimal solution, there must exist a quintuple $(\mathbf{u}^\star(\cdot), \mathbf{x}^\star(\cdot), \boldsymbol{\lambda}^\star(\cdot), \boldsymbol{\nu}^\star, t_\mathrm{f}^\star)$ which satisfies the Euler-Lagrange

equations

$$\dot{\mathbf{x}}^\star(t) = \mathcal{H}_{\boldsymbol{\lambda}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t), \boldsymbol{\lambda}^\star(t)); \qquad \mathbf{x}^\star(t_0) = \mathbf{x}_0 \tag{3.194}$$

$$\dot{\boldsymbol{\lambda}}^\star(t) = -\mathcal{H}_{\mathbf{x}}(t, \mathbf{x}^\star(t), \mathbf{u}^\star(t), \boldsymbol{\lambda}^\star(t)); \qquad \boldsymbol{\lambda}^\star(t_{\mathrm{f}}^\star) = \Phi_{\mathbf{x}}(t_{\mathrm{f}}^\star, \mathbf{x}^\star(t_{\mathrm{f}}^\star)) \tag{3.195}$$

$$\mathbf{0} = \mathcal{H}_{\mathbf{u}}(t, \mathbf{x}(t), \mathbf{u}(t), \boldsymbol{\lambda}(t)), \tag{3.196}$$

for all $t_0 \leq t \leq t_{\mathrm{f}}^\star$, along with the transversal conditions

$$\boldsymbol{\psi}(t_{\mathrm{f}}^\star, \mathbf{x}^\star(t_{\mathrm{f}}^\star)) = \mathbf{0} \tag{3.197}$$

$$\Phi_t(t_{\mathrm{f}}^\star, \mathbf{x}(t_{\mathrm{f}}^\star)) + \mathcal{H}(t_{\mathrm{f}}^\star, \mathbf{x}^\star(t_{\mathrm{f}}^\star), \mathbf{u}^\star(t_{\mathrm{f}}^\star), \boldsymbol{\lambda}^\star(t_{\mathrm{f}}^\star)) = \mathbf{0}, \tag{3.198}$$

with $\Phi := \phi + \boldsymbol{\nu}^{\star\mathsf{T}}\boldsymbol{\psi}$ and $\mathcal{H} := \ell + \boldsymbol{\lambda}^{\star\mathsf{T}}\mathbf{f}$.

Observe that if the adjoint values $\boldsymbol{\lambda}^\star(t_0)$ at initial time, the Lagrange multipliers $\boldsymbol{\nu}^\star$, and the terminal time $t_{\mathrm{f}}^\star$ were known, the Euler-Lagrange equations could be integrated, all together, forward in time. Hence, the idea of indirect shooting is to guess the values of $\boldsymbol{\lambda}^\star(t_0)$, $\boldsymbol{\nu}^\star$, and $t_{\mathrm{f}}^\star$, and then iteratively improve these estimates to satisfy the adjoint terminal conditions and the transversal conditions. (For this reason, this approach is also referred to as *boundary conditions iteration (BCI)* in the literature.) In other words, one wants to find $(\boldsymbol{\lambda}^\star(t_0), \boldsymbol{\nu}^\star, t_{\mathrm{f}}^\star)$ such that

$$\mathbf{b}(\boldsymbol{\lambda}^\star(t_0), \boldsymbol{\nu}^\star, t_{\mathrm{f}}^\star) := \left(\begin{array}{c} \boldsymbol{\lambda}^\star + \boldsymbol{\phi}_{\mathbf{x}} + \boldsymbol{\nu}^{\star\mathsf{T}}\boldsymbol{\phi}_{\mathbf{x}} \\ \boldsymbol{\psi} \\ \ell + \boldsymbol{\lambda}^{\star\mathsf{T}}\mathbf{f} + \boldsymbol{\phi}_t + \boldsymbol{\nu}^\star\boldsymbol{\phi}_t \end{array} \right)_{t=t_{\mathrm{f}}^\star} = \mathbf{0}.$$

In particular, a Newton iteration can be used to improve the estimates. An algorithm implementing the indirect shooting approach for optimal control problems having equality terminal constraints is given in the following.

Initialization Step

Let $\varepsilon > 0$ be a termination scalar, and choose initial values $\boldsymbol{\lambda}_0^0$, $\boldsymbol{\nu}^0$, t_{f}^0. Let $k = 0$ and go to the main step.

Main Step

1. Calculate the defect $\mathbf{b}(\boldsymbol{\lambda}_0^k, \boldsymbol{\nu}^k, t_{\mathrm{f}}^k)$ in the boundary and transversal conditions. If $\|\mathbf{b}(\boldsymbol{\lambda}_0^k, \boldsymbol{\nu}^k, t_{\mathrm{f}}^k)\| < \varepsilon$, stop;
2. Calculate the gradients $\boldsymbol{\nabla}_{\boldsymbol{\lambda}_0}\mathbf{b}(\boldsymbol{\lambda}_0^k, \boldsymbol{\nu}^k, t_{\mathrm{f}}^k)$, $\boldsymbol{\nabla}_{\boldsymbol{\nu}}\mathbf{b}(\boldsymbol{\lambda}_0^k, \boldsymbol{\nu}^k, t_{\mathrm{f}}^k)$, $\nabla_{t_{\mathrm{f}}}\mathbf{b}(\boldsymbol{\lambda}_0^k, \boldsymbol{\nu}^k, t_{\mathrm{f}}^k)$ of the defect;
3. Solve the linear system

$$\left(\begin{array}{c} \boldsymbol{\nabla}_{\boldsymbol{\lambda}_0}\mathbf{b}(\boldsymbol{\lambda}_0^k, \boldsymbol{\nu}^k, t_{\mathrm{f}}^k)^{\mathsf{T}} \\ \boldsymbol{\nabla}_{\boldsymbol{\nu}}\mathbf{b}(\boldsymbol{\lambda}_0^k, \boldsymbol{\nu}^k, t_{\mathrm{f}}^k)^{\mathsf{T}} \\ \nabla_{t_{\mathrm{f}}}\mathbf{b}(\boldsymbol{\lambda}_0^k, \boldsymbol{\nu}^k, t_{\mathrm{f}}^k)^{\mathsf{T}} \end{array} \right)^{\mathsf{T}} \left(\begin{array}{c} \mathbf{d}_{\boldsymbol{\lambda}}^k \\ \mathbf{d}_{\boldsymbol{\nu}}^k \\ d_{t_{\mathrm{f}}}^k \end{array} \right) = -\mathbf{b}(\boldsymbol{\lambda}_0^k, \boldsymbol{\nu}^k, t_{\mathrm{f}}^k)$$

to get the directions $\mathbf{d}_{\boldsymbol{\lambda}}^k$, $\mathbf{d}_{\boldsymbol{\nu}}^k$ and $d_{t_{\mathrm{f}}}^k$.

