NONLINEAR AND DYNAMIC OPTIMIZATION

From Theory to Practice

IC-32: Winter Semester 2006/2007

Benoît C. CHACHUAT
Laboratoire d'Automatique, École Polytechnique Fédérale de Lausanne

CONTENTS

CHAPTER 1

NONLINEAR PROGRAMMING

"Since the fabric of the universe is most perfect, and is the work of a most wise Creator, nothing whatsoever takes place in the universe in which some form of maximum and minimum does not appear."

—Leonhard Euler

1.1 INTRODUCTION

In this chapter, we introduce the nonlinear programming (NLP) problem. Our purpose is to provide some background on nonlinear problems; indeed, an exhaustive discussion of both theoretical and practical aspects of nonlinear programming can be the subject matter of an entire book.

There are several reasons for studying nonlinear programming in an optimal control class. First and foremost, anyone interested in optimal control should know about a number of fundamental results in nonlinear programming. As optimal control problems are optimization problems in (infinite-dimensional) functional spaces, while nonlinear programming are optimization problems in Euclidean spaces, optimal control can indeed be seen as a generalization of nonlinear programming.

Second and as we shall see in Chapter 3, NLP techniques are used routinely and are particularly efficient in solving optimal control problems. In the case of a *discrete* control problem, i.e., when the controls are exerted at discrete points, the problem can be directly stated as a NLP problem. In a *continuous* control problem, on the other hand, i.e., when

the controls are *functions* to be exerted over a prescribed planning horizon, an *approximate* solution can be found by solving a NLP problem.

Throughout this section, we shall consider the following NLP problem:

$$\begin{array}{ll} \min\limits_{\mathbf{x}} & f(\mathbf{x}) \\ \text{s.t.} & \mathbf{g}(\mathbf{x}) \leq \mathbf{0} \\ & \mathbf{h}(\mathbf{x}) = \mathbf{0} \\ & \mathbf{x} \in X \end{array} \tag{NLP}$$

where X is a subset of $\mathbb{R}^{n_x}$, $\mathbf{x}$ is a vector of n_x components $x_1, \ldots, x_{n_x}$, and $f : X \to \mathbb{R}$, $\mathbf{g} : X \to \mathbb{R}^{n_g}$ and $\mathbf{h} : X \to \mathbb{R}^{n_h}$ are defined on X.

The function f is usually called the *objective function* or *criterion function*. Each of the constraints $g_i(\mathbf{x}) \leq 0$, $i = 1, \ldots, n_g$, is called an *inequality constraint*, and each of the constraints $h_i(\mathbf{x}) = 0$, $i = 1, \ldots, n_h$, is called an *equality constraint*. Note also that the set X typically includes lower and upper bounds on the variables; the reason for separating variable bounds from the other inequality constraints is that they can play a useful role in some algorithms, i.e., they are handled in a specific way. A vector $\mathbf{x} \in X$ satisfying all the constraints is called a *feasible solution* to the problem; the collection of all such points forms the *feasible region*. The NLP problem, then, is to find a feasible point $\mathbf{x}^\star$ such that $f(\mathbf{x}) \geq f(\mathbf{x}^\star)$ for each feasible point $\mathbf{x}$. Needless to say, a NLP problem can be stated as a maximization problem, and the inequality constraints can be written in the form $\mathbf{g}(\mathbf{x}) \geq \mathbf{0}$.

Example 1.1. Consider the following problem

$$\begin{array}{lrcl} \min\limits_{\mathbf{x}} & (x_1 - 3)^2 + (x_2 - 2)^2 & & \\ \text{s.t.} & x_1^2 - x_2 - 3 & \leq & 0 \\ & x_2 - 1 & \leq & 0 \\ & -x_1 & \leq & 0. \end{array}$$

The objective function and the three inequality constraints are:

$$\begin{aligned} f(x_1, x_2) &= (x_1 - 3)^2 + (x_2 - 2)^2 \\ g_1(x_1, x_2) &= x_1^2 - x_2 - 3 \\ g_2(x_1, x_2) &= x_2 - 1 \\ g_3(x_1, x_2) &= -x_1. \end{aligned}$$

Fig. 1.1. illustrates the feasible region. The problem, then, is to find a point in the feasible region with the smallest possible value of $(x_1 - 3)^2 + (x_2 - 2)^2$. Note that points (x_1, x_2) with $(x_1 - 3)^2 + (x_2 - 2)^2 = c$ are circles with radius $\sqrt{c}$ and center $(3, 2)$. This circle is called the *contour* of the objective function having the value c. In order to minimize c, we must find the circle with the smallest radius that intersects the feasible region. As shown in Fig. 1.1., the smallest circle corresponds to $c = 2$ and intersects the feasible region at the point $(2, 1)$. Hence, the optimal solution occurs at the point $(2, 1)$ and has an objective value equal to 2.

The graphical approach used in Example 1.1 above, i.e., find an optimal solution by determining the objective contour with the smallest objective value that intersects the feasible region, is only suitable for small problems; it becomes intractable for problems containing

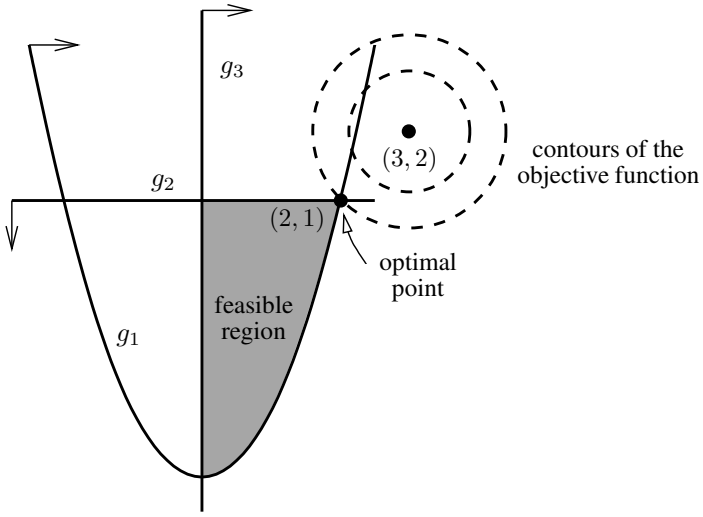


Figure 1.1. Geometric solution of a nonlinear problem.

more than three variables, as well as for problems having complicated objective and/or constraint functions.

This chapter is organized as follows. We start by defining what is meant by optimality, and give conditions under which a minimum (or a maximum) exists for a nonlinear program in §1.2. The special properties of convex programs are then discussed in §1.3. Then, both necessary and sufficient conditions of optimality are presented for NLP problems. We successively consider unconstrained problems (§1.4), problems with inequality constraints (§1.5), and problems with both equality and inequality constraints (§1.7). Finally, several numerical optimization techniques will be presented in §1.8, which are instrumental to solve a great variety of NLP problems.

1.2 DEFINITIONS OF OPTIMALITY

A variety of different definitions of optimality are used in different contexts. It is important to understand fully each definition and the context within which it is appropriately used.

1.2.1 Infimum and Supremum

Let $S \subset \mathbb{R}$ be a nonempty set.

Definition 1.2 (Infimum, Supremum). *The* infimum *of S, denoted as* $\inf S$*, provided it exists, is the* greatest lower bound *for S, i.e., a number α satisfying:*

(i) $z \geq \alpha \quad \forall z \in S$,

(ii) $\forall \bar{\alpha} > \alpha, \exists z \in S$ *such that* $z < \bar{\alpha}$.

Similarly, the supremum *of S, denoted as* $\sup S$*, provided it exists, is the* least upper bound *for S, i.e., a number α satisfying:*

(i) $z \leq \alpha \quad \forall z \in S$,

(ii) $\forall \bar{\alpha} < \alpha, \exists z \in S$ *such that* $z > \bar{\alpha}$.

The first question one may ask concerns the existence of infima and suprema in $\mathbb{R}$. In particular, one **cannot** prove that in $\mathbb{R}$, every set bounded from above has a supremum, and every set bounded from below has an infimum. This is an **axiom**, known as the *axiom of completeness*:

Axiom 1.3 (Axiom of Completeness). *If a nonempty subset of real numbers has an upper bound, then it has a least upper bound. If a nonempty subset of real numbers has a lower bound, it has a greatest lower bound.*

It is important to note that the real number $\inf S$ (resp. $\sup S$), with S a nonempty set in $\mathbb{R}$ bounded from below (resp. from above), although it exist, need not be an element of S.

Example 1.4. Let $S = (0, +\infty) = \{z \in \mathbb{R} : z > 0\}$. Clearly, $\inf S = 0$ and $0 \notin S$.

Notation 1.5. Let $S := \{f(\mathbf{x}) : \mathbf{x} \in D\}$ be the image of the feasible set $D \subset \mathbb{R}^n$ of an optimization problem under the objective function f. Then, the notation

$$\inf_{\mathbf{x} \in D} f(\mathbf{x}) \quad \text{or} \quad \inf\{f(\mathbf{x}) : \mathbf{x} \in D\}$$

refers to the number $\inf S$. Likewise, the notation

$$\sup_{\mathbf{x} \in D} f(\mathbf{x}) \quad \text{or} \quad \sup\{f(\mathbf{x}) : \mathbf{x} \in D\}$$

refers to $\sup S$.

Clearly, the numbers $\inf S$ and $\sup S$ may not be attained by the value $f(\mathbf{x})$ at any $\mathbf{x} \in D$. This is illustrated in an example below.

Example 1.6. Clearly, $\inf\{\exp(x) : x \in (0, +\infty)\} = 1$, but $\exp(x) > 1$ for all $x \in (0, +\infty)$.

By convention, the infimum of an empty set is $+\infty$, while the supremum of an empty set is $-\infty$. That is, if the values $\pm\infty$ are allowed, then infima and suprema always exist.

1.2.2 Minimum and Maximum

Consider the standard problem formulation

$$\min_{\mathbf{x} \in D} f(\mathbf{x})$$

where $D \subset \mathbb{R}^n$ denotes the *feasible set*. Any $\mathbf{x} \in D$ is said to be a *feasible point*; conversely, any $\mathbf{x} \in \mathbb{R}^n \setminus D := \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \notin D\}$ is said to be *infeasible*.

Definition 1.7 ((Global) Minimum, Strict (Global) Minimum). *A point* $\mathbf{x}^\star \in D$ *is said to be a (global)*[1] minimum *of* f *on* D *if*

$$f(\mathbf{x}) \geq f(\mathbf{x}^\star) \quad \forall \mathbf{x} \in D, \tag{1.1}$$

i.e., a minimum is a feasible point whose objective function value is less than or equal to the objective function value of all other feasible points. It is said to be a strict *(global)* minimum *of* f *on* D *if*

$$f(\mathbf{x}) > f(\mathbf{x}^\star) \quad \forall \mathbf{x} \in D,\ \mathbf{x} \neq \mathbf{x}^\star.$$

A (global) maximum is defined by reversing the inequality in Definition 1.7:

Definition 1.8 ((Global) Maximum, Strict (Global) Maximum). *A point* $\mathbf{x}^\star \in D$ *is said to be a (global)* maximum *of* f *on* D *if*

$$f(\mathbf{x}) \leq f(\mathbf{x}^\star) \quad \forall \mathbf{x} \in D. \tag{1.2}$$

It is said to be a strict *(global)* maximum *of* f *on* D *if*

$$f(\mathbf{x}) < f(\mathbf{x}^\star) \quad \forall \mathbf{x} \in D,\ \mathbf{x} \neq \mathbf{x}^\star.$$

The important distinction between minimum/maximum and infimum/supremum is that the value $\min\{f(\mathbf{x}) : \mathbf{x} \in D\}$ **must** be attained at **one or more** points $\mathbf{x} \in D$, whereas the value $\inf\{f(\mathbf{x}) : \mathbf{x} \in D\}$ does not necessarily have to be attained at any points $\mathbf{x} \in D$. Yet, if a minimum (resp. maximum) exists, then its optimal value will equal the infimum (resp. supremum).

Note also that if a minimum exists, it is **not** necessarily **unique**. That is, there may be multiple or even an infinite number of feasible points that satisfy the inequality (1.1) and are thus minima. Since there is in general a set of points that are minima, the notation

$$\arg\min\{f(\mathbf{x}) : \mathbf{x} \in D\} := \{\mathbf{x} \in D : f(\mathbf{x}) = \inf\{f(\mathbf{x}) : \mathbf{x} \in D\}\}$$

is introduced to denote the set of minima; this is a (possibly empty) **set** in $\mathbb{R}^n$.[2]

A minimum $\mathbf{x}^\star$ is often referred to as an *optimal solution*, a *global optimal solution*, or simply a *solution* of the optimization problem. The real number $f(\mathbf{x}^\star)$ is known as the (global) *optimal value* or *optimal solution value*. Regardless of the number of minima, there is always a **unique** real number that is the optimal value (if it exists). (The notation $\min\{f(\mathbf{x}) : \mathbf{x} \in D\}$ is used to refer to this real value.)

Unless the objective function f and the feasible set D possess special properties (e.g., convexity), it is usually very hard to devise algorithms that are capable of locating or estimating a global minimum or a global maximum with certainty. This motivates the definition of local minima and maxima, which, by the nature of their definition in terms of local information, are much more convenient to locate with an algorithm.

Definition 1.9 (Local Minimum, Strict Local Minimum). *A point* $\mathbf{x}^\star \in D$ *is said to be a* local minimum *of* f *on* D *if*

$$\exists \varepsilon > 0 \text{ such that } f(\mathbf{x}) \geq f(\mathbf{x}^\star) \quad \forall \mathbf{x} \in \mathcal{B}_\varepsilon(\mathbf{x}^\star) \cap D.$$

[1] Strictly, it is not necessary to qualify minimum with 'global' because minimum means a feasible point at which the smallest objective function value is attained. Yet, the qualification global minimum is often made to emphasize that a local minimum is not adequate.

[2] The notation $\bar{\mathbf{x}} = \arg\min\{f(\mathbf{x}) : \mathbf{x} \in D\}$ is also used by some authors. In this case, $\arg\min\{f(\mathbf{x}) : \mathbf{x} \in D\}$ should be understood as a function returning a point $\bar{\mathbf{x}}$ that minimizes f on D. (See, e.g., `http://planetmath.org/encyclopedia/ArgMin.html`.)

$\mathbf{x}^\star$ *is said to be a* strict local minimum *if*

$$\exists \varepsilon > 0 \text{ such that } f(\mathbf{x}) > f(\mathbf{x}^\star) \quad \forall \mathbf{x} \in \mathcal{B}_\varepsilon(\mathbf{x}^\star) \setminus \{\mathbf{x}^\star\} \cap D.$$

The qualifier 'local' originates from the requirement that $\mathbf{x}^\star$ be a minimum only for those feasible points in a neighborhood around the local minimum.

Remark 1.10. Trivially, the property of $\mathbf{x}^\star$ being a global minimum implies that $\mathbf{x}^\star$ is also a local minimum because a global minimum is local minimum with ε set arbitrarily large.

A local maximum is defined by reversing the inequalities in Definition 1.9:

Definition 1.11 (Local Maximum, Strict Local Maximum). *A point* $\mathbf{x}^\star \in D$ *is said to be a* local maximum *of* f *on* D *if*

$$\exists \varepsilon > 0 \text{ such that } f(\mathbf{x}) \leq f(\mathbf{x}^\star) \quad \forall \mathbf{x} \in \mathcal{B}_\varepsilon(\mathbf{x}^\star) \cap D.$$

$\mathbf{x}^\star$ *is said to be a* strict local maximum *if*

$$\exists \varepsilon > 0 \text{ such that } f(\mathbf{x}) < f(\mathbf{x}^\star) \quad \forall \mathbf{x} \in \mathcal{B}_\varepsilon(\mathbf{x}^\star) \setminus \{\mathbf{x}^\star\} \cap D.$$

Remark 1.12. It is important to note that the concept of a global minimum or a global maximum of a function on a set is defined without the notion of a *distance* (or a norm in the case of a vector space). In contrast, the definition of a local minimum or a local maximum requires that a distance be specified on the set of interest. In $\mathbb{R}^{n_x}$, norms are equivalent, and it is readily shown that local minima (resp. maxima) in $(\mathbb{R}^{n_x}, \|\cdot\|_\alpha)$ match local minima (resp. maxima) in $(\mathbb{R}^{n_x}, \|\cdot\|_\beta)$, for any two arbitrary norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ in $\mathbb{R}^{n_x}$ (e.g., the Euclidean norm $\|\cdot\|_2$ and the infinite norm $\|\cdot\|_\infty$). Yet, this nice property does not hold in linear functional spaces, as those encountered in problems of the calculus of variations (§2) and optimal control (§3).