4. Compute the new estimates

$$\begin{aligned}\boldsymbol{\lambda}_0^{k+1} &= \boldsymbol{\lambda}_0^k + \mathbf{d}_{\boldsymbol{\lambda}}^k \\ \boldsymbol{\nu}^{k+1} &= \boldsymbol{\nu}^k + \mathbf{d}_{\boldsymbol{\nu}}^k \\ t_{\mathrm{f}}^{k+1} &= t_{\mathrm{f}}^k + \mathbf{d}_{t_{\mathrm{f}}}^k.\end{aligned}$$

5. Replace $k \leftarrow k+1$, and go to step 1.

A number of remarks are in order.

- Obviously, the same procedure applies in the absence of terminal constraints, or if the terminal time is fixed. It suffices to reduce the number of free variables, and keep only the relevant necessary conditions in the defect function $\mathbf{b}$.
- Instead of iterating on the adjoint initial conditions $\boldsymbol{\lambda}_0^k$, one may as well guess the state terminal conditions $\mathbf{x}_{\mathrm{f}}^k = \mathbf{x}^k(t_{\mathrm{f}}^k)$, and integrate the Euler-Lagrange equations backward in time (starting from the guessed terminal time t_{f}^k).
- At each Newton iterate, one must supply the gradient of the functionals $\mathbf{b}(\boldsymbol{\lambda}_0^k, \boldsymbol{\nu}^k, t_{\mathrm{f}}^k)$. Although any of the methods described in §3.6.1.2 through §3.6.1.4 can be used, the forward sensitivity analysis approach is often the most efficient, since the number of parameters is usually small, and there are as many functionals as parameters. Note that, for problems with unspecified final time, it is necessary to make a change of variables so that the time range be fixed, e.g., to $[0,1]$; this way, the final time $t_{\mathrm{f}}^\star$ becomes a parameter of the problem, and both the forward and adjoint sensitivity analysis can be conducted. Yet another alternative to avoid gradient calculation is to apply a quasi-Newton method using a DFP or BFGS update scheme for estimating the Jacobian matrix. (see §1.8.3.2).

An illustration of the indirect shooting method is presented in Example 3.42 below.

Example 3.42 (Indirect Shooting Method). Consider the following scalar optimal control problem with terminal state equality constraint:

$$\text{minimize:} \quad \mathcal{J}(u) := \int_0^1 \tfrac{1}{2} u(t)^2 \, \mathrm{d}t \tag{3.199}$$

$$\text{subject to:} \quad \dot{x}(t) = u(t)[1 - x(t)]; \quad x(0) = -1; \quad x(1) = 0, \tag{3.200}$$

with $u \in \mathcal{C}[0,1]$. Provided that $u^\star$ is an optimal for this problem, there must exist a quadruple $(u^\star, x^\star, \lambda^\star, \nu^\star)$ such that

$$\begin{aligned} x^\star(t) &= \mathcal{H}_\lambda = u^\star(t)[1 - x^\star(t)]; \quad x^\star(0) = -1; \quad x^\star(1) = 0 \\ \lambda^\star(t) &= -\mathcal{H}_x = u^\star(t)\lambda^\star(t); \quad \lambda^\star(1) = \nu^\star \\ 0 &= \mathcal{H}_u = u(t) + \lambda(t)[1 - x(t)]. \end{aligned}$$

We now apply the indirect shooting approach to find one such quadruple. A full-step Newton algorithm is used here to find the unspecified adjoint conditions $\lambda^\star(0)$ and the Lagrange multiplier $\nu^\star$ such that

$$\mathbf{b}(\lambda^\star(0), \nu^\star) = \begin{pmatrix} \lambda(1) - \nu \\ x(1) \end{pmatrix} = \mathbf{0},$$

starting from the initial guess $(\lambda_0^0, \nu^0) = (0, 0)$; moreover, the Jacobian matrix of $\mathbf{b}$ is calculated via forward sensitivity analysis.

The resulting Newton iterates are shown on the left plot of Fig. 3.13.. Clearly, the terminal state is not affected by the value of the Lagrange multiplier. Note that, due to the quadratic rate of convergence of the Newton algorithm, the boundary conditions are satisfied within 10^{-4} after 4 iterations in this case. We also display the optimal control, state and adjoint trajectories on the left plot of Fig. 3.13.; the optimal control is constant over $[0, 1]$ in this example.

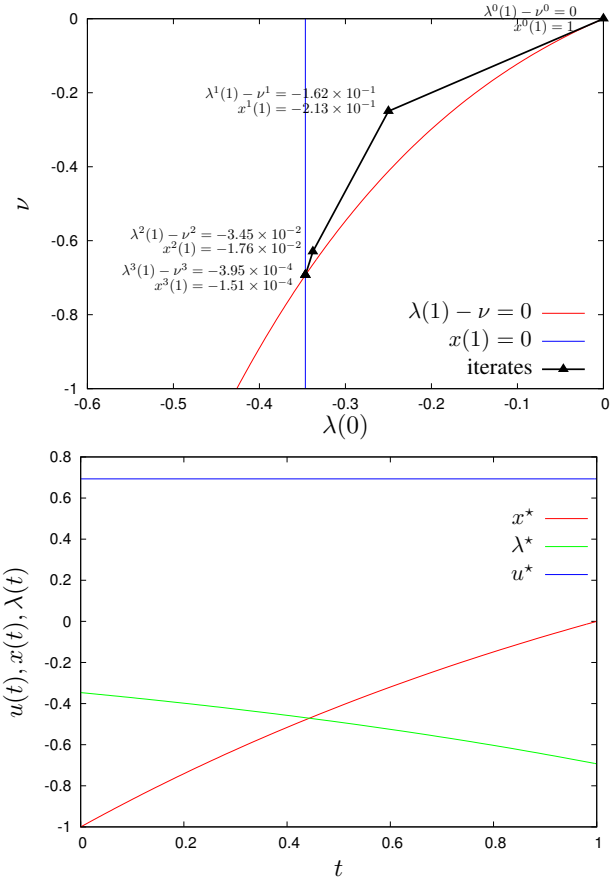


Figure 3.13. Illustration of the indirect shooting method in Example 3.42. Top plot: Newton iterates; bottom plot: Optimal trajectories

The main difficulty with indirect shooting methods is finding a first estimate of the adjoint variables $\boldsymbol{\lambda}_0^0$ (and possibly the final time t_f^0) that produce a solution reasonably close to the specified conditions $\boldsymbol{\lambda}^0(t_\mathrm{f}^0)$ at final time. The reason of this difficulty lies in the high sensitivity of extremal solution to small changes in the unspecified boundary conditions. Since the Euler-Lagrange equations are strongly coupled together, it is not unusual for the numerical integration, with poor initial guesses, to produce 'wild' trajectories in the state/adjoint space. Besides starting difficulty, the indirect shooting approach becomes more difficult in the presence of inequality state constraints as discussed subsequently.

3.6.2.2 Indirect Shooting with Inequality State Constraints For those simple optimal control problems having no inequality constraints and whose solution consists of a single arc, the indirect shooting method proceeds by iteratively refining the estimates of the adjoint variables at initial time, together with the Lagrange multipliers and the terminal time. The situation gets more complicated for problems containing either terminal or path inequality constraints. In this case, one has to postulate the sequence of constrained/unconstrained arcs and the active terminal constraints *a priori*, then calculate the control, state, adjoint, and multiplier functions that satisfy the Euler-Lagrange equations, and finally check that the multipliers satisfy all of the sign conditions. If part of the sign conditions are not met, then the postulated sequence of arcs cannot be optimal; a new sequence of arcs must be selected, and the optimization procedure is repeated until all the NCOs are satisfied.

Particularly difficult to solve are problems having inequality state constraints, since the adjoint trajectories may be discontinuous at the entry time of boundary arcs. In this case, the necessary conditions of optimality yield a nonlinear *multi-point boundary value problem* (MPBVP), i.e., additional conditions must also hold at interior points (see §3.5.6).

Besides the need of specifying rather accurate estimates for the adjoint variables at initial time and entry/contact times, another severe drawback of indirect shooting methods is that detailed *a priori* knowledge of the stucture of the optimal solution must be available. In particular, the direct methods presented in the following subsection do not require such an *a priori* knowledge, and can be used to identify the various types of arcs present in an optimal solution, as well as the active set of terminal constraints.

3.6.3 Direct Methods

Numerical methods that avoid the drawbacks of indirect methods can be found in the so-called *direct optimization methods*, which have been studied extensively over the last 30 years, and have proved to be powerful tools for solving practical optimal control problems. The basic idea of direct optimization methods is to *discretize* the control problem, and then apply NLP techniques to the resulting finite-dimensional optimization problem. These methods use only control and/or state variables as optimization variables and dispense completely with adjoint variables. Moreover, adjoint variables can be obtained by post-optimal calculations using the Lagrange multipliers of the resulting nonlinear optimization problem [14]. Another advantage is that they can be readily extended to problems described by DAE systems. Of course, an obvious drawback of direct methods is that they provide suboptimal solutions only, due to the discretization of the control profiles.

In this section, we shall present two popular direct methods, namely, the sequential method (3.6.3.1) and the simultaneous method (3.6.3.2), which differ in the level of discretization. The pros and cons of either methods will be discussed, and several illustrative examples will be given. In order to set forth the principles of direct methods, we shall consider the following optimal control problem throughout:

Problem 3.43. *Find* $\mathbf{u}^\star \in \hat{\mathcal{C}}^1[t_0, t_\mathrm{f}]^{n_u}$ *and* $\mathbf{v}^\star \in \mathbb{R}^{n_v}$ *to*

$$\textit{minimize:}\quad \mathcal{J}(\mathbf{u}) := \int_{t_0}^{t_\mathrm{f}} \ell(t, \mathbf{x}(t), \mathbf{u}(t), \mathbf{v})\, \mathrm{d}t + \phi(\mathbf{x}(t_\mathrm{f}), \mathbf{v}) \tag{3.201}$$

$$\textit{subject to:}\quad \dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t), \mathbf{v});\quad \mathbf{x}(t_0) = \mathbf{h}(\mathbf{v}) \tag{3.202}$$

$$\psi_j(\mathbf{x}(t_\mathrm{f}), \mathbf{v}) = 0,\quad j = 1, \ldots, n_\psi \tag{3.203}$$

$$\kappa_j(\mathbf{x}(t_\mathrm{f}), \mathbf{v}) \leq 0,\quad j = 1, \ldots, n_\kappa \tag{3.204}$$

$$g_j(t, \mathbf{x}(t), \mathbf{u}(t), \mathbf{v}) \leq 0,\quad j = 1, \ldots, n_g \tag{3.205}$$

$$\mathbf{u}(t) \in [\mathbf{u}^L, \mathbf{u}^U],\qquad \mathbf{v} \in [\mathbf{v}^L, \mathbf{v}^U]. \tag{3.206}$$

In this formulation, the optimization horizon is fixed. However, free final time problems can be easily handled by normalizing the time horizon, and considering the actual final time t_f as an extra time-invariant parameter in the time-invariant parameter vector $\mathbf{v}$.

3.6.3.1 Direct Sequential Methods In direct sequential methods, the control variables $\mathbf{u}(\cdot)$ are parameterized by a finite set of parameters, and the optimization is carried out in the parameter space; hence, this approach is often referred to as *control vector parameterization (CVP)* in the literature.

A convenient way to parameterize the controls is by subdividing the optimization horizon $[t_0, t_\mathrm{f}]$ into $n_s \geq 1$ control stages,

$$t_0 < t_1 < t_2 < \cdots < t_{n_s} = t_\mathrm{f},$$

and using low-order polynomials on each interval, so that

$$\mathbf{u}(t) = \boldsymbol{\mathcal{U}}^k(t, \boldsymbol{\omega}^k),\quad t_{k-1} \leq t \leq t_k.$$

with $\boldsymbol{\omega}^k \in \mathbb{R}^{n_\omega^k}$. Clearly, different orders may be used for different control variables and/or for different control intervals. For simplicity, we shall assume here that the same polynomial order M is employed for each control variable in all stages.