Fig. 1.2. illustrates the various definitions of minima and maxima. Point x^1 is the unique global maximum; the objective value at this point is also the supremum. Points a, x^2, and b are strict local minima because there exists a neighborhood around each of these point for which a, x^2, or b is the unique minimum (on the intersection of this neighborhood with the feasible set D). Likewise, point x^3 is a strict local maximum. Point x^4 is the unique global minimum; the objective value at this point is also the infimum. Finally, point x^5 is simultaneously a local minimum and a local maximum because there are neighborhoods for which the objective function remains constant over the entire neighborhood; it is neither a strict local minimum, nor a strict local maximum.

Example 1.13. Consider the function

$$f(x) = \begin{cases} +1 & \text{if } x < 0 \\ -1 & \text{otherwise.} \end{cases} \tag{1.3}$$

The point $x^\star = -1$ is a local minimum for

$$\min_{x \in [-2,2]} f(x)$$

with value $f(x^\star) = +1$. The optimal value of (1.3) is -1, and $\arg\min\{f(x) : x \in [-2,2]\} = [0,2]$.

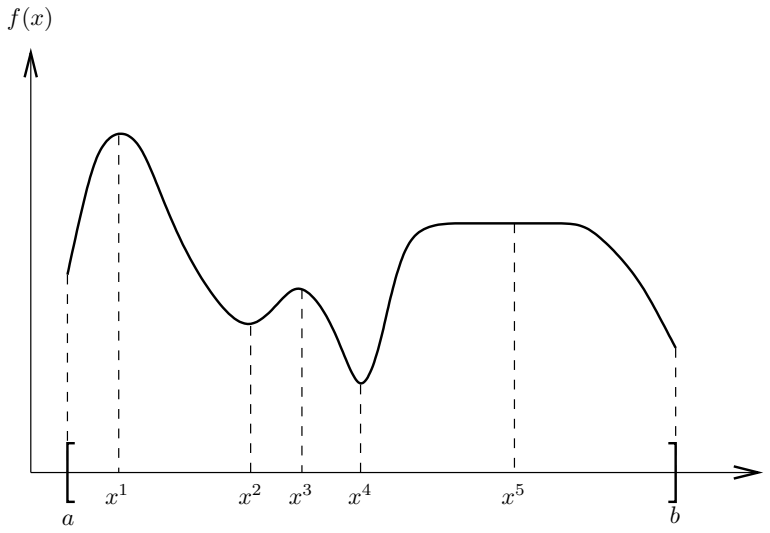


Figure 1.2. The various types of minima and maxima.

1.2.3 Existence of Minima and Maxima

A crucial question when it comes to optimizing a function on a given set, is whether a minimizer or a maximizer exist for that function in that set. Strictly, a minimum or maximum should only be referred to when it is known to exist.

Fig 1.3. illustrates three instances where a minimum does not exist. In Fig 1.3.(a), the infimum of f over $S := (a, b)$ is given by $f(b)$, but since S is not closed and, in particular, $b \notin S$, a minimum does not exist. In Fig 1.3.(b), the infimum of f over $S := [a, b]$ is given by the limit of $f(x)$ as x approaches c from the left, i.e., $\inf\{f(x) : x \in S\} = \lim_{x \to c^-} f(x)$. However, because f is discontinuous at c, a minimizing solution does not exist. Finally, Fig 1.3.(c) illustrates a situation within which f is unbounded over the unbounded set $S := \{x \in \mathbb{R} : x \geq a\}$.

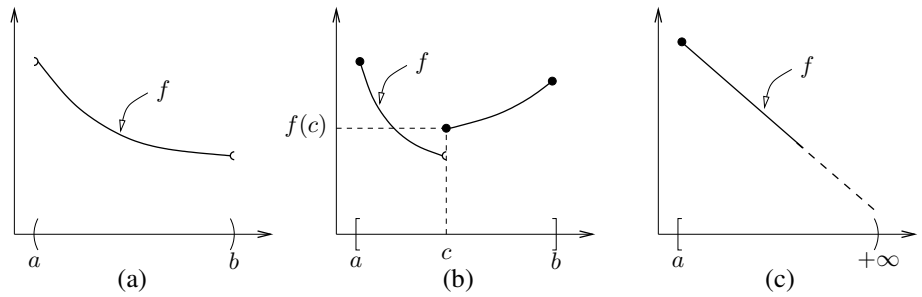


Figure 1.3. The nonexistence of a minimizing solution.

We now formally state and prove the result that if S is nonempty, closed, and bounded, and if f is continuous on S, then, unlike the various situations of Fig. 1.3., a minimum exists.

Theorem 1.14 (Weierstrass' Theorem). *Let S be a nonempty, compact set, and let $f : S \to \mathbb{R}$ be continuous on S. Then, the problem $\min\{f(\mathbf{x}) : \mathbf{x} \in S\}$ attains its minimum, that is, there exists a minimizing solution to this problem.*

Proof. Since f is continuous on S and S is both closed and bounded, f is bounded below on S. Consequently, since $S \neq \emptyset$, there exists a greatest lower bound $\alpha := \inf\{f(\mathbf{x} : \mathbf{x} \in S\}$ (see Axiom 1.3). Now, let $0 < \varepsilon < 1$, and consider the set $S_k := \{\mathbf{x} \in S : \alpha \leq f(\mathbf{x}) \leq \alpha + \varepsilon^k\}$ for $k = 1, 2, \ldots$. By the definition of an infimum, $S_k \neq \emptyset$ for each k, and so we may construct a sequence of points $\{\mathbf{x}_k\} \subset S$ by selecting a point $\mathbf{x}_k$ for each $k = 1, 2, \ldots$. Since S is bounded, there exists a convergent subsequence $\{\mathbf{x}_k\}_{\mathcal{K}} \subset S$ indexed by the set $\mathcal{K} \subset \mathbb{N}$; let $\bar{\mathbf{x}}$ denote its limit. By the closedness of S, we have $\bar{\mathbf{x}} \in S$; and by the continuity of f on S, since $\alpha \leq f(\mathbf{x}_k) \leq \alpha + \varepsilon^k$, we have $\alpha = \lim_{k\to\infty, k\in\mathcal{K}} f(\mathbf{x}_k) = f(\bar{\mathbf{x}})$. Hence, we have shown that there exist a solution $\bar{\mathbf{x}} \in S$ such that $f(\bar{\mathbf{x}}) = \alpha = \inf\{f(\mathbf{x} : \mathbf{x} \in S\}$, i.e., $\bar{\mathbf{x}}$ is a minimizing solution. □

The hypotheses of Theorem 1.14 can be justified as follows: (i) the feasible set must be **nonempty**, otherwise there are no feasible points at which to attain the minimum; (ii) the feasible set must contain its boundary points, which is ensured by assuming that the feasible set is **closed**; (iii) the objective function must be **continuous** on the feasible set, otherwise the limit at a point may not exist or be different from the value of the function at that point; and (iv) the feasible set must be **bounded** because otherwise even a continuous function can be unbounded on the feasible set.

Example 1.15. Theorem 1.14 establishes that a minimum (and a maximum) of

$$\min_{x\in[-1,1]} x^2$$

exists, since $[-1, 1]$ is a nonempty, compact set and $x \mapsto x^2$ is a continuous function on $[-1, 1]$. On the other hand, minima can still exist even though the set is not compact or the function is not continuous, for Theorem 1.14 only provides a sufficient condition. This is the case for the problem

$$\min_{x\in(-1,1)} x^2,$$

which has a minimum at $x = 0$. (See also Example 1.13.)

Example 1.16. Consider the NLP problem of Example 1.1 (p. 2),

$$\begin{array}{lrcl} \min\limits_{\mathbf{x}} & (x_1 - 3)^2 + (x_2 - 2)^2 & & \\ \text{s.t.} & x_1^2 - x_2 - 3 & \leq & 0 \\ & x_2 - 1 & \leq & 0 \\ & -x_1 & \leq & 0. \end{array}$$

The objective function being continuous and the feasible region being nonempty, closed and bounded, the existence of a minimum to this problem directly follows from Theorem 1.14.

1.3 CONVEX PROGRAMMING

A particular class of nonlinear programs is that of convex programs (see Appendix A.3 for a general overview on convex sets and convex functions):

Definition 1.17 (Convex Program). *Let C be a nonempty convex set in $\mathbb{R}^n$, and let $f : C \to \mathbb{R}$ be convex on C. Then,*

$$\min_{\mathbf{x}\in C} f(\mathbf{x})$$

is said to be a convex program *(or a* convex optimization problem*).*

Convex programs possess nicer theoretical properties than general nonconvex problems. The following theorem is a fundamental result in convex programming:

Theorem 1.18. *Let $\mathbf{x}^\star$ be a local minimum of a convex program. Then, $\mathbf{x}^\star$ is also a global minimum.*

Proof. $\mathbf{x}^\star$ being a local minimum,

$$\exists \varepsilon > 0 \text{ such that } f(\mathbf{x}) \geq f(\mathbf{x}^\star), \forall \mathbf{x} \in \mathcal{B}_\varepsilon(\mathbf{x}^\star).$$

By contradiction, suppose that $\mathbf{x}^\star$ is not a global minimum. Then,

$$\exists \bar{\mathbf{x}} \in C \text{ such that } f(\bar{\mathbf{x}}) < f(\mathbf{x}^\star). \tag{1.4}$$

Let $\lambda \in (0,1)$ be chosen such that $\mathbf{y} := \lambda\bar{\mathbf{x}} + (1-\lambda)\mathbf{x}^\star \in \mathcal{B}_\varepsilon(\mathbf{x}^\star)$. By convexity of C, $\mathbf{y}$ is in C. Next, by convexity of f on C and (1.4),

$$f(\mathbf{y}) \leq \lambda f(\bar{\mathbf{x}}) + (1-\lambda) f(\mathbf{x}^\star) < \lambda f(\mathbf{x}^\star) + (1-\lambda) f(\mathbf{x}^\star) = f(\mathbf{x}^\star),$$

hence contradicting the assumption that $\mathbf{x}^\star$ is a local minimum. □

Example 1.19. Consider once again the NLP problem of Example 1.1 (p. 2),

$$\begin{array}{rll} \min\limits_{\mathbf{x}} & (x_1-3)^2 + (x_2-2)^2 & \\ \text{s.t.} & x_1^2 - x_2 - 3 & \leq \quad 0 \\ & x_2 - 1 & \leq \quad 0 \\ & -x_1 & \leq \quad 0. \end{array}$$

The objective function f and the inequality constraints g_1, g_2 and g_3 being convex, every local solution to this problem is also a global solution by Theorem 1.18; henceforth, $(1,2)$ is a global solution and the global solution value is 4.

In convex programming, any local minimum is therefore a local optimum. This is a powerful result that makes any local optimization algorithm a global optimization algorithm when applied to a convex optimization problem. Yet, Theorem 1.18 only gives a sufficient condition for that property to hold. That is, a nonlinear program with nonconvex participating functions may not necessarily have local minima that are not global minima.

1.4 UNCONSTRAINED PROBLEMS

An unconstrained problem is a problem of the form to minimize (or maximize) $f(\mathbf{x})$ without any constraints on the variables $\mathbf{x}$:

$$\min\{f(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^{n_x}\}.$$

Note that the feasible domain of $\mathbf{x}$ being unbounded, Weierstrass' Theorem 1.14 does not apply, and one does not know with certainty, whether a minimum actually exists for that problem.[3] Moreover, even if the objective function is convex, one such minimum may not exist (think of $f : x \mapsto \exp x$!). Hence, we shall proceed with the theoretically unattractive task of seeking minima and maxima of functions which need not have them!

Given a point $\mathbf{x}$ in $\mathbb{R}^{n_x}$, necessary conditions help determine whether or not a point is a local or a global minimum of a function f. For this purpose, we are mostly interested in obtaining conditions that can be checked algebraically.

Definition 1.20 (Descent Direction). *Suppose that $f : \mathbb{R}^{n_x} \to \mathbb{R}$ is continuous at $\bar{\mathbf{x}}$. A vector $\mathbf{d} \in \mathbb{R}^{n_x}$ is said to be a* descent direction*, or an* improving direction*, for f at $\bar{\mathbf{x}}$ if*

$$\exists \delta > 0 : f(\bar{\mathbf{x}} + \lambda \mathbf{d}) < f(\bar{\mathbf{x}}) \quad \forall \lambda \in (0, \delta).$$

Moreover, the cone of descent directions *at $\bar{\mathbf{x}}$, denoted by $\mathcal{F}(\bar{\mathbf{x}})$, is given by*

$$\mathcal{F}(\bar{\mathbf{x}}) := \{\mathbf{d} : \exists \delta > 0 \text{ such that } f(\bar{\mathbf{x}} + \lambda \mathbf{d}) < f(\bar{\mathbf{x}}) \quad \forall \lambda \in (0, \delta)\}.$$

The foregoing definition provides a **geometrical** characterization for a descent direction. yet, an **algebraic** characterization for a descent direction would be more useful from a practical point of view. In response to this, let us assume that f is differentiable and define the following set at $\bar{\mathbf{x}}$:

$$\mathcal{F}_0(\bar{\mathbf{x}}) := \{\mathbf{d} : \boldsymbol{\nabla} f(\bar{\mathbf{x}})^{\mathsf{T}} \mathbf{d} < 0\}.$$

This is illustrated in Fig. 1.4., where the half-space $\mathcal{F}_0(\bar{\mathbf{x}})$ and the gradient $\boldsymbol{\nabla} f(\bar{\mathbf{x}})$ are translated from the origin to $\bar{\mathbf{x}}$ for convenience.

The following lemma proves that every element $\mathbf{d} \in \mathcal{F}_0(\bar{\mathbf{x}})$ is a descent direction at $\bar{\mathbf{x}}$.

Lemma 1.21 (Algebraic Characterization of a Descent Direction). *Suppose that $f : \mathbb{R}^{n_x} \to \mathbb{R}$ is differentiable at $\bar{\mathbf{x}}$. If there exists a vector $\mathbf{d}$ such that $\boldsymbol{\nabla} f(\bar{\mathbf{x}})^{\mathsf{T}} \mathbf{d} < 0$, then $\mathbf{d}$ is a descent direction for f at $\bar{\mathbf{x}}$. That is,*

$$\mathcal{F}_0(\bar{\mathbf{x}}) \subseteq \mathcal{F}(\bar{\mathbf{x}}).$$

Proof. f being differentiable at $\bar{\mathbf{x}}$,

$$f(\bar{\mathbf{x}} + \lambda \mathbf{d}) = f(\bar{\mathbf{x}}) + \lambda \boldsymbol{\nabla} f(\bar{\mathbf{x}})^{\mathsf{T}} \mathbf{d} + \lambda \|\mathbf{d}\| \alpha(\lambda \mathbf{d})$$

where $\lim_{\lambda \to 0} \alpha(\lambda \mathbf{d}) = 0$. Rearranging the terms and dividing by $\lambda \neq 0$, we get

$$\frac{f(\bar{\mathbf{x}} + \lambda \mathbf{d}) - f(\bar{\mathbf{x}})}{\lambda} = \boldsymbol{\nabla} f(\bar{\mathbf{x}})^{\mathsf{T}} \mathbf{d} + \|\mathbf{d}\| \alpha(\lambda \mathbf{d}).$$

[3] For unconstrained optimization problems, the existence of a minimum can actually be guaranteed if the objective objective function is such that $\lim_{\|\mathbf{x}\| \to +\infty} f(\mathbf{x}) = +\infty$ (O-coercive function).

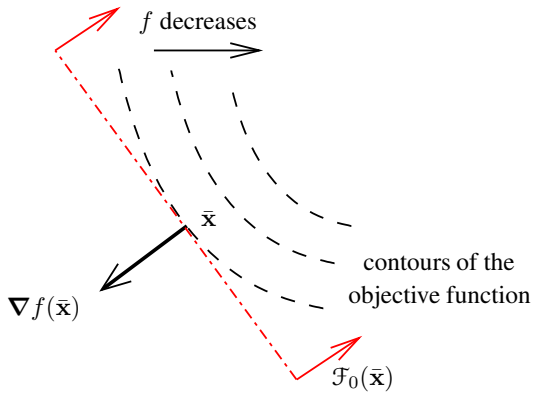


Figure 1.4. Illustration of the set $\mathcal{F}_0(\bar{\mathbf{x}})$.

Since $\nabla f(\bar{\mathbf{x}})^\mathsf{T}\mathbf{d} < 0$ and $\lim_{\lambda\to 0}\alpha(\lambda\mathbf{d}) = 0$, there exists a $\delta > 0$ such that $\nabla f(\bar{\mathbf{x}})^\mathsf{T}\mathbf{d} + \|\mathbf{d}\|\alpha(\lambda\mathbf{d}) < 0$ for all $\lambda \in (0, \delta)$. □

We are now ready to derive a number of necessary conditions for a point to be a local minimum of an unconstrained optimization problem.