In practice, *Lagrange polynomials* are often employed to approximate the controls. In stage k, the jth control variable is then given by

$$u_j(t) = \mathcal{U}_j^k(t, \boldsymbol{\omega}^k) = \sum_{i=0}^{M} \omega_{ij}^k \phi_i^{(M)}(\tau^{(k)}),\quad t_{k-1} \leq t \leq t_k \tag{3.207}$$

where $\tau^{(k)} = \frac{t - t_{k-1}}{t_k - t_{k-1}} \in [0, 1]$ stands for the normalized time in stage k, and $\phi_i^{(M)}(\cdot)$ denotes the Lagrange polynomial of order M,[18]

$$\phi_i^{(M)}(\tau) := \begin{cases} 1, & \text{if } M = 0 \\ \displaystyle\prod_{\substack{q=0 \\ q \neq i}}^{M} \frac{\tau - \tau_q}{\tau_i - \tau_q}, & \text{if } M \geq 1, \end{cases} \tag{3.208}$$

[18]Lagrange polynomials have the property that $\phi_i^{(M)}(\tau_q) = \delta_{i,q}$. Hence, at each collocation point τ_q, $q = 0, \ldots, M$, we have:

$$u_j(t_{k-1} + \tau_q(t_k - t_{k-1})) = \omega_{qj}^k.$$

with collocation points $0 \leq \tau_0 < \tau_1 < \cdots < \tau_M \leq 1$. Note that piecewise constant controls are obtained for $M = 0$, and piecewise linear controls for $M = 1$.

- The choice of the set of normalized points τ_i used for construction of the Lagrange polynomials does *not* affect the solution obtained. However, to a certain extent, the choice of some points may be dictated by the need to enforce the control variable bounds given in (3.206). For piecewise linear controls ($M = 1$), it is useful to set $\tau_0 = 0$ and $\tau_1 = 1$, in which case the bounds on variable $u_j(t)$ in stage k can be enforced simply through

$$u_j^L \leq \omega_{ij}^k \leq u_j^U, \quad i = 0, 1.$$

  In fact, bounds can also be enforced for piecewise quadratic or cubic controls through inequality constraints expressed in terms of the parameters $\omega_{i,j,k}$. However, such bounding is problematic for polynomials of higher order.

- For some applications, it may be desirable to enforce some degree of continuity in the control profiles across stage boundaries. Continuity of the jth control variable between stages $k-1$ and k can be achieved simply by constraints of the form

$$\omega_{Mj}^{k-1} = \omega_{0j}^k, \qquad \text{if } \tau_0 = 0 \text{ and } \tau_M = 1.$$

  Higher-order continuity can also be enforced by adding linear constraints derived upon differentiating the Lagrange polynomials with respect to time.

Examples of control profile of various degrees and continuity orders are shown in Fig. 3.14. below.

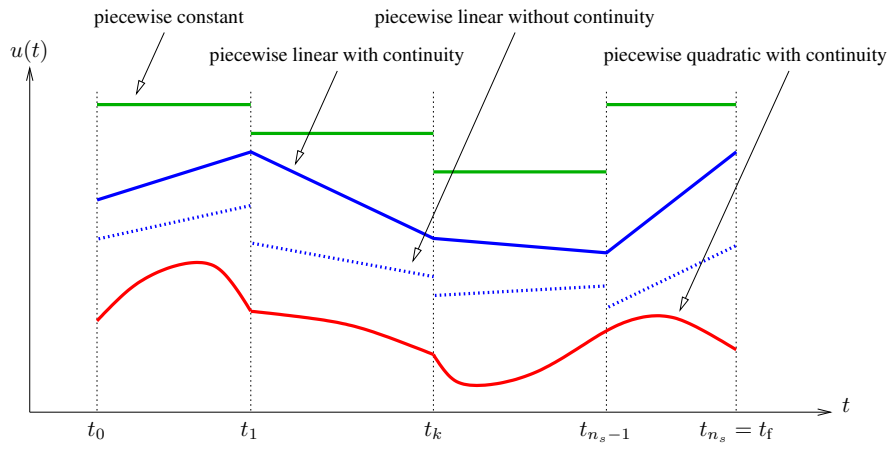


Figure 3.14. Examples of control variable profiles.

Upon parameterization of the controls by (3.207), the problem (3.201–3.206) is transformed into a finite dimensional optimization problem of the following form:

$$\text{minimize:} \quad \sum_{k=1}^{n_s} \int_{t_{k-1}}^{t_k} \ell(t, \mathbf{x}(t), \boldsymbol{\mathcal{U}}^k(t, \boldsymbol{\omega}^k), \mathbf{v}) \, \mathrm{d}t + \phi(\mathbf{x}(t_{n_s}), \mathbf{v}) \tag{3.209}$$

$$\text{subject to:} \quad \dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \boldsymbol{\mathcal{U}}^k(t, \boldsymbol{\omega}^k), \mathbf{v}), \quad t_{k-1} \le t \le t_k, \ k = 1, \ldots, n_s \tag{3.210}$$

$$\mathbf{x}(t_0) = \mathbf{h}(\mathbf{v}) \tag{3.211}$$

$$\boldsymbol{\psi}(\mathbf{x}(t_{n_s}), \mathbf{v}) = \mathbf{0} \tag{3.212}$$

$$\boldsymbol{\kappa}(\mathbf{x}(t_{n_s}), \mathbf{v}) \le \mathbf{0} \tag{3.213}$$

$$\mathbf{g}(t, \mathbf{x}(t), \boldsymbol{\mathcal{U}}^k(t, \boldsymbol{\omega}^k), \mathbf{v}) \le \mathbf{0}, \quad t_{k-1} \le t \le t_k, \ k = 1, \ldots, n_s \tag{3.214}$$

$$\boldsymbol{\omega}^k \in [\boldsymbol{\omega}^L, \boldsymbol{\omega}^U], \quad \mathbf{v} \in [\mathbf{v}^L, \mathbf{v}^U], \tag{3.215}$$

where the optimization parameters are $(\boldsymbol{\omega}^1, \ldots, \boldsymbol{\omega}^{n_s}, \mathbf{v})$. Observe that the path inequality constraints (3.214) consist of an *infinite* number of constraints, since they must hold at each $t_0 \le t \le t_\mathrm{f}$. Several approaches have been suggested to make path constraints tractable:

- *Transcription as integral constraints* [54]: One possible measure of the violation of the jth path constraint (3.214) is

$$\mathcal{G}_j(\boldsymbol{\omega}^1, \ldots, \boldsymbol{\omega}^{n_s}, \mathbf{v}) := \sum_{k=1}^{n_s} \int_{t_{k-1}}^{t_k} \max\{0; g_j(t, \mathbf{x}(t), \boldsymbol{\mathcal{U}}^k(t, \boldsymbol{\omega}^k), \mathbf{v})\}^2 \, \mathrm{d}t.$$

  However, this transcription has the disadvantage that the equality constraint so obtained, $\mathcal{G}_j = 0$, does not satisfy the usual constraint qualification because $\boldsymbol{\nabla}\mathcal{G}_j = 0$ whenever $\mathcal{G}_j = 0$ (see Remark 1.52, p. 25). A possible workaround consists of relaxing the equality constraint as

$$\mathcal{G}_j \le \epsilon,$$

  where $\epsilon > 0$ is a small nonnegative constant. Although this makes the problem regular, slow convergence is often observed in practice.

- *Discretization as interior-point constraints* [57]: Another straightforward technique is to approximate the path inequality constraints (3.214) by imposing pointwise inequality constraints,

$$g_j(t_{k,q}, \mathbf{x}(t_{k,q}), \boldsymbol{\mathcal{U}}^k(t_{k,q}, \boldsymbol{\omega}^k), \mathbf{v}) \le 0,$$

  at a given set of points $t_{k,q} \in [t_{k-1}, t_k]$, in each stage $k = 1, \ldots, n_s$. A disadvantage of this approach is that a rather large number of points $t_{k,q}$ might be necessary to ensure that the path constraints (3.214) are actually not violated between consecutive $t_{k,q}$'s. A hybrid approach combining the discretization approach with the former transcription approach is also possible.

With these reformulations, the parameterized problem (3.209–3.215) consists of a finite number of functionals in either the Mayer or the Lagrange form, subject to an IVP in ODEs. For fixed values of the parameters, one can thus calculate the values of the objective and constraint functionals by using standard numerical integration algorithms (§3.6.1.1). Further, the gradient of the objective and constraint functionals can be calculated with the sensitivity equations of the ODE system (§3.6.1.3) or by integration of the adjoint equations

(§3.6.1.4). Overall, one can therefore apply the numerical methods for NLP problems presented in §1.8, such as SQP, in order to find optimal values $\mathbf{p}^\star$ for the parameters. The direct sequential procedure is illustrated in Fig. 3.15. below and its application is illustrated in Example 3.44.

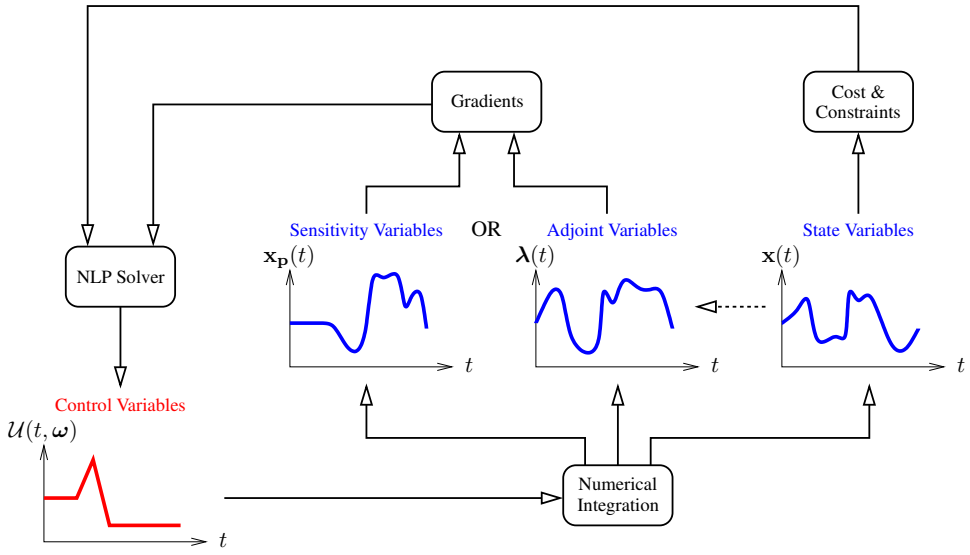


Figure 3.15. Direct sequential methods.

Example 3.44 (Direct Sequential Method). Consider the following scalar optimal control problem:

$$\text{minimize:} \quad \mathcal{J}(u) := \int_0^1 \left([x_1(t)]^2 + [x_2(t)]^2 + \varrho [u(t)]^2 \right) \mathrm{d}t \tag{3.216}$$

$$\text{subject to:} \quad \dot{x}_1(t) = x_2(t); \qquad x_1(0) = 0 \tag{3.217}$$

$$\dot{x}_2(t) = -x_2(t) + u(t); \qquad x_2(0) = -1 \tag{3.218}$$

$$x_2(t) + 0.5 - 8[t - 0.5]^2 \leq 0, \qquad 0 \leq t \leq 1 \tag{3.219}$$

$$-20 \leq u(t) \leq 20, \qquad 0 \leq t \leq 1, \tag{3.220}$$

with $u \in \mathcal{C}[0,1]$, and $\varrho \geq 0$. A rapid analysis of this problems shows that the inequality state constraint is of order $p = 1$. Moreover, the control problem is nonsingular for $\varrho > 0$, and singular for $\varrho = 0$; we investigate these two situations.

Case A: $\varrho = 5 \times 10^{-3}$**.** The direct sequential approach is applied for piecewise constant controls, and $n_s = 10$, 20, 40, and 100 stages. In each case, the gradients are calculated via forward sensitivity analysis, and a SQP solver is used to solve the NLP problem. Note also that the state constraint (3.219) is transcribed as an integral constraint, which is then relaxed as an inequality constraint

$$\int_0^1 \max\{0; x_2(t) + 0.5 - 8[t - 0.5]^2\}^2 \, \mathrm{d}t \leq \epsilon,$$

with $\epsilon = 10^{-6}$.

The results of the direct sequential approach are presented in Fig. 3.16., with the optimal piecewise controls $u^\star$ and the optimal response $x_2^\star$ shown on the left and right plots, respectively. The corresponding optimal cost values are reported the following table:

n_s	10	20	40	100
$\mathcal{J}(u^\star)$	1.79751×10^{-1}	1.71482×10^{-1}	1.69614×10^{-1}	1.69161×10^{-1}

Observe that the reduction in the optimal cost value is negligible when $n_s \geq 40$ stages; this behavior is in fact typical of control parameterization methods. However, by refining the control profile, one can obtain a better idea of the sequence of constrained/unconstrained arcs within the optimal solution. While it is unclear which arcs compose the solution for $n_s = 10$, it becomes obvious, e.g., for $n_s = 100$, that we have three arcs: the first and third arcs are (nonsingular) interior arcs; and the second arc is a boundary arc for the state inequality constraint. Furthermore, the optimal control appears to be continuous at both the entry and exit times of the boundary arc. To confirm it, a possibility would be to apply the indirect shooting method (§3.6.2.1), by using the present results as initial guess, and check whether all the necessary conditions of optimality are satisfied.