Theorem 1.22 (First-Order Necessary Condition for a Local Minimum). *Suppose that $f : \mathbb{R}^{n_x} \to \mathbb{R}$ is differentiable at $\mathbf{x}^\star$. If $\mathbf{x}^\star$ is a local minimum, then $\nabla f(\mathbf{x}^\star) = \mathbf{0}$.*

Proof. The proof proceeds by contraposition. Suppose that $\nabla f(\mathbf{x}^\star) \neq \mathbf{0}$. Then, letting $\mathbf{d} = -\nabla f(\mathbf{x}^\star)$, we get $\nabla f(\mathbf{x}^\star)^\mathsf{T}\mathbf{d} = -\|\nabla f(\mathbf{x}^\star)\|^2 < 0$. By Lemma 1.21,

$$\exists \delta > 0 : f(\mathbf{x}^\star + \lambda\mathbf{d}) < f(\mathbf{x}^\star) \quad \forall \lambda \in (0, \delta),$$

hence contradicting the assumption that $\mathbf{x}^\star$ is a local minimum for f. □

Remark 1.23 (Obtaining Candidate Solution Points). The above condition is called a *first-order necessary condition* because it uses the first-order derivatives of f. This condition indicates that the candidate solutions to an unconstrained optimization problem can be found by solving a system of n_x algebraic (nonlinear) equations. Points $\bar{\mathbf{x}}$ such that $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$ are known as *stationary points*. Yet, a stationary point need *not* be a local minimum as illustrated by the following example; it could very well be a local maximum, or even a *saddle point*.

Example 1.24. Consider the problem

$$\min_{x\in\mathbb{R}} \quad x^2 - x^4.$$

The gradient vector of the objective function is given by

$$\nabla f(x) = 2x - 4x^3,$$

which has three distinct roots $x_1^\star = 0$, $x_2^\star = \frac{1}{\sqrt{2}}$ and $x_2^\star = -\frac{1}{\sqrt{2}}$. Out of these values, $x_1^\star$ gives the smallest cost value, $f(x_1^\star) = 0$. Yet, we cannot declare $x_1^\star$ to be the global

minimum, because we do not know whether a (global) minimum exists for this problem. Indeed, as shown in Fig. 1.5., none of the stationary points is a global minimum, because f decreases to $-\infty$ as $|x| \to \infty$.

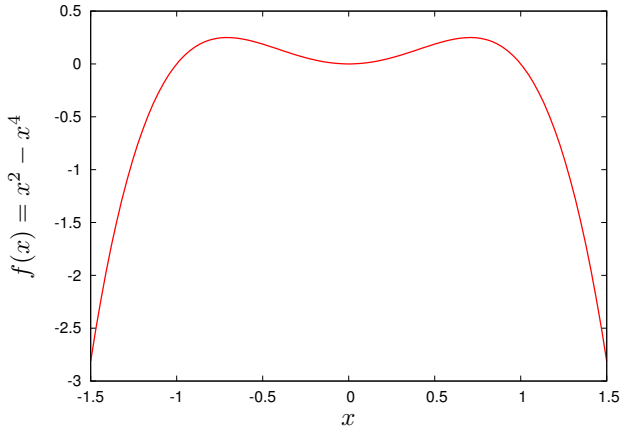


Figure 1.5. Illustration of the objective function in Example 1.24.

More restrictive necessary conditions can also be derived in terms of the Hessian matrix H whose elements are the second-order derivatives of f. One such second-order condition is given below.

Theorem 1.25 (Second-Order Necessary Conditions for a Local Minimum). *Suppose that $f : \mathbb{R}^{n_x} \to \mathbb{R}$ is twice differentiable at $\mathbf{x}^\star$. If $\mathbf{x}^\star$ is a local minimum, then $\nabla f(\mathbf{x}^\star) = \mathbf{0}$ and $\mathsf{H}(\mathbf{x}^\star)$ is positive semidefinite.*

Proof. Consider an arbitrary direction $\mathbf{d}$. Then, from the differentiability of f at $\mathbf{x}^\star$, we have

$$f(\mathbf{x}^\star + \lambda\mathbf{d}) = f(\mathbf{x}^\star) + \lambda\nabla f(\mathbf{x}^\star)^\mathsf{T}\mathbf{d} + \frac{\lambda^2}{2}\mathbf{d}^\mathsf{T}\mathsf{H}(\mathbf{x}^\star)\mathbf{d} + \lambda^2\|\mathbf{d}\|^2\alpha(\lambda\mathbf{d}), \tag{1.5}$$

where $\lim_{\lambda\to 0}\alpha(\lambda\mathbf{d}) = 0$. Since $\mathbf{x}^\star$ is a local minimum, from Theorem 1.22, $\nabla f(\mathbf{x}^\star) = 0$. Rearranging the terms in (1.5) and dividing by λ^2, we get

$$\frac{f(\mathbf{x}^\star + \lambda\mathbf{d}) - f(\mathbf{x}^\star)}{\lambda^2} = \frac{1}{2}\mathbf{d}^\mathsf{T}\mathsf{H}(\mathbf{x}^\star)\mathbf{d} + \|\mathbf{d}\|^2\alpha(\lambda\mathbf{d}).$$

Since $\mathbf{x}^\star$ is a local minimum, $f(\mathbf{x}^\star + \lambda\mathbf{d}) \geq f(\mathbf{x}^\star)$ for λ sufficiently small. By taking the limit as $\lambda \to 0$, it follows that $\mathbf{d}^\mathsf{T}\mathsf{H}(\mathbf{x}^\star)\mathbf{d} \geq 0$. Since $\mathbf{d}$ is arbitrary, $\mathsf{H}(\mathbf{x}^\star)$ is therefore positive semidefinite. □

Example 1.26. Consider the problem

$$\min_{\mathbf{x}\in\mathbb{R}^2} \quad x_1x_2.$$

The gradient vector of the objective function is given by

$$\nabla f(\mathbf{x}) = \begin{bmatrix} x_2 & x_1 \end{bmatrix}^{\mathsf{T}}$$

so that the only stationary point in $\mathbb{R}^2$ is $\bar{\mathbf{x}} = (0, 0)$. Now, consider the Hessian matrix of the objective function at $\bar{\mathbf{x}}$:

$$\mathsf{H}(\bar{\mathbf{x}}) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \forall \mathbf{x} \in \mathbb{R}^2.$$

It is easily checked that $\mathsf{H}(\bar{\mathbf{x}})$ is indefinite, therefore, by Theorem 1.25, the stationary point $\bar{\mathbf{x}}$ is not a (local) minimum (nor is it a local maximum). Such stationary points are called *saddle points* (see Fig. 1.6. below).

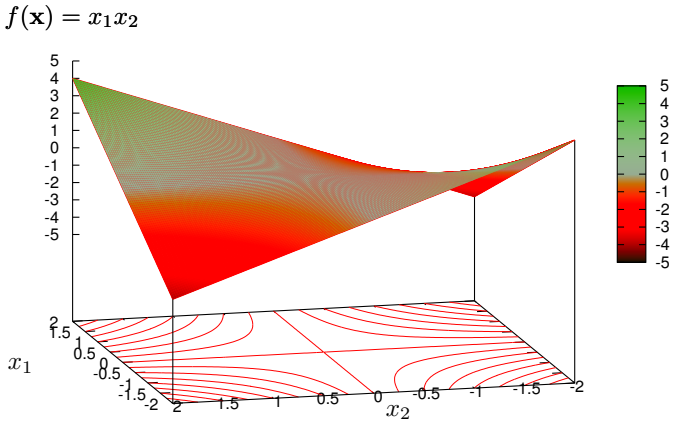


Figure 1.6. Illustration of the objective function in Example 1.26.

The conditions presented in Theorems 1.22 and 1.25 are necessary conditions. That is, they must hold true at every local optimal solution. Yet, a point satisfying these conditions need not be a local minimum. The following theorem gives sufficient conditions for a stationary point to be a *global* minimum point, provided the objective function is convex on $\mathbb{R}^{n_x}$.

Theorem 1.27 (First-Order Sufficient Conditions for a Strict Local Minimum). *Suppose that $f : \mathbb{R}^{n_x} \to \mathbb{R}$ is differentiable at $\mathbf{x}^\star$ and convex on $\mathbb{R}^{n_x}$. If $\nabla f(\mathbf{x}^\star) = \mathbf{0}$, then $\mathbf{x}^\star$ is a global minimum of f on $\mathbb{R}^{n_x}$.*

Proof. f being convex on $\mathbb{R}^{n_x}$ and differentiable at $\mathbf{x}^\star$, by Theorem A.17, we have

$$f(\mathbf{x}) \geq f(\mathbf{x}^\star) + \nabla f(\mathbf{x}^\star)^{\mathsf{T}}[\mathbf{x} - \mathbf{x}^\star] \quad \forall \mathbf{x} \in \mathbb{R}^{n_x}.$$

But since $\mathbf{x}^\star$ is a stationary point,

$$f(\mathbf{x}) \geq f(\mathbf{x}^\star) \quad \forall \mathbf{x} \in \mathbb{R}^{n_x}.$$

□

The convexity condition required by the foregoing theorem is actually very restrictive, in the sense that many practical problems are nonconvex. In the following theorem, we give sufficient conditions for characterizing a local minimum point, provided the objective function is strictly convex in a neighborhood of that point.

Theorem 1.28 (Second-Order Sufficient Conditions for a Strict Local Minimum). *Suppose that $f : \mathbb{R}^{n_x} \to \mathbb{R}$ is twice differentiable at $\mathbf{x}^\star$. If $\nabla f(\mathbf{x}^\star) = \mathbf{0}$ and $\mathsf{H}(\mathbf{x}^\star)$ is positive definite, then $\mathbf{x}^\star$ is a local minimum of f.*

Proof. f being twice differentiable at $\mathbf{x}^\star$, we have

$$f(\mathbf{x}^\star + \mathbf{d}) \;=\; f(\mathbf{x}^\star) + \nabla f(\mathbf{x}^\star)^\mathsf{T}\mathbf{d} + \frac{1}{2}\mathbf{d}^\mathsf{T}\mathsf{H}(\mathbf{x}^\star)\mathbf{d} + \|\mathbf{d}\|^2\alpha(\mathbf{d}),$$

for each $\mathbf{d} \in \mathbb{R}^{n_x}$, where $\lim_{\mathbf{d}\to\mathbf{0}} \alpha(\mathbf{d}) = 0$. Let λ^L denote the smallest eigenvalue of $\mathsf{H}(\mathbf{x}^\star)$. Then, $\mathsf{H}(\mathbf{x}^\star)$ being positive definite we have $\lambda^L > 0$, and $\mathbf{d}^\mathsf{T}\mathsf{H}(\mathbf{x}^\star)\mathbf{d} \;\geq\; \lambda^L\|\mathbf{d}\|^2$. Moreover, from $\nabla f(\mathbf{x}^\star) = \mathbf{0}$, we get

$$f(\mathbf{x}^\star + \mathbf{d}) - f(\mathbf{x}^\star) \;\geq\; \left[\frac{\lambda}{2} + \alpha(\mathbf{d})\right]\|\mathbf{d}\|^2.$$

Since $\lim_{\mathbf{d}\to\mathbf{0}} \alpha(\mathbf{d}) = 0$,

$$\exists \eta > 0 \text{ such that } |\alpha(\mathbf{d})| < \frac{\lambda}{4} \quad \forall \mathbf{d} \in \mathcal{B}_\eta(\mathbf{0}),$$

and finally,

$$f(\mathbf{x}^\star + \mathbf{d}) - f(\mathbf{x}^\star) \;\geq\; \frac{\lambda}{4}\|\mathbf{d}\|^2 \;>\; 0 \quad \forall \mathbf{d} \in \mathcal{B}_\eta(\mathbf{0}) \setminus \{\mathbf{0}\},$$

i.e., $\mathbf{x}^\star$ is a strict local minimum of f. □

Example 1.29. Consider the problem

$$\min_{\mathbf{x}\in\mathbb{R}^2} \quad (x_1 - 1)^2 - x_1x_2 + (x_2 - 1)^2.$$

The gradient vector and Hessian matrix at $\bar{\mathbf{x}} = (2, 2)$ are given by

$$\begin{aligned}\nabla f(\bar{\mathbf{x}}) \;&=\; \begin{bmatrix} 2(\bar{x}_1 - 1) - \bar{x}_2 & 2(\bar{x}_2 - 1) - \bar{x}_1 \end{bmatrix}^\mathsf{T} \;=\; \mathbf{0} \\ \mathsf{H}(\bar{\mathbf{x}}) \;&=\; \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \succ \mathbf{0}\end{aligned}$$

Hence, by Theorem 1.25, $\bar{\mathbf{x}}$ is a local minimum of f. ($\bar{\mathbf{x}}$ is also a global minimum of f on $\mathbb{R}^2$ since f is convex.) The objective function is pictured in Fig. 1.7. below.

We close this subsection by reemphasizing the fact that every local minimum of an unconstrained optimization problem $\min\{f(\mathbf{x} : \mathbf{x} \in \mathbb{R}^{n_x}\}$ is a global minimum if f is a convex function on $\mathbb{R}^{n_x}$ (see Theorem 1.18). Yet, convexity of f is *not* a necessary condition for each local minimum to be a global minimum. As just an example, consider the function $x \mapsto \exp\left(-\frac{1}{x^2}\right)$ (see Fig 1.8.). In fact, such functions are said to be *pseudoconvex*.

$f(\mathbf{x}) = (x_1 - 1)^2 - x_1 x_2 + (x_2 - 1)^2$

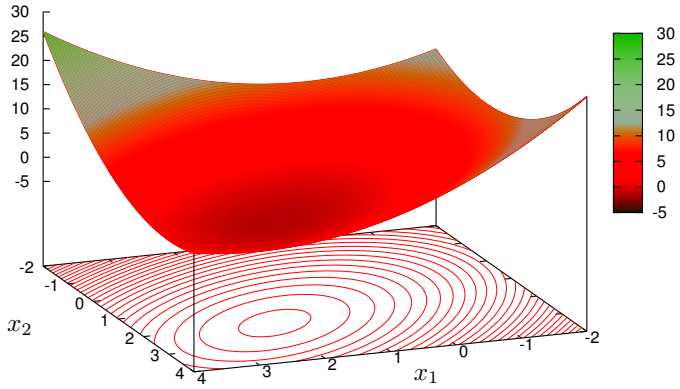


Figure 1.7. Illustration of the objective function in Example 1.29.

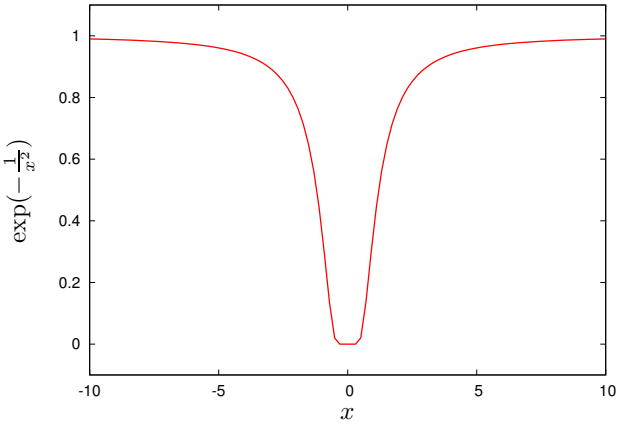


Figure 1.8. Plot of the pseudoconvex function $x \mapsto \exp\left(-\frac{1}{x^2}\right)$.

1.5 PROBLEMS WITH INEQUALITY CONSTRAINTS

In practice, few problems can be formulated as unconstrained programs. This is because the feasible region is generally restricted by imposing constraints on the optimization variables.

In this section, we first present theoretical results for the problem to:

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & \mathbf{x} \in S, \end{aligned}$$

for a general set S (geometric optimality conditions). Then, we let S be more specifically defined as the feasible region of a NLP of the form to minimize $f(\mathbf{x})$, subject to $\mathbf{g}(\mathbf{x}) \leq 0$ and $\mathbf{x} \in X$, and derive the Karush-Kuhn-Tucker (KKT) conditions of optimality.

1.5.1 Geometric Optimality Conditions

Definition 1.30 (Feasible Direction). *Let S be a nonempty set in $\mathbb{R}^{n_x}$. A vector $\mathbf{d} \in \mathbb{R}^{n_x}$, $\mathbf{d} \neq \mathbf{0}$, is said to be a* feasible direction *at $\bar{\mathbf{x}} \in \mathrm{cl}\,(S)$ if*

$$\exists \delta > 0 \text{ such that } \bar{\mathbf{x}} + \eta \mathbf{d} \in S \quad \forall \eta \in (0, \delta).$$

Moreover, the cone of feasible directions *at $\bar{\mathbf{x}}$, denoted by $\mathcal{D}(\bar{\mathbf{x}})$, is given by*

$$\mathcal{D}(\bar{\mathbf{x}}) := \{\mathbf{d} \neq \mathbf{0} : \exists \delta > 0 \text{ such that } \bar{\mathbf{x}} + \eta \mathbf{d} \in S \quad \forall \eta \in (0, \delta)\}.$$

From the above definition and Lemma 1.21, it is clear that a small movement from $\bar{\mathbf{x}}$ along a direction $\mathbf{d} \in \mathcal{D}(\bar{\mathbf{x}})$ leads to feasible points, whereas a similar movement along a direction $\mathbf{d} \in \mathcal{F}_0(\bar{\mathbf{x}})$ leads to solutions of improving objective value (see Definition 1.20). As shown in Theorem 1.31 below, a (geometric) necessary condition for local optimality is that: *"Every improving direction is not a feasible direction."* This fact is illustrated in Fig. 1.9., where both the half-space $\mathcal{F}_0(\bar{\mathbf{x}})$ and the cone $\mathcal{D}(\bar{\mathbf{x}})$ (see Definition A.10) are translated from the origin to $\bar{\mathbf{x}}$ for clarity.

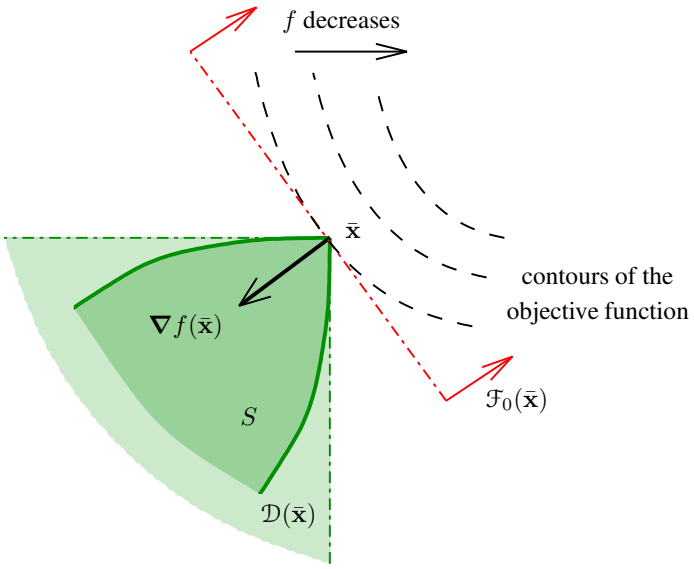


Figure 1.9. Illustration of the (geometric) necessary condition $\mathcal{F}_0(\bar{\mathbf{x}}) \cap \mathcal{D}(\bar{\mathbf{x}}) = \emptyset$.

Theorem 1.31 (Geometric Necessary Condition for a Local Minimum). *Let S be a nonempty set in $\mathbb{R}^{n_x}$, and let $f : \mathbb{R}^{n_x} \to \mathbb{R}$ be a differentiable function. Suppose that $\bar{\mathbf{x}}$ is a local minimizer of the problem to minimize $f(\mathbf{x})$ subject to $\mathbf{x} \in S$. Then, $\mathcal{F}_0(\bar{\mathbf{x}}) \cap \mathcal{D}(\bar{\mathbf{x}}) = \emptyset$.*

Proof. By contradiction, suppose that there exists a vector $\mathbf{d} \in \mathcal{F}_0(\bar{\mathbf{x}}) \cap \mathcal{D}(\bar{\mathbf{x}})$, $\mathbf{d} \neq \mathbf{0}$. Then, by Lemma 1.21,

$$\exists \delta_1 > 0 \text{ such that } f(\bar{\mathbf{x}} + \eta \mathbf{d}) < f(\bar{\mathbf{x}}) \quad \forall \eta \in (0, \delta_1).$$

Moreover, by Definition 1.30,

$$\exists \delta_2 > 0 \text{ such that } \bar{\mathbf{x}} + \eta \mathbf{d} \in S \quad \forall \eta \in (0, \delta_2).$$

Hence,

$$\exists \mathbf{x} \in \mathcal{B}_\eta(\bar{\mathbf{x}}) \cap S \text{ such that } f(\bar{\mathbf{x}} + \eta \mathbf{d}) < f(\bar{\mathbf{x}}),$$

for every $\eta \in (0, \min\{\delta_1, \delta_2\})$, which contradicts the assumption that $\bar{\mathbf{x}}$ is a local minimum of f on S (see Definition 1.9). □

1.5.2 KKT Conditions

We now specify the feasible region as

$$S := \{\mathbf{x} : g_i(\mathbf{x}) \leq 0 \quad \forall i = 1, \ldots, n_g\},$$

where $g_i : \mathbb{R}^{n_x} \to \mathbb{R}$, $i = 1, \ldots, n_g$, are continuous functions. In the geometric optimality condition given by Theorem 1.31, $\mathcal{D}(\bar{\mathbf{x}})$ is the cone of feasible directions. From a practical viewpoint, it is desirable to convert this geometric condition into a more usable condition involving algebraic equations. As Lemma 1.33 below indicates, we can define a cone $\mathcal{D}_0(\bar{\mathbf{x}})$ in terms of the gradients of the *active constraints* at $\bar{\mathbf{x}}$, such that $\mathcal{D}_0(\bar{\mathbf{x}}) \subseteq \mathcal{D}(\bar{\mathbf{x}})$. For this, we need the following:

Definition 1.32 (Active Constraint, Active Set). *Let $g_i : \mathbb{R}^{n_x} \to \mathbb{R}$, $i = 1, \ldots, n_g$, and consider the set $S := \{\mathbf{x} : g_i(\mathbf{x}) \leq 0, i = 1, \ldots, n_g\}$. Let $\bar{\mathbf{x}} \in S$ be a feasible point. For each $i = 1, \ldots, n_g$, the constraint g_i is said to* active *or* binding *at $\bar{\mathbf{x}}$ if $g_i(\bar{\mathbf{x}}) = 0$; it is said to be* inactive *at $\bar{\mathbf{x}}$ if $g_i(\bar{\mathbf{x}}) < 0$. Moreover,*

$$\mathcal{A}(\bar{\mathbf{x}}) \;:=\; \{i : g_i(\bar{\mathbf{x}}) = 0\},$$

denotes the set of active constraints *at $\bar{\mathbf{x}}$.*

Lemma 1.33 (Algebraic Characterization of a Feasible Direction). *Let $g_i : \mathbb{R}^{n_x} \to \mathbb{R}$, $i = 1, \ldots, n_g$ be differentiable functions, and consider the set $S := \{\mathbf{x} : g_i(\mathbf{x}) \leq 0, i = 1, \ldots, n_g\}$. For any feasible point $\bar{\mathbf{x}} \in S$, we have*

$$\mathcal{D}_0(\bar{\mathbf{x}}) \;:=\; \{\mathbf{d} : \nabla g_i(\bar{\mathbf{x}})^\mathsf{T} \mathbf{d} < 0 \quad \forall i \in \mathcal{A}(\bar{\mathbf{x}})\} \;\subseteq\; \mathcal{D}(\bar{\mathbf{x}}).$$

Proof. Suppose $\mathcal{D}_0(\bar{\mathbf{x}})$ is nonempty, and let $\mathbf{d} \in \mathcal{D}_0(\bar{\mathbf{x}})$. Since $\nabla g_i(\bar{\mathbf{x}})\mathbf{d} < 0$ for each $i \in \mathcal{A}(\bar{\mathbf{x}})$, then by Lemma 1.21, $\mathbf{d}$ is a descent direction for g_i at $\bar{\mathbf{x}}$, i.e.,

$$\exists \delta_2 > 0 \text{ such that } g_i(\bar{\mathbf{x}} + \eta \mathbf{d}) < g_i(\bar{\mathbf{x}}) = 0 \quad \forall \eta \in (0, \delta_2),\ \forall i \in \mathcal{A}(\bar{\mathbf{x}}).$$

Furthermore, since $g_i(\bar{\mathbf{x}}) < 0$ and g_i is continuous at $\bar{\mathbf{x}}$ (for it is differentiable) for each $i \notin \mathcal{A}(\bar{\mathbf{x}})$,

$$\exists \delta_1 > 0 \text{ such that } g_i(\bar{\mathbf{x}} + \eta \mathbf{d}) < 0 \quad \forall \eta \in (0, \delta_1),\ \forall i \notin \mathcal{A}(\bar{\mathbf{x}}).$$

Furthermore, Overall, it is clear that the points $\bar{\mathbf{x}} + \eta \mathbf{d}$ are in S for all $\eta \in (0, \min\{\delta_1, \delta_2\})$. Hence, by Definition 1.30, $\mathbf{d} \in \mathcal{D}(\bar{\mathbf{x}})$. □

Remark 1.34. This lemma together with Theorem 1.31 directly leads to the result that $\mathcal{F}_0(\bar{\mathbf{x}}) \cap \mathcal{D}_0(\bar{\mathbf{x}}) = \emptyset$ for any local solution point $\bar{\mathbf{x}}$, i.e.,

$$\arg\min\{f(\mathbf{x}) : \mathbf{x} \in S\} \subset \{\mathbf{x} \in \mathbb{R}^{n_x} : \mathcal{F}_0(\bar{\mathbf{x}}) \cap \mathcal{D}_0(\bar{\mathbf{x}}) = \emptyset\}.$$

The foregoing geometric characterization of local solution points applies equally well to either interior points $\text{int}(S) := \{\mathbf{x} \in \mathbb{R}^{n_x} : g_i(\mathbf{x}) < 0, \forall i = 1, \ldots, n_g\}$, or boundary

points being at the boundary of the feasible domain. At an interior point, in particular, any direction is feasible, and the necessary condition $\mathcal{F}_0(\bar{\mathbf{x}}) \cap \mathcal{D}_0(\bar{\mathbf{x}}) = \emptyset$ reduces to $\boldsymbol{\nabla} f(\bar{\mathbf{x}}) = \mathbf{0}$, which gives the same condition as in unconstrained optimization (see Theorem 1.22).

Note also that there are several cases where the condition $\mathcal{F}_0(\bar{\mathbf{x}}) \cap \mathcal{D}_0(\bar{\mathbf{x}}) = \emptyset$ is satisfied by non-optimal points. In other words, this condition is necessary but *not sufficient* for a point $\bar{\mathbf{x}}$ to be a local minimum of f on S. For instance, any point $\bar{\mathbf{x}}$ with $\boldsymbol{\nabla} g_i(\bar{\mathbf{x}}) = 0$ for some $i \in \mathcal{A}(\bar{\mathbf{x}})$ trivially satisfies the condition $\mathcal{F}_0(\bar{\mathbf{x}}) \cap \mathcal{D}_0(\bar{\mathbf{x}}) = \emptyset$. Another example is given below.

Example 1.35. Consider the problem

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^2} \; & f(\mathbf{x}) := x_1^2 + x_2^2 \\ \text{s.t.} \; & g_1(\mathbf{x}) := x_1 \leq 0 \\ & g_2(\mathbf{x}) := -x_1 \leq 0. \end{aligned} \tag{1.6}$$

Clearly, this problem is convex and $\mathbf{x}^\star = (0,0)^\mathsf{T}$ is the unique global minimum.

Now, let $\bar{\mathbf{x}}$ be any point on the line $\mathcal{C} := \{\mathbf{x} : x_1 = 0\}$. Both constraints g_1 and g_2 are active at $\bar{\mathbf{x}}$, and we have $\boldsymbol{\nabla} g_1(\bar{\mathbf{x}}) = -\boldsymbol{\nabla} g_2(\bar{\mathbf{x}}) = (1,0)^\mathsf{T}$. Therefore, no direction $\mathbf{d} \neq \mathbf{0}$ can be found such that $\boldsymbol{\nabla} g_1(\bar{\mathbf{x}})^\mathsf{T} \mathbf{d} < 0$ and $\boldsymbol{\nabla} g_2(\bar{\mathbf{x}})^\mathsf{T} \mathbf{d} < 0$ simultaneously, i.e., $\mathcal{D}_0(\bar{\mathbf{x}}) = \emptyset$. In turn, this implies that $\mathcal{F}_0(\bar{\mathbf{x}}) \cap \mathcal{D}_0(\bar{\mathbf{x}}) = \emptyset$ is trivially satisfied for any point on $\mathcal{C}$.

On the other hand, observe that the condition $\mathcal{F}_0(\bar{\mathbf{x}}) \cap \mathcal{D}(\bar{\mathbf{x}}) = \emptyset$ in Theorem 1.31 excludes all the points on $\mathcal{C}$, but the origin, since a feasible direction at $\bar{\mathbf{x}}$ is given, e.g., by $\mathbf{d} = (0,1)^\mathsf{T}$.

Next, we reduce the geometric necessary optimality condition $\mathcal{F}_0(\bar{\mathbf{x}}) \cap \mathcal{D}_0(\bar{\mathbf{x}}) = \emptyset$ to a statement in terms of the gradients of the objective function and of the active constraints. The resulting first-order optimality conditions are known as the *Karush-Kuhn-Tucker (KKT) necessary conditions*. Beforehand, we introduce the important concepts of a *regular point* and of a *KKT point*.

Definition 1.36 (Regular Point (for a Set of Inequality Constraints)). *Let $g_i : \mathbb{R}^{n_x} \to \mathbb{R}$, $i = 1, \ldots, n_g$, be differentiable functions on $\mathbb{R}^{n_x}$ and consider the set $S := \{\mathbf{x} \in \mathbb{R}^{n_x} : g_i(\mathbf{x}) \leq 0, i = 1, \ldots, n_g\}$. A point $\bar{\mathbf{x}} \in S$ is said to be a* regular point *if the gradient vectors $\boldsymbol{\nabla} g_i(\bar{\mathbf{x}})$, $i \in \mathcal{A}(\bar{\mathbf{x}})$, are linearly independent,*

$$\operatorname{rank} \left(\boldsymbol{\nabla} g_i(\bar{\mathbf{x}}), i \in \mathcal{A}(\bar{\mathbf{x}}) \right) \; = \; |\mathcal{A}(\bar{\mathbf{x}})|.$$

Definition 1.37 (KKT Point). *Let $f : \mathbb{R}^{n_x} \to \mathbb{R}$ and $g_i : \mathbb{R}^{n_x} \to \mathbb{R}$, $i = 1, \ldots, n_g$ be differentiable functions. Consider the problem to minimize $f(\mathbf{x})$ subject to $\mathbf{g}(\mathbf{x}) \leq \mathbf{0}$. If a point $(\bar{\mathbf{x}}, \bar{\boldsymbol{\nu}}) \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_g}$ satisfies the conditions:*

$$\boldsymbol{\nabla} f(\bar{\mathbf{x}}) + \bar{\boldsymbol{\nu}}^\mathsf{T} \boldsymbol{\nabla} \mathbf{g}(\bar{\mathbf{x}}) = \mathbf{0} \tag{1.7}$$

$$\bar{\boldsymbol{\nu}} \geq \mathbf{0} \tag{1.8}$$

$$\mathbf{g}(\bar{\mathbf{x}}) \leq \mathbf{0} \tag{1.9}$$

$$\bar{\boldsymbol{\nu}}^\mathsf{T} \mathbf{g}(\bar{\mathbf{x}}) = 0, \tag{1.10}$$

then $(\bar{\mathbf{x}}, \bar{\boldsymbol{\nu}})$ *is said to be a* KKT point.

Remark 1.38. The scalars ν_i, $i = 1, \ldots, n_g$, are called the *Lagrange multipliers*. The condition (1.7), i.e., the requirement that $\bar{\mathbf{x}}$ be feasible, is called the *primal feasibility* (PF) condition; the conditions (1.8) and (1.9) are referred to as the *dual feasibility* (DF) conditions; finally, the condition (1.10) is called the *complementarity slackness*[4] (CS) condition.