Case B: $\varrho = 0$**.** The results of the direct sequential approach, obtained with the same assumptions as in case A, are presented in Fig. 3.16.. The corresponding optimal cost values are reported the following table:

n_s	10	20	40	100
$\mathcal{J}(u^\star)$	1.13080×10^{-1}	0.97320×10^{-1}	0.96942×10^{-1}	0.96893×10^{-1}

Similar to case A, the improvement of the optimal cost value becomes marginal for $n_s \geq 40$. (The optimal cost values are lower here since the control is no longer penalized in the cost functional.) Interestingly, the optimal solution is now composed of four arcs. By inspection, it is not hard to see that: the first arc is a boundary arc for the control constraint $u(t) \leq 20$; the second and fourth arcs are interior arcs; and the third arc is a boundary arc for the state inequality constraint. Regarding interior arcs in particular, we have:

$$
\begin{aligned}
0 &= \mathcal{H}_u = \lambda_2(t) \\
0 &= \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}_u = \dot{\lambda}_2(t) = -2x_2(t) - \lambda_1(t) + \lambda_2(t) \\
0 &= \frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathcal{H}_u = -2\dot{x}_2(t) - \dot{\lambda}_1(t) + \dot{\lambda}_2(t) = -2[u(t) - x_2(t)] - 2x_1(t)
\end{aligned}
$$

Hence, both interior arcs are singular arcs of order $p = 1$, and $u^\star(t) = x_2^\star(t) - x_1^\star(t)$ along these arcs. Moreover, the optimal control appears to be discontinuous at the function between boundary and interior arcs. Again, it would be interesting to confirm these results upon application of an indirect shooting method (§3.6.2.1).

Since the ODEs are solved at each iteration of the NLP solver, direct sequential methods are often called *feasible path methods*. These methods are known to be very robust as

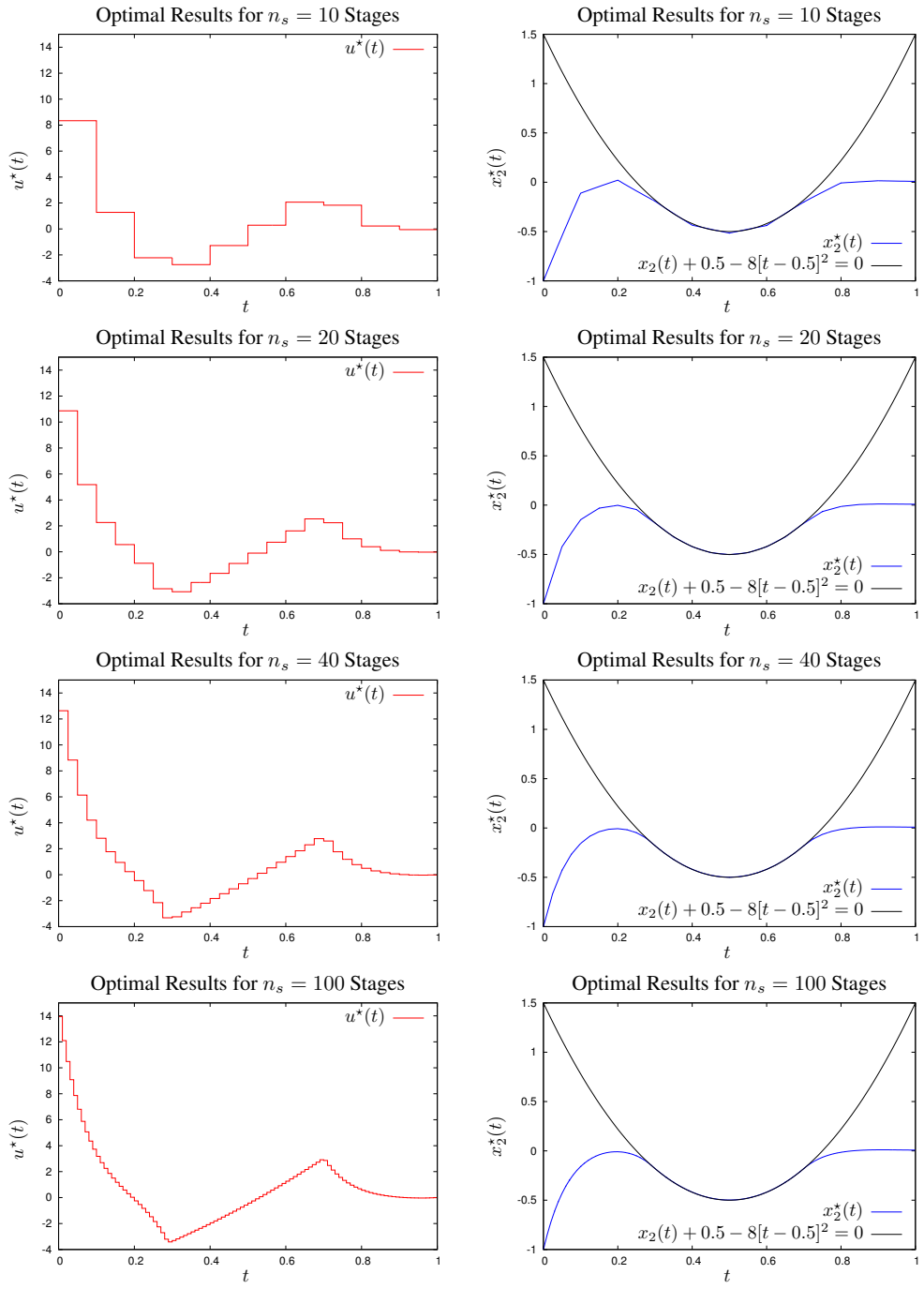


Figure 3.16. Results of the direct sequential approach as applied to Example 3.44 with $\varrho = 5 \times 10^{-3}$ (case A), for $n_s = 10$, 20, 40, and 100. Left plots: optimal piecewise control $u^\star$; right plots: optimal response ${x_2}^\star$