Theorem 1.39 (KKT Necessary Conditions). *Let* $f : \mathbb{R}^{n_x} \to \mathbb{R}$ *and* $g_i : \mathbb{R}^{n_x} \to \mathbb{R}$, $i = 1, \ldots, n_g$ *be differentiable functions. Consider the problem to minimize* $f(\mathbf{x})$ *subject to* $\mathbf{g}(\mathbf{x}) \leq \mathbf{0}$. *If* $\mathbf{x}^\star$ *is a local minimum and a regular point of the constraints, then there exists a* unique *vector* $\boldsymbol{\nu}^\star$ *such that* $(\mathbf{x}^\star, \boldsymbol{\nu}^\star)$ *is a KKT point.*

Proof. Since $\mathbf{x}^\star$ solves the problem, then there exists no direction $\mathbf{d} \in \mathbb{R}^{n_x}$ such that $\boldsymbol{\nabla} f(\bar{\mathbf{x}})^\mathsf{T}\mathbf{d} < 0$ and $\boldsymbol{\nabla} g_i(\bar{\mathbf{x}})^\mathsf{T}\mathbf{d} < 0$, $\forall i \in \mathcal{A}(\mathbf{x}^\star)$ simultaneously (see Remark 1.34). Let $\mathsf{A} \in \mathbb{R}^{(|\mathcal{A}(\mathbf{x}^\star)|+1)\times n_x}$ be the matrix whose rows are $\boldsymbol{\nabla} f(\bar{\mathbf{x}})^\mathsf{T}$ and $\boldsymbol{\nabla} g_i(\bar{\mathbf{x}})^\mathsf{T}$, $i \in \mathcal{A}(\mathbf{x}^\star)$. Clearly, the statement $\{\exists \mathbf{d} \in \mathbb{R}^{n_x} : \mathsf{A}\mathbf{d} < \mathbf{0}\}$ is false, and by Gordan's Theorem 1.A.78, there exists a nonzero vector $\mathbf{p} \geq \mathbf{0}$ in $\mathbb{R}^{|\mathcal{A}(\mathbf{x}^\star)|+1}$ such that $\mathsf{A}^\mathsf{T}\mathbf{p} = \mathbf{0}$. Denoting the components of $\mathbf{p}$ by u_0 and u_i for $i \in \mathcal{A}(\mathbf{x}^\star)$, we get:

$$u_0 \boldsymbol{\nabla} f(\mathbf{x}^\star) + \sum_{i \in \mathcal{A}(\mathbf{x}^\star)} u_i \boldsymbol{\nabla} g_i(\mathbf{x}^\star) = \mathbf{0}$$

where $u_0 \geq 0$ and $u_i \geq 0$ for $i \in \mathcal{A}(\mathbf{x}^\star)$, and $(u_0, \mathbf{u}_{\mathcal{A}(\mathbf{x}^\star)}) \neq (0, \mathbf{0})$ (here $\mathbf{u}_{\mathcal{A}(\mathbf{x}^\star)}$ is the vector whose components are the u_i's for $i \in \mathcal{A}(\mathbf{x}^\star)$). Letting $u_i = 0$ for $i \notin \mathcal{A}(\mathbf{x}^\star)$, we then get the conditions:

$$\begin{aligned} u_0 \boldsymbol{\nabla} f(\mathbf{x}^\star) + \mathbf{u}^\mathsf{T} \boldsymbol{\nabla} \mathbf{g}(\mathbf{x}^\star) &= \mathbf{0} \\ \mathbf{u}^\mathsf{T} \mathbf{g}(\mathbf{x}^\star) &= \mathbf{0} \\ u_0, \mathbf{u} &\geq 0 \\ (u_0, \mathbf{u}) &\neq (0, \mathbf{0}), \end{aligned}$$

where $\mathbf{u}$ is the vector whose components are u_i for $i = 1, \ldots, n_g$. Note that $u_0 \neq 0$, for otherwise the assumption of linear independence of the active constraints at $\mathbf{x}^\star$ would be violated. Then, letting $\boldsymbol{\nu}^\star = \frac{1}{u_0}\mathbf{u}$, we obtain that $(\mathbf{x}^\star, \boldsymbol{\nu}^\star)$ is a KKT point. □

Remark 1.40. One of the major difficulties in applying the foregoing result is that we do not know *a priori* which constraints are active and which constraints are inactive, i.e., the active set is *unknown*. Therefore, it is necessary to investigate *all* possible active sets for finding candidate points satisfying the KKT conditions. This is illustrated in Example 1.41 below.

Example 1.41 (Regular Case). Consider the problem

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^3} \; & f(\mathbf{x}) := \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) \\ \text{s.t.} \; & g_1(\mathbf{x}) := x_1 + x_2 + x_3 + 3 \leq 0 \\ & g_2(\mathbf{x}) := x_1 \leq 0. \end{aligned} \tag{1.11}$$

[4] Often, the condition (1.10) is replaced by the equivalent conditions:

$$\bar{\nu}_i g_i(\bar{\mathbf{x}}) = 0 \quad \text{for } i = 1, \ldots, n_g.$$

Note that every feasible point is regular (the point (0,0,0) being infeasible), so $\mathbf{x}^\star$ must satisfy the dual feasibility conditions:

$$\begin{aligned} x_1^\star + \nu_1^\star + \nu_2^\star &= 0 \\ x_2^\star + \nu_1^\star &= 0 \\ x_3^\star + \nu_1^\star &= 0. \end{aligned}$$

Four cases can be distinguished:

(i) The constraints g_1 and g_2 are both *inactive*, i.e., $x_1^\star + x_2^\star + x_3^\star < -3$, $x_1^\star < 0$, and $\nu_1^\star = \nu_2^\star = 0$. From the latter together with the dual feasibility conditions, we get $x_1^\star = x_2^\star = x_3^\star = 0$, hence contradicting the former.

(ii) The constraint g_1 is *inactive*, while g_2 is *active*, i.e., $x_1^\star + x_2^\star + x_3^\star < -3$, $x_1^\star = 0$, $\nu_2^\star \geq 0$, and $\nu_1^\star = 0$. From the latter together with the dual feasibility conditions, we get $x_2^\star = x_3^\star = 0$, hence contradicting the former once again.

(iii) The constraint g_1 is *active*, while g_2 is *inactive*, i.e., $x_1^\star + x_2^\star + x_3^\star = -3$, $x_1^\star < 0$, and $\nu_1^\star \geq 0$, and $\nu_2^\star = 0$. Then, the point $(\mathbf{x}^\star, \boldsymbol{\nu}^\star)$ such that ${x_1}^\star = {x_2}^\star = {x_3}^\star = -1$, $\nu_1^\star = 1$ and $\nu_2^\star = 0$ is a KKT point.

(iv) The constraints g_1 and g_2 are both *active*, i.e., $x_1^\star + x_2^\star + x_3^\star = -3$, $x_1^\star = 0$, and $\nu_1^\star, \nu_2^\star > 0$. Then, we obtain ${x_2}^\star = {x_3}^\star = -\frac{3}{2}$, $\nu_1^\star = \frac{3}{2}$, and $\nu_2^\star = -\frac{3}{2}$, hence contradicting the dual feasibility condition $\nu_2^\star \geq 0$.

Overall, there is a *unique* candidate for a local minimum. Yet, it cannot be concluded as to whether this point is actually a global minimum, or even a local minimum, of (1.11). This question will be addressed later on in Example 1.45.

Remark 1.42 (Constraint Qualification). It is *very* important to note that for a local minimum $\mathbf{x}^\star$ to be a KKT point, an additional condition must be placed on the behavior of the constraint, i.e., **not every local minimum is a KKT point**; such a condition is known as a *constraint qualification*. In Theorem 1.39, it is shown that one possible constraint qualification is that $\mathbf{x}^\star$ be a regular point, which is the well known *linear independence constraint qualification* (LICQ). A weaker constraint qualification (i.e., implied by LICQ) known as the *Mangasarian-Fromovitz constraint qualification* (MFCQ) requires that there exits (at least) one direction $\mathbf{d} \in \mathcal{D}_0(\mathbf{x}^\star)$, i.e., such that $\boldsymbol{\nabla} g_i(\mathbf{x}^\star)^\mathsf{T}\mathbf{d} < 0$, for each $i \in \mathcal{A}(\mathbf{x}^\star)$. Note, however, that the Lagrange multipliers are guaranteed to be unique if LICQ holds (as stated in Theorem 1.39), while this uniqueness property may be lost under MFCQ.

The following example illustrates the necessity of having a constraint qualification for a KKT point to be a local minimum point of an NLP.

Example 1.43 (Non-Regular Case). Consider the problem

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^2} \; & f(\mathbf{x}) := -x_1 \\ \text{s.t.} \; & g_1(\mathbf{x}) := x_2 - (1-x_1)^3 \leq 0 \\ & g_2(\mathbf{x}) := -x_2 \leq 0. \end{aligned} \tag{1.12}$$

The feasible region is shown in Fig. 1.10. below. Note that a minimum point of (1.12) is $\mathbf{x}^\star = (1,0)^\mathsf{T}$. The dual feasibility condition relative to variable x_1 reads:

$$-1 + 3\nu_1(1-x_1)^2 \;=\; 0.$$

It is readily seen that this condition cannot be met at any point on the straight line $\mathcal{C} := \{\mathbf{x} : x_1 = 1\}$, including the minimum point $\mathbf{x}^\star$. In other words, the KKT conditions are not necessary in this example. This is because no constraint qualification can hold at $\mathbf{x}^\star$. In particular, $\mathbf{x}^\star$ not being a regular point, LICQ does not hold; moreover, the set $\mathcal{D}_0(\mathbf{x}^\star)$ being empty (the direction $\mathbf{d} = (-1,0)^\mathsf{T}$ gives $\nabla g_1(\mathbf{x}^\star)^\mathsf{T}\mathbf{d} = \nabla g_2(\mathbf{x}^\star)^\mathsf{T}\mathbf{d} = 0$, while any other direction induces a violation of either one of the constraints), MFCQ does not hold at $\mathbf{x}^\star$ either.

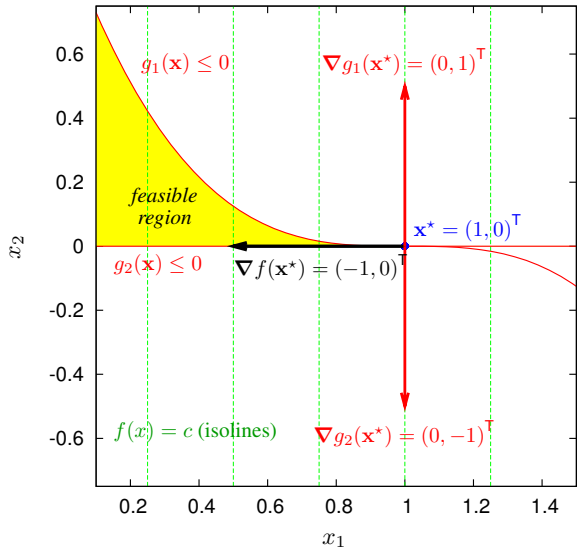


Figure 1.10. Solution of Example 1.43.

The following theorem provides a sufficient condition under which any KKT point of an inequality constrained NLP problem is guaranteed to be a global minimum of that problem.

Theorem 1.44 (KKT sufficient Conditions). *Let* $f : \mathbb{R}^{n_x} \to \mathbb{R}$ *and* $g_i : \mathbb{R}^{n_x} \to \mathbb{R}$, $i = 1, \ldots, n_g$, *be convex and differentiable functions. Consider the problem to minimize* $f(\mathbf{x})$ *subject to* $\mathbf{g}(\mathbf{x}) \le \mathbf{0}$. *If* $(\mathbf{x}^\star, \boldsymbol{\nu}^\star)$ *is a KKT point, then* $\mathbf{x}^\star$ *is a global minimum of that problem.*

Proof. Consider the function $\mathcal{L}(\mathbf{x}) := f(\mathbf{x}) + \sum_{i=1}^{n_g} \nu_i^\star g_i(\mathbf{x})$. Since f and g_i, $i = 1, \ldots, n_g$, are convex functions, and $\nu_i \ge 0$, $\mathcal{L}$ is also convex. Moreover, the dual feasibility conditions impose that we have $\nabla\mathcal{L}(\mathbf{x}^\star) = \mathbf{0}$. Hence, by Theorem 1.27, $\mathbf{x}^\star$ is a global minimizer for $\mathcal{L}$ on $\mathbb{R}^{n_x}$, i.e.,

$$\mathcal{L}(\mathbf{x}) \;\ge\; \mathcal{L}(\mathbf{x}^\star) \quad \forall \mathbf{x} \in \mathbb{R}^{n_x}.$$

In particular, for each $\mathbf{x}$ such that $g_i(\mathbf{x}) \leq g_i(\mathbf{x}^\star) = 0$, $i \in \mathcal{A}(\mathbf{x}^\star)$, we have

$$f(\mathbf{x}) - f(\mathbf{x}^\star) \;\geq\; -\sum_{i \in \mathcal{A}(\mathbf{x}^\star)} \mu_i^\star \left[g_i(\mathbf{x}) - g_i(\mathbf{x}^\star) \right] \;\geq\; 0.$$

Noting that $\{\mathbf{x} \in \mathbb{R}^{n_x} : g_i(\mathbf{x}) \leq 0, i \in \mathcal{A}(\mathbf{x}^\star)\}$ contains the feasible domain $\{\mathbf{x} \in \mathbb{R}^{n_x} : g_i(\mathbf{x}) \leq 0, i = 1, \ldots, n_g\}$, we therefore showed that $\mathbf{x}^\star$ is a global minimizer for the problem. □

Example 1.45. Consider the same Problem (1.11) as in Example 1.41 above. The point $(\mathbf{x}^\star, \boldsymbol{\nu}^\star)$ with ${x_1}^\star = {x_2}^\star = {x_3}^\star = -1$, $\nu_1^\star = 1$ and $\nu_2^\star = 0$, being a KKT point, and both the objective function and the feasible set being convex, by Theorem 1.44, $\mathbf{x}^\star$ is a global minimum for the Problem (1.11).

Both second-order necessary and sufficient conditions for inequality constrained NLP problems will be presented later on in §1.7.

1.6 PROBLEMS WITH EQUALITY CONSTRAINTS

In this section, we shall consider nonlinear programming problems with equality constraints of the form:

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & h_i(\mathbf{x}) \;=\; 0 \quad i = 1, \ldots, n_h. \end{array}$$

Based on the material presented in §1.5, it is tempting to convert this problem into an inequality constrained problem, by replacing each equality constraints $h_i(\mathbf{x}) = 0$ by two inequality constraints $h_i^+(\mathbf{x}) = h_i(\mathbf{x}) \leq 0$ and $h_i^-(\mathbf{x}) = -h(\mathbf{x}) \leq 0$. Given a feasible point $\bar{\mathbf{x}} \in \mathbb{R}^{n_x}$, we would have $h_i^+(\bar{\mathbf{x}}) = h_i^-(\bar{\mathbf{x}}) = 0$ and $\boldsymbol{\nabla} h_i^+(\bar{\mathbf{x}}) = -\boldsymbol{\nabla} h_i^-(\bar{\mathbf{x}})$. Therefore, there could exist no vector $\mathbf{d}$ such that $\boldsymbol{\nabla} h_i^+(\bar{\mathbf{x}}) < 0$ and $\boldsymbol{\nabla} h_i^-(\bar{\mathbf{x}}) < 0$ simultaneously, i.e., $\mathcal{D}_0(\bar{\mathbf{x}}) = \emptyset$. In other words, the geometric conditions developed in the previous section for inequality constrained problems are satisfied by all feasible solutions and, hence, are not informative (see Example 1.35 for an illustration). A different approach must therefore be used to deal with equality constrained problems. After a number of preliminary results in §1.6.1, we shall describe the method of Lagrange multipliers for equality constrained problems in §1.6.2.

1.6.1 Preliminaries

An equality constraint $h(\mathbf{x}) = 0$ defines a set on $\mathbb{R}^{n_x}$, which is best view as a hypersurface. When considering $n_h \geq 1$ equality constraints $h_1(\mathbf{x}), \ldots, h_{n_h}(\mathbf{x})$, their intersection forms a (possibly empty) set $S := \{\mathbf{x} \in \mathbb{R}^{n_x} : h_i(\mathbf{x}) = 0, i = 1, \ldots, n_h\}$.

Throughout this section, we shall assume that the equality constraints are differentiable; that is, the set $S := \{\mathbf{x} \in \mathbb{R}^{n_x} : h_i(\mathbf{x}) = 0, i = 1, \ldots, n_h\}$ is said to be *differentiable manifold* (or *smooth manifold*). Associated with a point on a differentiable manifold is the *tangent set* at that point. To formalize this notion, we start by defining *curves* on a manifold. A curve $\boldsymbol{\xi}$ on a manifold S is a continuous application $\boldsymbol{\xi} : \mathcal{I} \subset \mathbb{R} \to S$, i.e., a family of

points $\boldsymbol{\xi}(t) \in S$ continuously parameterized by t in an interval $\mathcal{I}$ of $\mathbb{R}$. A curve is said to pass through the point $\bar{\mathbf{x}}$ if $\bar{\mathbf{x}} = \boldsymbol{\xi}(\bar{t})$ for some $\bar{t} \in \mathcal{I}$; the *derivative* of a curve at $\bar{t}$, provided it exists, is defined as $\dot{\boldsymbol{\xi}}(\bar{t}) := \lim_{h\to 0} \frac{\boldsymbol{\xi}(\bar{t}+h)-\boldsymbol{\xi}(\bar{t})}{h}$. A curve is said to be *differentiable* (or *smooth*) if a derivative exists for each $t \in \mathcal{I}$.

Definition 1.46 (Tangent Set). *Let S be a (differentiable) manifold in $\mathbb{R}^{n_x}$, and let $\bar{\mathbf{x}} \in S$. Consider the collection of* all *the continuously differentiable curves on S passing through $\bar{\mathbf{x}}$. Then, the collection of* all *the vectors tangent to these curves at $\bar{\mathbf{x}}$ is said to be the* tangent set *to S at $\bar{\mathbf{x}}$, denoted by $\mathscr{T}(\bar{\mathbf{x}})$.*

If the constraints are *regular*, in the sense of Definition 1.47 below, then S is (locally) of dimension $n_x - n_h$, and $\mathscr{T}(\bar{\mathbf{x}})$ constitutes a subspace of dimension $n_x - n_h$, called the *tangent space*.

Definition 1.47 (Regular Point (for a Set of Equality Constraints)). *Let $h_i : \mathbb{R}^{n_x} \to \mathbb{R}$, $i = 1, \ldots, n_h$, be differentiable functions on $\mathbb{R}^{n_x}$ and consider the set $S := \{\mathbf{x} \in \mathbb{R}^{n_x} : h_i(\mathbf{x}) = 0, i = 1, \ldots, n_h\}$. A point $\bar{\mathbf{x}} \in S$ is said to be a* regular point *if the gradient vectors $\boldsymbol{\nabla} h_i(\bar{\mathbf{x}})$, $i = 1, \ldots, n_h$, are linearly independent, i.e.,*

$$\operatorname{rank}\begin{pmatrix} \boldsymbol{\nabla} h_1(\bar{\mathbf{x}}) & \boldsymbol{\nabla} h_2(\bar{\mathbf{x}}) & \cdots & \boldsymbol{\nabla} h_{n_h}(\bar{\mathbf{x}}) \end{pmatrix} = n_h.$$

Lemma 1.48 (Algebraic Characterization of a Tangent Space). *Let $h_i : \mathbb{R}^{n_x} \to \mathbb{R}$, $i = 1, \ldots, n_h$, be differentiable functions on $\mathbb{R}^{n_x}$ and consider the set $S := \{\mathbf{x} \in \mathbb{R}^{n_x} : h_i(\mathbf{x}) = 0, i = 1, \ldots, n_h\}$. At a regular point $\bar{\mathbf{x}} \in S$, the tangent space is such that*

$$\mathscr{T}(\bar{\mathbf{x}}) = \{\mathbf{d} : \boldsymbol{\nabla}\mathbf{h}(\bar{\mathbf{x}})^{\mathsf{T}}\mathbf{d} = \mathbf{0}\}.$$

Proof. The proof is technical and is omitted here (see, e.g., [36, §10.2]). □

1.6.2 The Method of Lagrange Multipliers

The idea behind the method of Lagrange multipliers for solving equality constrained NLP problems of the form

$$\begin{aligned} &\text{minimize} && f(\mathbf{x}) \\ &\text{subject to} && h_i(\mathbf{x}) \;=\; 0 \quad i = 1, \ldots, n_h. \end{aligned}$$

is to restrict the search of a minimum on the manifold $S := \{\mathbf{x} \in \mathbb{R}^{n_x} : h_i(\mathbf{x}) = 0, \forall i = 1, \ldots, n_h\}$. In other words, we derive optimality conditions by considering the value of the objective function along curves on the manifold S passing through the optimal point.

The following Theorem shows that the tangent space $\mathscr{T}(\bar{\mathbf{x}})$ at a regular (local) minimum point $\bar{\mathbf{x}}$ is orthogonal to the gradient of the objective function at $\bar{\mathbf{x}}$. This important fact is illustrated in Fig. 1.11. in the case of a single equality constraint.

Theorem 1.49 (Geometric Necessary Condition for a Local Minimum). *Let $f : \mathbb{R}^{n_x} \to \mathbb{R}$ and $h_i : \mathbb{R}^{n_x} \to \mathbb{R}$, $i = 1, \ldots, n_h$, be continuously differentiable functions on $\mathbb{R}^{n_x}$. Suppose that $\mathbf{x}^\star$ is a local minimum point of the problem to minimize $f(\mathbf{x})$ subject to the constraints $\mathbf{h}(\mathbf{x}) = \mathbf{0}$. Then, $\boldsymbol{\nabla} f(\mathbf{x}^\star)$ is orthogonal to the tangent space $\mathscr{T}(\mathbf{x}^\star)$,*

$$\mathcal{F}_0(\mathbf{x}^\star) \cap \mathscr{T}(\mathbf{x}^\star) \;=\; \emptyset.$$

Proof. By contradiction, assume that there exists a $\mathbf{d} \in \mathscr{T}(\mathbf{x}^\star)$ such that $\boldsymbol{\nabla} f(\mathbf{x}^\star)^{\mathsf{T}}\mathbf{d} \neq 0$. Let $\boldsymbol{\xi} : \mathcal{I} = [-a, a] \to S$, $a > 0$, be any smooth curve passing through $\mathbf{x}^\star$ with $\boldsymbol{\xi}(0) = \mathbf{x}^\star$

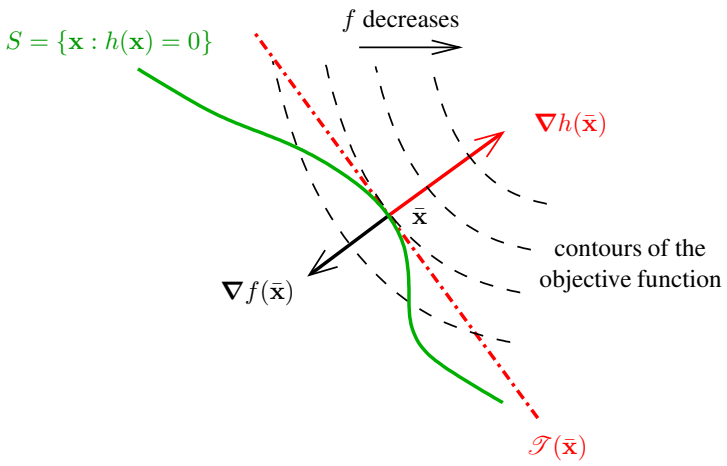


Figure 1.11. Illustration of the necessary conditions of optimality with equality constraints.

and $\dot{\boldsymbol{\xi}}(0) = \mathbf{d}$. Let also φ be the function defined as $\varphi(t) := f(\boldsymbol{\xi}(t))$, $\forall t \in \mathcal{I}$. Since $\mathbf{x}^\star$ is a local minimum of f on $S := \{\mathbf{x} \in \mathbb{R}^{n_x} : \mathbf{h}(\mathbf{x}) = \mathbf{0}\}$, by Definition 1.9, we have

$$\exists \eta > 0 \text{ such that } \varphi(t) = f(\boldsymbol{\xi}(t)) \geq f(\mathbf{x}^\star) = \varphi(0) \quad \forall t \in \mathcal{B}_\eta(0) \cap \mathcal{I}.$$

It follows that $t^\star = 0$ is an unconstrained (local) minimum point for φ, and

$$0 \;=\; \nabla\varphi(0) \;=\; \boldsymbol{\nabla} f(\mathbf{x}^\star)^\mathsf{T} \dot{\boldsymbol{\xi}}(0) \;=\; \boldsymbol{\nabla} f(\mathbf{x}^\star)^\mathsf{T} \mathbf{d}.$$

We thus get a contradiction with the fact that $\boldsymbol{\nabla} f(\mathbf{x}^\star)^\mathsf{T} \mathbf{d} \neq 0$. □

Next, we take advantage of the forgoing geometric characterization, and derive first-order necessary conditions for equality constrained NLP problems.

Theorem 1.50 (First-Order Necessary Conditions). *Let $f : \mathbb{R}^{n_x} \to \mathbb{R}$ and $h_i : \mathbb{R}^{n_x} \to \mathbb{R}$, $i = 1, \ldots, n_h$, be continuously differentiable functions on $\mathbb{R}^{n_x}$. Consider the problem to minimize $f(\mathbf{x})$ subject to the constraints $\mathbf{h}(\mathbf{x}) = \mathbf{0}$. If $\mathbf{x}^\star$ is a local minimum and is a regular point of the constraints, then there exists a unique vector $\boldsymbol{\lambda}^\star \in \mathbb{R}^{n_h}$ such that*

$$\boldsymbol{\nabla} f(\mathbf{x}^\star) + \boldsymbol{\nabla} \mathbf{h}(\mathbf{x}^\star)^\mathsf{T} \boldsymbol{\lambda}^\star = \mathbf{0}.$$

Proof. [5] Since $\mathbf{x}^\star$ is a local minimum of f on $S := \{\mathbf{x} \in \mathbb{R}^{n_x} : \mathbf{h}(\mathbf{x}) = \mathbf{0}\}$, by Theorem 1.49, we have $\mathcal{F}_0(\mathbf{x}^\star) \cap \mathscr{T}(\mathbf{x}^\star) = \emptyset$, i.e., the system

$$\boldsymbol{\nabla} f(\mathbf{x}^\star)^\mathsf{T} \mathbf{d} < 0 \qquad \boldsymbol{\nabla} \mathbf{h}(\mathbf{x}^\star)^\mathsf{T} \mathbf{d} = 0,$$

is inconsistent. Consider the following two sets:

$$\begin{aligned} C_1 &:= \{(z_1, \mathbf{z}_2) \in \mathbb{R}^{n_h+1} : z_1 = \boldsymbol{\nabla} f(\mathbf{x}^\star)^\mathsf{T} \mathbf{d}, \quad \mathbf{z}_2 = \boldsymbol{\nabla} \mathbf{h}(\mathbf{x}^\star)^\mathsf{T} \mathbf{d}\} \\ C_2 &:= \{(z_1, \mathbf{z}_2) \in \mathbb{R}^{n_h+1} : z_1 < 0, \quad \mathbf{z}_2 = \mathbf{0}.\} \end{aligned}$$

[5] See also in Appendix of §1 for an alternative proof of Theorem 1.50 that does not use the concept of tangent sets.

Clearly, C_1 and C_2 are convex, and $C_1 \cap C_2 = \emptyset$. Then, by the separation Theorem A.9, there exists a nonzero vector $(\mu, \boldsymbol{\lambda}) \in \mathbb{R}^{n_h+1}$ such that

$$\mu \boldsymbol{\nabla} f(\mathbf{x}^\star)^\mathsf{T} \mathbf{d} + \boldsymbol{\lambda}^\mathsf{T} [\boldsymbol{\nabla} \mathbf{h}(\mathbf{x}^\star)^\mathsf{T} \mathbf{d}] \;\geq\; \mu z_1 + \boldsymbol{\lambda}^\mathsf{T} \mathbf{z}_2 \qquad \forall \mathbf{d} \in \mathbb{R}^{n_x},\; \forall (z_1, \mathbf{z}_2) \in C_2.$$

Letting $\mathbf{z}_2 = \mathbf{0}$ and since z_1 can be made an arbitrarily large negative number, it follows that $\mu \geq 0$. Also, letting $(z_1, \mathbf{z}_2) = (0, \mathbf{0})$, we must have $[\mu \boldsymbol{\nabla} f(\mathbf{x}^\star) + \boldsymbol{\lambda}^\mathsf{T} \boldsymbol{\nabla} \mathbf{h}(\mathbf{x}^\star)]^\mathsf{T} \mathbf{d} \geq \mathbf{0}$, for each $\mathbf{d} \in \mathbb{R}^{n_x}$. In particular, letting $\mathbf{d} = -[\mu \boldsymbol{\nabla} f(\mathbf{x}^\star) + \boldsymbol{\lambda}^\mathsf{T} \boldsymbol{\nabla} \mathbf{h}(\mathbf{x}^\star)]$, it follows that $-\|\mu \boldsymbol{\nabla} f(\mathbf{x}^\star) + \boldsymbol{\lambda}^\mathsf{T} \boldsymbol{\nabla} \mathbf{h}(\mathbf{x}^\star)\|^2 \geq 0$, and thus,

$$\mu \boldsymbol{\nabla} f(\mathbf{x}^\star) + \boldsymbol{\lambda}^\mathsf{T} \boldsymbol{\nabla} \mathbf{h}(\mathbf{x}^\star) = \mathbf{0} \qquad \text{with } (\mu, \boldsymbol{\lambda}) \neq (0, \mathbf{0}). \tag{1.13}$$

Finally, note that $\mu > 0$, for otherwise (1.13) would contradict the assumption of linear independence of $\boldsymbol{\nabla} h_i(\mathbf{x}^\star)$, $i = 1, \ldots, n_h$. The result follows by letting $\boldsymbol{\lambda}^\star := \frac{1}{\mu}\boldsymbol{\lambda}$, and noting that the linear independence assumption implies the uniqueness of these Lagrangian multipliers. □

Remark 1.51 (Obtaining Candidate Solution Points). The first-order necessary conditions

$$\boldsymbol{\nabla} f(\mathbf{x}^\star) + \boldsymbol{\nabla} \mathbf{h}(\mathbf{x}^\star)^\mathsf{T} \boldsymbol{\lambda}^\star = \mathbf{0},$$

together with the constraints

$$\mathbf{h}(\mathbf{x}^\star) = \mathbf{0},$$

give a total of $n_x + n_h$ (typically nonlinear) equations in the variables $(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)$. Hence, these conditions are complete in the sense that they determine, at least locally, a unique solution. However, as in the unconstrained case, a solution to the first-order necessary conditions need not be a (local) minimum of the original problem; it could very well correspond to a (local) maximum point, or some kind of saddle point. These consideration are illustrated in Example 1.54 below.

Remark 1.52 (Regularity-Type Assumption). It is important to note that for a local minimum to satisfy the foregoing first-order conditions and, in particular, for a unique Lagrange multiplier vector to exist, it is necessary that the equality constraint satisfy a regularity condition. In other word, the first-order conditions may not hold at a local minimum point that is non-regular. An illustration of these considerations is provided in Example 1.55.

There exists a number of similarities with the constraint qualification needed for a local minimizer of an inequality constrained NLP problem to be KKT point; in particular, the condition that the minimum point be a regular point for the constraints corresponds to LICQ (see Remark 1.42).

Remark 1.53 (Lagrangian). It is convenient to introduce the *Lagrangian* $\mathcal{L} : \mathbb{R}^{n_x} \times \mathbb{R}^{n_h} \to \mathbb{R}$ associated with the constrained problem, by adjoining the cost and constraint functions as:

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) \;:=\; f(\mathbf{x}) + \boldsymbol{\lambda}^\mathsf{T} \mathbf{h}(\mathbf{x}).$$

Thus, if $\mathbf{x}^\star$ is a local minimum which is regular, the first-order necessary conditions are written as

$$\boldsymbol{\nabla}_\mathbf{x} \mathcal{L}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star) = \mathbf{0} \tag{1.14}$$

$$\boldsymbol{\nabla}_{\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star) = \mathbf{0}, \tag{1.15}$$

the latter equations being simply a restatement of the constraints. Note that the solution of the original problem typically corresponds to a saddle point of the Lagrangian function.

Example 1.54 (Regular Case). Consider the problem

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^2} \; & f(\mathbf{x}) := x_1 + x_2 \\ \text{s.t.} \; & h(\mathbf{x}) := x_1^2 + x_2^2 - 2 = 0. \end{aligned} \tag{1.16}$$

Observe first that every feasible point is a regular point for the equality constraint (the point (0,0) being infeasible). Therefore, every local minimum is a stationary point of the Lagrangian function by Theorem 1.50.

The gradient vectors $\nabla f(\mathbf{x})$ and $\nabla h(\mathbf{x})$ are given by

$$\nabla f(\mathbf{x}) = \begin{pmatrix} 1 & 1 \end{pmatrix}^\mathsf{T} \quad \text{and} \quad \nabla h(\mathbf{x}) = \begin{pmatrix} 2x_1 & 2x_2 \end{pmatrix}^\mathsf{T},$$

so that the first-order necessary conditions read

$$\begin{aligned} 2\lambda x_1 &= -1 \\ 2\lambda x_2 &= -1 \\ x_1^2 + x_2^2 &= 2. \end{aligned}$$

These three equations can be solved for the three unknowns x_1, x_2 and λ. Two candidate local minimum points are obtained: (i) ${x_1}^\star = {x_2}^\star = -1$, $\lambda^\star = \frac{1}{2}$, and (ii) ${x_1}^\star = {x_2}^\star = 1$, $\lambda^\star = -\frac{1}{2}$. These results are illustrated on Fig. 1.12.. It can be seen that only the former actually corresponds to a local minimum point, while the latter gives a local maximum point.

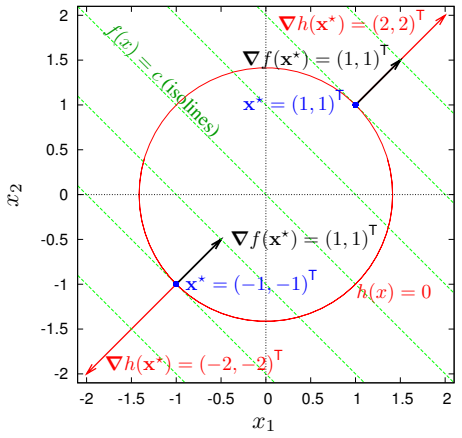


Figure 1.12. Solution of Example 1.54.