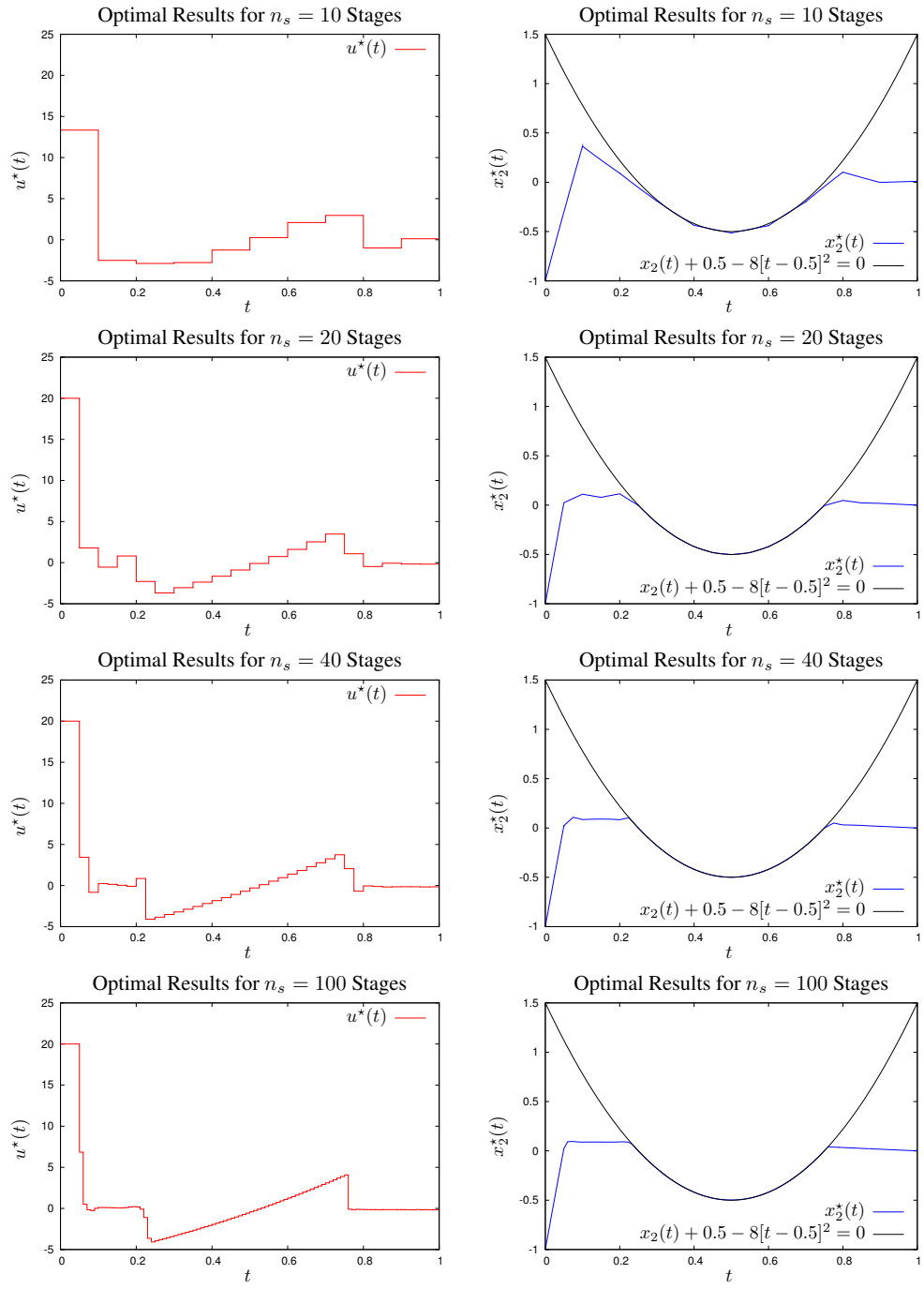


Figure 3.17. Results of the direct sequential approach as applied to Example 3.44 with $\varrho = 0$ (case B), for $n_s = 10$, 20, 40, and 100. Left plots: optimal piecewise control $u^\star$; right plots: optimal response $x_2{}^\star$

long as the dynamic system is stable in the control/parameter range of interest. Moreover, they guarantee the accuracy of the state variables through the error control mechanism of the numerical integration solvers. However, finding an optimal, or even a feasible solution, may prove difficult when the system is unstable or the response is undefined for certain control/parameter values. In addition, much computational effort may be wasted in obtaining accurate state values when the controls/parameters are far from their optimal values. These considerations motivate the direct simultaneous methods which are presented in the following subsection.

3.6.3.2 Direct Simultaneous Methods In direct simultaneous methods, the original optimal control problem (3.201–3.206) is transcribed into a finite dimensional NLP problem through the discretization of *all* the variables, i.e., both the control *and* the state variables. Accordingly, this approach is often referred to as *full discretization* in the literature.

Regarding the control variables first, similar parameterizations as those described in the direct sequential approach §(3.6.3.1) can be used; for the jth control variable in the control stage k, we have

$$u_j(t) = \mathcal{U}_j^k(t, \boldsymbol{\omega}^k) = \sum_{i=0}^{N} \omega_{i,j}^k \phi_i^{(M)}(\tfrac{t-t_{k-1}}{t_k - t_{k-1}}), \qquad t_{k-1} \le t \le t_k,$$

with $\phi_i^{(M)}(\cdot)$ given by (3.208). In particular, piecewise constant or piecewise linear control parameterizations are often considered in practical applications.

To approximate the state variables, on the other hand, a family of polynomials is also considered on each interval, so that

$$\mathbf{x}(t) = \boldsymbol{\mathcal{X}}^k(t, \boldsymbol{\xi}^k), \quad t_{k-1} \le t \le t_k, \; k = 1, \ldots, n_s,$$

with $\boldsymbol{\xi}^k \in \mathbb{R}^{n_\xi^k}$. For simplicity, we shall assume subsequently that the polynomials have the same order N for each state variable and each stage. Different polynomial representations have been suggested in the literature:

- *Lagrange Polynomials* [35]: Similar to the control parameterization, the jth state variable in stage k is calculated as

$$x_j(t) = \mathcal{X}_j^k(t, \boldsymbol{\xi}^k) = \sum_{i=0}^{N} \xi_{i,j}^k \phi_i^{(N)}(\tfrac{t-t_{k-1}}{t_k - t_{k-1}}), \quad t_{k-1} \le t \le t_k, \tag{3.221}$$

  with $\phi_i^{(N)}(\cdot)$ given by (3.208).

- *Monomial Basis Representation* [4]: The jth state variable in stage k is calculated as

$$x_j(t) = \mathcal{X}_j^k(t, \boldsymbol{\xi}^k) = \xi_{0,j}^k + (t_k - t_{k-1}) \sum_{i=1}^{N} \xi_{i,j}^k \Omega_i^{(N)}(\tfrac{t-t_{k-1}}{t_k - t_{k-1}}), \quad t_{k-1} \le t \le t_k, \tag{3.222}$$

  where $\Omega_i^{(N)}(\cdot)$ is a polynomial of order N satisfying

$$\Omega_i^{(N)}(0) := 0$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\Omega_i^{(N)}(\tau_q) := \delta_{i,q}, \quad q = 1, \ldots, N,$$

with collocation points $0 = \tau_0 \leq \tau_1 < \tau_2 < \cdots < \tau_N \leq 1$.