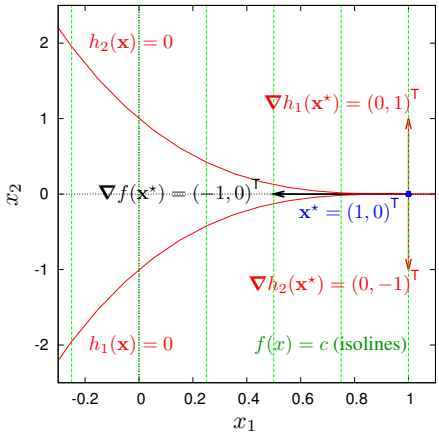


Figure 1.13. Solution of Example 1.55.

Example 1.55 (Non-Regular Case). Consider the problem

$$\begin{array}{ll} \min\limits_{\mathbf{x}\in\mathbb{R}^2} & f(\mathbf{x}) := -x_1 \\ \text{s.t.} & h_1(\mathbf{x}) := (1-x_1)^3 + x_2 = 0 \\ & h_2(\mathbf{x}) := (1-x_1)^3 - x_2 = 0. \end{array} \tag{1.17}$$

As shown by Fig. 1.13., this problem has only one feasible point, namely, $\mathbf{x}^\star = (1,0)^\mathsf{T}$; that is, $\mathbf{x}^\star$ is also the unique global minimum of (1.17). However, at this point, we have

$$\nabla f(\mathbf{x}^\star) = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad \nabla h_1(\mathbf{x}^\star) = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \text{and} \quad \nabla h_2(\mathbf{x}^\star) = \begin{pmatrix} 0 \\ -1 \end{pmatrix},$$

hence the first-order conditions

$$\lambda_1 \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \lambda_2 \begin{pmatrix} 0 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

cannot be satisfied. This illustrates the fact that a minimum point may not be a stationary point for the Lagrangian if that point is non-regular.

The following theorem provides second-order necessary conditions for a point to be a local minimum of a NLP problem with equality constraints.

Theorem 1.56 (Second-Order Necessary Conditions). *Let $f : \mathbb{R}^{n_x} \to \mathbb{R}$ and $h_i : \mathbb{R}^{n_x} \to \mathbb{R}$, $i = 1, \ldots, n_h$, be twice continuously differentiable functions on $\mathbb{R}^{n_x}$. Consider the problem to minimize $f(\mathbf{x})$ subject to the constraints $\mathbf{h}(\mathbf{x}) = \mathbf{0}$. If $\mathbf{x}^\star$ is a local minimum and is a regular point of the constraints, then there exists a unique vector $\boldsymbol{\lambda}^\star \in \mathbb{R}^{n_h}$ such that*

$$\nabla f(\mathbf{x}^\star) + \nabla \mathbf{h}(\mathbf{x}^\star)^\mathsf{T} \boldsymbol{\lambda}^\star = \mathbf{0},$$

and

$$\mathbf{d}^\mathsf{T} \left(\nabla^2 f(\mathbf{x}^\star) + \nabla^2 \mathbf{h}(\mathbf{x}^\star)^\mathsf{T} \boldsymbol{\lambda}^\star \right) \mathbf{d} \geq 0 \qquad \forall \mathbf{d} \text{ such that } \nabla \mathbf{h}(\mathbf{x}^\star)^\mathsf{T} \mathbf{d} = 0.$$

Proof. Note first that $\nabla f(\mathbf{x}^\star) + \nabla \mathbf{h}(\mathbf{x}^\star)^\mathsf{T} \boldsymbol{\lambda}^\star = \mathbf{0}$ directly follows from Theorem 1.50.

Let $\mathbf{d}$ be an arbitrary direction in $\mathscr{T}(\mathbf{x}^\star)$; that is, $\nabla \mathbf{h}(\mathbf{x}^\star)^\mathsf{T} \mathbf{d} = 0$ since $\mathbf{x}^\star$ is a regular point (see Lemma 1.48). Consider any twice-differentiable curve $\boldsymbol{\xi} : \mathcal{I} = [-a, a] \to S$, $a > 0$, passing through $\mathbf{x}^\star$ with $\boldsymbol{\xi}(0) = \mathbf{x}^\star$ and $\dot{\boldsymbol{\xi}}(0) = \mathbf{d}$. Let φ be the function defined as $\varphi(t) := f(\boldsymbol{\xi}(t))$, $\forall t \in \mathcal{I}$. Since $\mathbf{x}^\star$ is a local minimum of f on $S := \{\mathbf{x} \in \mathbb{R}^{n_x} : \mathbf{h}(\mathbf{x}) = \mathbf{0}\}$, $t^\star = 0$ is an unconstrained (local) minimum point for φ. By Theorem 1.25, it follows that

$$0 \;\leq\; \nabla^2 \varphi(0) \;=\; \dot{\boldsymbol{\xi}}(0)^\mathsf{T} \nabla^2 f(\mathbf{x}^\star) \dot{\boldsymbol{\xi}}(0) + \nabla f(\mathbf{x}^\star)^\mathsf{T} \ddot{\boldsymbol{\xi}}(0).$$

Furthermore, differentiating the relation $\mathbf{h}(\boldsymbol{\xi}(\mathbf{t}))^\mathsf{T} \boldsymbol{\lambda} = 0$ twice, we obtain

$$\dot{\boldsymbol{\xi}}(0)^\mathsf{T} \left(\nabla^2 \mathbf{h}(\mathbf{x}^\star)^\mathsf{T} \boldsymbol{\lambda} \right) \dot{\boldsymbol{\xi}}(0) + \left(\nabla \mathbf{h}(\mathbf{x}^\star)^\mathsf{T} \boldsymbol{\lambda} \right)^\mathsf{T} \ddot{\boldsymbol{\xi}}(0) \;=\; 0.$$

Adding the last two equations yields

$$\mathbf{d}^\mathsf{T} \left(\nabla^2 f(\mathbf{x}^\star) + \nabla^2 \mathbf{h}(\mathbf{x}^\star)^\mathsf{T} \boldsymbol{\lambda}^\star \right) \mathbf{d} \geq 0,$$

and this condition must hold for every $\mathbf{d}$ such that $\boldsymbol{\nabla}\mathbf{h}(\mathbf{x}^\star)^\mathsf{T}\mathbf{d} = 0$. □

Remark 1.57 (Eigenvalues in Tangent Space). In the foregoing theorem, it is shown that the matrix $\boldsymbol{\nabla}^2_{\mathbf{xx}}\mathcal{L}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)$ restricted to the subspace $\mathscr{T}(\mathbf{x}^\star)$ plays a key role. Geometrically, the restriction of $\boldsymbol{\nabla}^2_{\mathbf{xx}}\mathcal{L}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)$ to $\mathscr{T}(\mathbf{x}^\star)$ corresponds to the projection $\mathcal{P}_{\mathscr{T}(\mathbf{x}^\star)}[\boldsymbol{\nabla}^2_{\mathbf{xx}}\mathcal{L}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)]$.
A vector $\mathbf{y} \in \mathscr{T}(\mathbf{x}^\star)$ is said to be an *eigenvector* of $\mathcal{P}_{\mathscr{T}(\mathbf{x}^\star)}[\boldsymbol{\nabla}^2_{\mathbf{xx}}\mathcal{L}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)]$ if there is a real number μ such that

$$\mathcal{P}_{\mathscr{T}(\mathbf{x}^\star)}[\boldsymbol{\nabla}^2_{\mathbf{xx}}\mathcal{L}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)]\mathbf{y} \;=\; \mu\mathbf{y};$$

the corresponding μ is said to be an *eigenvalue* of $\mathcal{P}_{\mathscr{T}(\mathbf{x}^\star)}[\boldsymbol{\nabla}^2_{\mathbf{xx}}\mathcal{L}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)]$. (These definitions coincide with the usual definitions of eigenvector and eigenvalue for real matrices.)
Now, to obtain a matrix representation for $\mathcal{P}_{\mathscr{T}(\mathbf{x}^\star)}[\boldsymbol{\nabla}^2_{\mathbf{xx}}\mathcal{L}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)]$, it is necessary to introduce a basis of the subspace $\mathscr{T}(\mathbf{x}^\star)$, say $\mathsf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_{n_x - n_h})$. Then, the eigenvalues of $\mathcal{P}_{\mathscr{T}(\mathbf{x}^\star)}[\boldsymbol{\nabla}^2_{\mathbf{xx}}\mathcal{L}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)]$ are the same as those of the $(n_x - n_h) \times (n_x - n_h)$ matrix $\mathsf{E}^\mathsf{T}\boldsymbol{\nabla}^2_{\mathbf{xx}}\mathcal{L}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)\mathsf{E}$; in particular, they are independent of the particular choice of basis E.

Example 1.58 (Regular Case Continued). Consider the problem (1.16) addressed earlier in Example 1.54. Two candidate local minimum points, (i) ${x_1}^\star = {x_2}^\star = -1$, $\lambda^\star = \frac{1}{2}$, and (ii) ${x_1}^\star = {x_2}^\star = 1$, $\lambda^\star = -\frac{1}{2}$, were obtained on application of the first-order necessary conditions. The Hessian matrix of the Lagrangian function is given by

$$\boldsymbol{\nabla}^2_{\mathbf{xx}}\mathcal{L}(\mathbf{x}, \lambda) \;=\; \boldsymbol{\nabla}^2 f(\mathbf{x}) + \lambda\boldsymbol{\nabla}^2 h(\mathbf{x}) \;=\; \lambda\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix},$$

and a basis of the tangent subspace at a point $\mathbf{x} \in \mathscr{T}(\mathbf{x})$, $\mathbf{x} \neq (0, 0)$, is

$$\mathsf{E}(\mathbf{x}) \;:=\; \begin{pmatrix} -x_2 \\ x_1 \end{pmatrix}.$$

Therefore,

$$\mathsf{E}^\mathsf{T}\boldsymbol{\nabla}^2_{\mathbf{xx}}\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})\mathsf{E} \;=\; 2\lambda(x_1^2 + x_2^2).$$

In particular, for the candidate solution point (i), we have

$$\mathsf{E}^\mathsf{T}\boldsymbol{\nabla}^2_{\mathbf{xx}}\mathcal{L}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)\mathsf{E} \;=\; 2 \;>\; 0,$$

hence satisfying the second-order necessary conditions (in fact, this point also satisfies the second-order sufficient conditions of optimality discussed hereafter). On the other hand, for the candidate solution point (ii), we get

$$\mathsf{E}^\mathsf{T}\boldsymbol{\nabla}^2_{\mathbf{xx}}\mathcal{L}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)\mathsf{E} \;=\; -2 \;<\; 0$$

which does not satisfy the second-order requirement, so this point cannot be a local minimum.

The conditions given in Theorems 1.50 and 1.56 are necessary conditions that must hold at each local minimum point. Yet, a point satisfying these conditions may not be a local minimum. The following theorem provides sufficient conditions for a stationary point of the Lagrangian function to be a (local) minimum, provided that the Hessian matrix

of the Lagrangian function is locally convex along directions in the tangent space of the constraints.

Theorem 1.59 (Second-Order Sufficient Conditions). *Let $f : \mathbb{R}^{n_x} \to \mathbb{R}$ and $h_i : \mathbb{R}^{n_x} \to \mathbb{R}$, $i = 1, \ldots, n_h$, be twice continuously differentiable functions on $\mathbb{R}^{n_x}$. Consider the problem to minimize $f(\mathbf{x})$ subject to the constraints $\mathbf{h}(\mathbf{x}) = \mathbf{0}$. If $\mathbf{x}^\star$ and $\boldsymbol{\lambda}^\star$ satisfy*

$$\begin{aligned} \boldsymbol{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star) &= 0 \\ \boldsymbol{\nabla}_{\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star) &= 0, \end{aligned}$$

and

$$\mathbf{y}^\mathsf{T} \boldsymbol{\nabla}^2_{\mathbf{xx}} \mathcal{L}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star) \mathbf{y} > 0 \quad \forall \mathbf{y} \neq \mathbf{0} \text{ such that } \boldsymbol{\nabla} \mathbf{h}(\mathbf{x}^\star)^\mathsf{T} \mathbf{y} = 0,$$

where $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^\mathsf{T} \mathbf{h}(\mathbf{x})$, then $\mathbf{x}^\star$ is a strict local minimum.

Proof. Consider the augmented Lagrangian function

$$\bar{\mathcal{L}}(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^\mathsf{T} \mathbf{h}(\mathbf{x}) + \frac{c}{2} \|\mathbf{h}(\mathbf{x})\|^2,$$

where c is a scalar. We have

$$\begin{aligned} \boldsymbol{\nabla}_{\mathbf{x}} \bar{\mathcal{L}}(\mathbf{x}, \boldsymbol{\lambda}) &= \boldsymbol{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \bar{\boldsymbol{\lambda}}) \\ \boldsymbol{\nabla}^2_{\mathbf{xx}} \bar{\mathcal{L}}(\mathbf{x}, \boldsymbol{\lambda}) &= \boldsymbol{\nabla}^2_{\mathbf{xx}} \mathcal{L}(\mathbf{x}, \bar{\boldsymbol{\lambda}}) + c \boldsymbol{\nabla} \mathbf{h}(\mathbf{x})^\mathsf{T} \boldsymbol{\nabla} \mathbf{h}(\mathbf{x}), \end{aligned}$$

where $\bar{\boldsymbol{\lambda}} = \boldsymbol{\lambda} + c\mathbf{h}(\mathbf{x})$. Since $(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)$ satisfy the sufficient conditions and by Lemma 1.A.79, we obtain

$$\boldsymbol{\nabla}_{\mathbf{x}} \bar{\mathcal{L}}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star) = 0 \quad \text{and} \quad \boldsymbol{\nabla}^2_{\mathbf{xx}} \bar{\mathcal{L}}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star) \succ 0,$$

for sufficiently large c. $\bar{\mathcal{L}}$ being definite positive at $(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)$,

$$\exists \varrho > 0, \delta > 0 \text{ such that } \bar{\mathcal{L}}(\mathbf{x}, \boldsymbol{\lambda}^\star) \geq \bar{\mathcal{L}}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star) + \frac{\varrho}{2} \|\mathbf{x} - \mathbf{x}^\star\|^2 \quad \text{for } \|\mathbf{x} - \mathbf{x}^\star\| < \delta.$$

Finally, since $\bar{\mathcal{L}}(\mathbf{x}, \boldsymbol{\lambda}^\star) = f(\mathbf{x})$ when $\mathbf{h}(\mathbf{x}) = \mathbf{0}$, we get

$$f(\mathbf{x}) \geq f(\mathbf{x}^\star) + \frac{\varrho}{2} \|\mathbf{x} - \mathbf{x}^\star\|^2 \quad \text{if } \mathbf{h}(\mathbf{x}) = \mathbf{0}, \|\mathbf{x} - \mathbf{x}^\star\| < \delta,$$

i.e., $\mathbf{x}^\star$ is a strict local minimum. □

Example 1.60. Consider the problem

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^3} \; & f(\mathbf{x}) := -x_1 x_2 - x_1 x_3 - x_2 x_3 \\ \text{s.t.} \; & h(\mathbf{x}) := x_1 + x_2 + x_3 - 3 = 0. \end{aligned} \tag{1.18}$$

The first-order conditions for this problem are

$$\begin{aligned} -(x_2 + x_3) + \lambda &= 0 \\ -(x_1 + x_3) + \lambda &= 0 \\ -(x_1 + x_2) + \lambda &= 0, \end{aligned}$$

together with the equality constraint. It is easily checked that the point $x_1^\star = x_2^\star = x_3^\star = 1$, $\lambda^\star = 2$ satisfies these conditions. Moreover,

$$\boldsymbol{\nabla}^2_{\mathbf{xx}}\mathcal{L}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star) = \boldsymbol{\nabla}^2 f(\mathbf{x}^\star) = \begin{pmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \end{pmatrix},$$

and a basis of the tangent space to the constraint $h(\mathbf{x}) = 0$ at $\mathbf{x}^\star$ is

$$\mathsf{E} := \begin{pmatrix} 0 & 2 \\ 1 & -1 \\ -1 & -1 \end{pmatrix}.$$

We thus obtain

$$\mathsf{E}^\mathsf{T}\boldsymbol{\nabla}^2_{\mathbf{xx}}\mathcal{L}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)\mathsf{E} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix},$$

which is clearly a definite positive matrix. Hence, $\mathbf{x}^\star$ is a strict local minimum of (1.18). (Interestingly enough, the Hessian matrix of the objective function itself is indefinite at $\mathbf{x}^\star$ in this case.)

We close this section by providing insight into the Lagrange multipliers.

Remark 1.61 (Interpretation of the Lagrange Multipliers). The concept of Lagrange multipliers allows to adjoin the constraints to the objective function. That is, one can view constrained optimization as a search for a vector $\mathbf{x}^\star$ at which the gradient of the objective function is a linear combination of the gradients of constraints.