By using either of the foregoing polynomial representations, the problem (3.201–3.206) can be rewritten into the following form:

$$\text{minimize:} \quad \sum_{k=1}^{n_s} \int_{t_{k-1}}^{t_k} \ell(t, \boldsymbol{\mathcal{X}}^k(t, \boldsymbol{\xi}^k), \boldsymbol{\mathcal{U}}^k(t, \boldsymbol{\omega}^k), \mathbf{v})\, \mathrm{d}t + \phi(\boldsymbol{\mathcal{X}}^{n_s}(t_{n_s}, \boldsymbol{\xi}^{n_s}), \mathbf{v}) \tag{3.223}$$

$$\text{subject to:} \quad \boldsymbol{\mathcal{X}}_t^k(t_{k,q}, \boldsymbol{\xi}^k) = \mathbf{f}(t_{k,q}, \boldsymbol{\mathcal{X}}^k(t_{k,q}, \boldsymbol{\xi}^k), \boldsymbol{\mathcal{U}}^k(t_{k,q}, \boldsymbol{\omega}^k), \mathbf{v}) \tag{3.224}$$

$$\boldsymbol{\mathcal{X}}^1(t_0, \boldsymbol{\xi}^1) = \mathbf{h}(\mathbf{v}); \qquad \boldsymbol{\mathcal{X}}^k(t_k, \boldsymbol{\xi}^k) = \boldsymbol{\mathcal{X}}^{k-1}(t_k, \boldsymbol{\xi}^{k-1}) \tag{3.225}$$

$$\boldsymbol{\psi}(\boldsymbol{\mathcal{X}}^{n_s}(t_{n_s}, \boldsymbol{\xi}^{n_s}), \mathbf{v}) = 0 \tag{3.226}$$

$$\boldsymbol{\kappa}(\boldsymbol{\mathcal{X}}^{n_s}(t_{n_s}, \boldsymbol{\xi}^{n_s}), \mathbf{v}) \leq 0 \tag{3.227}$$

$$\mathbf{g}(t_{k,q}, \boldsymbol{\mathcal{X}}^k(t_{k,q}, \boldsymbol{\xi}^k), \boldsymbol{\mathcal{U}}^k(t_{k,q}, \boldsymbol{\omega}^k), \mathbf{v}) \leq 0 \tag{3.228}$$

$$\boldsymbol{\xi}^k \in [\boldsymbol{\xi}^L, \boldsymbol{\xi}^U], \quad \boldsymbol{\omega} \in [\boldsymbol{\omega}^L, \boldsymbol{\omega}^U], \quad \mathbf{v} \in [\mathbf{v}^L, \mathbf{v}^U], \tag{3.229}$$

where $t_{k,q} := t_{k-1} + \tau_q(t_k - t_{k-1})$, with $k = 1, \ldots, n_s$, and $q = 1, \ldots, N$.

A number of remarks are in order:

- The continuous differential equations (3.202) are discretized into $(N + 1)$ equality constraints (3.224) in each time stage, $k = 1, \ldots, n_s$; moreover, the conditions (3.225) are imposed so that the state variables are continuous at the junctions between consecutive time stages. That is, a common characteristic of direct simultaneous methods is that the differential equations are, in general, satisfied at the solution of the optimization problem only; for this reason, these methods are often called *infeasible path methods*.

- The inequality state constraints (3.205) are also discretized into a finite number of inequality constraints (3.228), which must hold at every collocation point in each time stage. Hence, an advantage of simultaneous methods over sequential methods is that they allow handling state inequality constraints more easily, i.e., by enforcing interior-point inequality constraint at collocation points.

- The time stages $t_1, \ldots, t_{n_s}$ can be optimized very easily in direct simultaneous methods, together with the other parameters $(\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^{n_s}, \boldsymbol{\omega}^1, \ldots, \boldsymbol{\omega}^{n_s}, \mathbf{v})$.

Unlike sequential methods, direct simultaneous methods have the advantage of not wasting computational effort in obtaining feasible solutions to the ODEs, away from the solution of the optimization problem. This also allows to handle efficiently those dynamic systems for which instabilities occur in the range of inputs, or for which a solution does not exist for certain inputs. On the other hand, however, only the converged solution satisfies the ODEs, while the intermediate solutions have no physical meaning. Moreover, one does not know *a priori* how many time stages and collocation points should be taken for obtaining an accurate solution to the ODEs. Finally, the resulting NLP problem in the variables $(\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^{n_s}, \boldsymbol{\omega}^1, \ldots, \boldsymbol{\omega}^{n_s}, \mathbf{v})$ is a large-scale NLP problem which may be difficult to solve. This difficulty has led to the development of special decomposition techniques to solve such NLP problems.

3.7 NOTES AND REFERENCES

Only a brief discussion on the problem of existence of an optimal solution has been given in §3.3. A comprehensive treatment of the existence problem can be found in the book by Macki and Strauss [38, Chapter 4]. For a more mathematical presentation of existence results, the interested reader is referred to the book by Cesari [16].

The seminal textbook for the variational approach presented in §3.4 is the book by Bryson and Ho [12]. The book by Kamien and Schwartz [28] has also been useful in writing this chapter, especially regarding the interpretation of the adjoint variables.

In this chapter, we have also limited our presentation of the sufficient conditions to the so-called Mangasarian conditions. More general sufficient conditions can be found in the survey paper by Seierstad and Sydstæter [49]. For local sufficient conditions, we refer the interested reader to the book by Bryson and Ho [12, Chapter 6].

Regarding §3.5 on maximum principles for optimal control problems, the reference textbook is that by Pontryagin and coworkers [41]; a proof of the Pontryagin Maximum Principle can also be found in [38, Chapter 5]. The indirect adjoining approach for optimal control problems with pure path constraints was originally introduced by Pontryagin [41]. For a recent survey on maximum principles for problems having both mixed and pure state constraints, see the paper by [25]. A comprehensive discussion of necessary conditions for singular control problems, including second order necessary conditions, can be found in the paper by Kelley and coworkers [29].

For a general introduction to numerical methods for initial value problems in ODEs and DAEs see the book by Brenan, Campbell and Petzold [11], or Ascher and Petzold [3]. Concerning the calculation of gradients for functionals with ODEs embedded, see, e.g., [45, 56] for the forward sensitivity approach and [54] for the adjoint sensitivity approach. A detailed discussion of indirect solution methods for optimal control problems, including the indirect shooting approach, can be found in [12]. Finally, for more information on direct solution methods, see [54, 56] for the sequential approach and [8, 9] for the simultaneous approach.