Another insightful interpretation of the Lagrange multipliers is as follows. Consider the set of perturbed problems $v^\star(y) := \min\{f(\mathbf{x}) : h(\mathbf{x}) = y\}$. Suppose that there is a unique regular solution point for each y, and let $\{\boldsymbol{\xi}^\star(y)\} := \arg\min\{f(\mathbf{x}) : h(\mathbf{x}) = y\}$ denote the evolution of the optimal solution point as a function of the perturbation parameter y. Clearly,

$$v(0) = f(\mathbf{x}^\star) \quad \text{and} \quad \boldsymbol{\xi}(0) = \mathbf{x}^\star.$$

Moreover, since $h(\boldsymbol{\xi}(y)) = y$ for each y, we have

$$\nabla_y h(\boldsymbol{\xi}(y)) \;=\; 1 \;=\; \boldsymbol{\nabla}_{\mathbf{x}} h(\boldsymbol{\xi}(y))^\mathsf{T} \boldsymbol{\nabla}_y \boldsymbol{\xi}(y).$$

Denoting by $\lambda^\star$ the Lagrange multiplier associated to the constraint $h(\mathbf{x}) = 0$ in the original problem, we have

$$\nabla_y v(0) \;=\; \boldsymbol{\nabla}_{\mathbf{x}} f(\mathbf{x}^\star)^\mathsf{T} \boldsymbol{\nabla}_y \boldsymbol{\xi}(0) \;=\; -\lambda^\star \boldsymbol{\nabla}_{\mathbf{x}} h(\mathbf{x}^\star)^\mathsf{T} \boldsymbol{\nabla}_y \boldsymbol{\xi}(0) \;=\; -\lambda^\star.$$

Therefore, the Lagrange multipliers $\lambda^\star$ can be interpreted as the sensitivity of the objective function f with respect to the constraint h. Said differently, $\lambda^\star$ indicates how much the optimal cost would change, if the constraints were perturbed.

This interpretation extends straightforwardly to NLP problems having inequality constraints. The Lagrange multipliers of an active constraints $g(\mathbf{x}) \leq 0$, say $\nu^\star > 0$, can be interpreted as the sensitivity of $f(\mathbf{x}^\star)$ with respect to a change in that constraints, as $g(\mathbf{x}) \leq y$; in this case, the positivity of the Lagrange multipliers follows from the fact that by increasing y, the feasible region is relaxed, hence the optimal cost cannot increase. Regarding inactive constraints, the sensitivity interpretation also explains why the Lagrange multipliers are zero, as any infinitesimal change in the value of these constraints leaves the optimal cost value unchanged.

1.7 GENERAL NLP PROBLEMS

In this section, we shall consider general, nonlinear programming problems with both equality and inequality constraints,

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & g_i(\mathbf{x}) \;\leq\; 0, \quad i = 1, \ldots, n_g \\ & h_i(\mathbf{x}) \;=\; 0, \quad i = 1, \ldots, n_h. \end{array}$$

Before stating necessary and sufficient conditions for such problems, we shall start by revisiting the KKT conditions for inequality constrained problems, based on the method of Lagrange multipliers described in §1.6.

1.7.1 KKT Conditions for Inequality Constrained NLP Problems Revisited

Consider the problem to minimize a function $f(\mathbf{x})$ for $\mathbf{x} \in S := \{\mathbf{x} \in \mathbb{R}^{n_x} : g_i(\mathbf{x}) \leq 0, i = 1, \ldots, n_g\}$, and suppose that $\mathbf{x}^\star$ is a local minimum point. Clearly, $\mathbf{x}^\star$ is also a local minimum of the inequality constrained problem where the inactive constraints $g_i(\mathbf{x}) \leq 0$, $i \notin \mathcal{A}(\mathbf{x}^\star)$, have been discarded. Thus, in effect, *inactive constraints at* $\mathbf{x}^\star$ *don't matter*; they can be ignored in the statement of optimality conditions.

On the other hand, *active inequality constraints can be treated to a large extend as equality constraints* at a local minimum point. In particular, it should be clear that $\mathbf{x}^\star$ is also a local minimum to the equality constrained problem

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & g_i(\mathbf{x}) \;=\; 0, \quad i \in \mathcal{A}(\mathbf{x}^\star). \end{array}$$

That is, it follows from Theorem 1.50 that, if $\mathbf{x}^\star$ is a regular point, there exists a unique Lagrange multiplier vector $\boldsymbol{\nu}^\star \in \mathbb{R}^{n_g}$ such that

$$\boldsymbol{\nabla} f(\mathbf{x}^\star) + \sum_{i \in \mathcal{A}(\mathbf{x}^\star)} \nu_i^\star \boldsymbol{\nabla} g_i(\mathbf{x}^\star) \;=\; \mathbf{0}.$$

Assigning zero Lagrange multipliers to the inactive constraints, we obtain

$$\boldsymbol{\nabla} f(\mathbf{x}^\star) + \boldsymbol{\nabla} \mathbf{g}(\mathbf{x}^\star)^{\mathsf{T}} \boldsymbol{\nu}^\star \;= \mathbf{0} \tag{1.19}$$

$$\nu_i^\star \;= 0, \quad \forall i \notin \mathcal{A}(\mathbf{x}^\star). \tag{1.20}$$

This latter condition can be rewritten by means of the following inequalities

$$\nu_i^\star g_i(\mathbf{x}^\star) = 0 \qquad \forall i = 1, \ldots, n_g.$$

The argument showing that $\boldsymbol{\nu} \geq \mathbf{0}$ is a little more elaborate. By contradiction, assume that $\nu_\ell < 0$ for some $\ell \in \mathcal{A}(\mathbf{x}^\star)$. Let $\mathsf{A} \in \mathbb{R}^{(n_g+1) \times n_x}$ be the matrix whose rows are $\boldsymbol{\nabla} f(\mathbf{x}^\star)$ and $\boldsymbol{\nabla} g_i(\mathbf{x}^\star)$, $i = 1, \ldots, n_g$. Since $\mathbf{x}^\star$ is a regular point, the Lagrange multiplier vector $\boldsymbol{\nu}^\star$ is unique. Therefore, the condition

$$\mathsf{A}^{\mathsf{T}} \mathbf{y} \;=\; \mathbf{0},$$

can only be satisfied by $\mathbf{y}^\star := \eta(\; 1 \quad \boldsymbol{\nu}^\star \;)^{\mathsf{T}}$ with $\eta \in \mathbb{R}$. Because $\nu_\ell < 0$, we know by Gordan's Theorem 1.A.78 that there exists a direction $\bar{\mathbf{d}} \in \mathbb{R}^{n_x}$ such that $\mathsf{A}\bar{\mathbf{d}} < \mathbf{0}$. In other words,

$$\bar{\mathbf{d}} \in \mathcal{F}_0(\mathbf{x}^\star) \cap \mathcal{D}_0(\mathbf{x}^\star) \neq \emptyset,$$

which contradicts the hypothesis that $\mathbf{x}^\star$ is a local minimizer of f on S (see Remark 1.34).

Overall, these results thus constitute the KKT optimality conditions as stated in Theorem 1.39. But although the foregoing development is straightforward, it is somewhat limited by the regularity-type assumption at the optimal solution. Obtaining more general constraint qualifications (see Remark 1.42) requires that the KKT conditions be derived using an alternative approach, e.g., the approach described earlier in §1.5. Still, the conversion to equality constrained problem proves useful in many situations, e.g., for deriving second-order sufficiency conditions for inequality constrained NLP problems.

1.7.2 Optimality Conditions for General NLP Problems

We are now ready to generalize the necessary and sufficient conditions given in Theorems 1.39, 1.50, 1.56 and 1.59 to general NLP problems.

Theorem 1.62 (First- and Second-Order Necessary Conditions). *Let $f : \mathbb{R}^{n_x} \to \mathbb{R}$, $g_i : \mathbb{R}^{n_x} \to \mathbb{R}$, $i = 1, \ldots, n_g$, and $h_i : \mathbb{R}^{n_x} \to \mathbb{R}$, $i = 1, \ldots, n_h$, be twice continuously differentiable functions on $\mathbb{R}^{n_x}$. Consider the problem P to minimize $f(\mathbf{x})$ subject to the constraints $\mathbf{g}(\mathbf{x}) = \mathbf{0}$ and $\mathbf{h}(\mathbf{x}) = \mathbf{0}$. If $\mathbf{x}^\star$ is a local minimum of P and is a regular point of the constraints, then there exist unique vectors $\boldsymbol{\nu}^\star \in \mathbb{R}^{n_g}$ and $\boldsymbol{\lambda}^\star \in \mathbb{R}^{n_h}$ such that*

$$\nabla f(\mathbf{x}^\star) + \nabla \mathbf{g}(\mathbf{x}^\star)^\mathsf{T} \boldsymbol{\nu}^\star + \nabla \mathbf{h}(\mathbf{x}^\star)^\mathsf{T} \boldsymbol{\lambda}^\star = \mathbf{0} \tag{1.21}$$

$$\boldsymbol{\nu}^\star \geq \mathbf{0} \tag{1.22}$$

$$\mathbf{g}(\mathbf{x}^\star) \leq \mathbf{0} \tag{1.23}$$

$$\mathbf{h}(\mathbf{x}^\star) = \mathbf{0} \tag{1.24}$$

$$\boldsymbol{\nu}^{\star\mathsf{T}} \mathbf{g}(\mathbf{x}^\star) = 0, \tag{1.25}$$

and

$$\mathbf{y}^\mathsf{T} \left(\nabla^2 f(\mathbf{x}^\star) + \nabla^2 \mathbf{g}(\mathbf{x}^\star)^\mathsf{T} \boldsymbol{\nu}^\star + \nabla^2 \mathbf{h}(\mathbf{x}^\star)^\mathsf{T} \boldsymbol{\lambda}^\star \right) \mathbf{y} \geq 0,$$

for all $\mathbf{y}$ such that $\nabla g_i(\mathbf{x}^\star)^\mathsf{T} \mathbf{y} = 0$, $i \in \mathcal{A}(\mathbf{x}^\star)$, and $\nabla \mathbf{h}(\mathbf{x}^\star)^\mathsf{T} \mathbf{y} = \mathbf{0}$.

Theorem 1.63 (Second-Order Sufficient Conditions). *Let $f : \mathbb{R}^{n_x} \to \mathbb{R}$, $g_i : \mathbb{R}^{n_x} \to \mathbb{R}$, $i = 1, \ldots, n_g$, and $h_i : \mathbb{R}^{n_x} \to \mathbb{R}$, $i = 1, \ldots, n_h$, be twice continuously differentiable functions on $\mathbb{R}^{n_x}$. Consider the problem P to minimize $f(\mathbf{x})$ subject to the constraints $\mathbf{g}(\mathbf{x}) = \mathbf{0}$ and $\mathbf{h}(\mathbf{x}) = \mathbf{0}$. If there exists $\mathbf{x}^\star$, $\boldsymbol{\nu}^\star$ and $\boldsymbol{\lambda}^\star$ satisfying the KKT conditions (1.21–1.25), and*

$$\mathbf{y}^\mathsf{T} \nabla^2_{\mathbf{xx}} \mathcal{L}(\mathbf{x}^\star, \boldsymbol{\nu}^\star, \boldsymbol{\lambda}^\star) \mathbf{y} > 0$$

for all $\mathbf{y} \neq \mathbf{0}$ such that

$$\begin{aligned} \nabla g_i(\mathbf{x}^\star)^\mathsf{T} \mathbf{y} &= 0 \quad && i \in \mathcal{A}(\mathbf{x}^\star) \text{ with } \nu_i^\star > 0 \\ \nabla g_i(\mathbf{x}^\star)^\mathsf{T} \mathbf{y} &\leq 0 && i \in \mathcal{A}(\mathbf{x}^\star) \text{ with } \nu_i^\star = 0 \\ \nabla \mathbf{h}(\mathbf{x}^\star)^\mathsf{T} \mathbf{y} &= 0, \end{aligned}$$

where $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\nu}^\mathsf{T} \mathbf{g}(\mathbf{x}) + \boldsymbol{\lambda}^\mathsf{T} \mathbf{h}(\mathbf{x})$, then $\mathbf{x}^\star$ is a strict local minimum of P.

Likewise, the KKT sufficient conditions given in Theorem 1.44 for convex, inequality constrained problems can be generalized to general convex problems as follows:

Theorem 1.64 (KKT sufficient Conditions). *Let* $f : \mathbb{R}^{n_x} \to \mathbb{R}$ *and* $g_i : \mathbb{R}^{n_x} \to \mathbb{R}$, $i = 1, \ldots, n_g$, *be convex and differentiable functions. Let also* $h_i : \mathbb{R}^{n_x} \to \mathbb{R}$, $i = 1, \ldots, n_h$, *be affine functions. Consider the problem to minimize* $f(\mathbf{x})$ *subject to* $\mathbf{x} \in S := \{\mathbf{x} \in \mathbb{R}^{n_x} : \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \mathbf{h}(\mathbf{x}) = \mathbf{0}\}$. *If* $(\mathbf{x}^\star, \boldsymbol{\nu}^\star, \boldsymbol{\lambda}^\star)$ *satisfies the KKT conditions (1.21–1.25), then* $\mathbf{x}^\star$ *is a global minimizer for* f *on* S.

1.8 NUMERICAL METHODS FOR NONLINEAR PROGRAMMING PROBLEMS

Nowadays, strong and efficient mathematical programming techniques are available for solving a great variety of nonlinear problems, which are based on solid theoretical results and extensive numerical studies. Approximated functions, derivatives and optimal solutions can be employed together with optimization algorithms to reduce the computational time.

The aim of this section is *not* to describe state-of-the-art algorithms in nonlinear programming, but to explain, in a simple way, a number of modern algorithms for solving nonlinear problems. These techniques are typically *iterative* in the sense that, given an initial point $\mathbf{x}^0$, a sequence of points, $\{\mathbf{x}^k\}$, is obtained by repeated application of an algorithmic rule. The objective is to make this sequence converge to a point $\bar{\mathbf{x}}$ in a pre-specified *solution set*. Typically, the solution set is specified in terms of optimality conditions, such as those presented in §1.4 through §1.7.

We start by recalling a number of concepts in §1.8.1. Then, we discuss the principles of Newton-like algorithms for nonlinear systems in §1.8.2, and use these concepts for the solution of unconstrained optimization problems in §1.8.3. Finally, algorithms for solving general, nonlinear problems are presented in §1.8.4, with emphasizes on sequential unconstrained minimization (SUM) and sequential quadratic programming (SQP) techniques.

1.8.1 Preliminaries

Two essential questions must be addressed concerning iterative algorithms. The first question, which is qualitative in nature, is whether a given algorithm in some sense yields, at least in the limit, a solution to the original problem; the second question, the more quantitative one, is concerned with how fast the algorithm converges to a solution. We elaborate on these concepts in this subsection.

The convergence of an algorithm is said to *asymptotic* when the solution is not achieved after a finite number of iterations; except for particular problems such as linear and quadratic programming, this is generally the case in nonlinear programming. That is, a very desirable property of an optimization algorithm is global convergence:

Definition 1.65 (Global Convergence, Local Convergence). *An algorithm is said to be* globally convergent *if, for any initial point* $\mathbf{x}^0$, *it generates a sequence of points that converges to a point* $\bar{\mathbf{x}}$ *in the solution set. It is said to be* locally convergent *if there exists a* $\varrho > 0$ *such that for any initial point* $\mathbf{x}^0$ *such that* $\|\bar{\mathbf{x}} - \mathbf{x}^0\| < \varrho$, *it generates a sequence of points converging to* $\bar{\mathbf{x}}$ *in the solution set.*

Most modern mathematical programming algorithms are globally convergent. Locally convergent algorithms are not useful in practice because the neighborhood of convergence is not known in advance and can be arbitrarily small.

Next, what distinguishes optimization algorithms with the global convergence property is the order of convergence:

Definition 1.66 (Order of Convergence). *The* order of convergence *of a sequence* $\{\mathbf{x}^k\} \to \bar{\mathbf{x}}$ *is the largest nonnegative integer* p *such that*

$$\lim_{k\to\infty} \frac{\|\mathbf{x}^{k+1} - \bar{\mathbf{x}}\|}{\|\mathbf{x}^k - \bar{\mathbf{x}}\|^p} \;=\; \beta \;<\; \infty,$$

When $p = 1$ *and the convergence ratio* $\beta < 1$*, the convergence is said to be* linear*; if* $\beta = 0$*, the convergence is said to be* superlinear. *When* $p = 2$*, the convergence is said to be* quadratic.

Since they involve the limit when $k \to \infty$, p and β are a measure of the *asymptotic* rate of convergence, i.e., as the iterates gets close to the solution; yet, a sequence with a good order of convergence may be very "slow" far from the solution. Clearly, the convergence is faster when p is larger and β is smaller. Near the solution, if the convergence rate is linear, then the error is multiplied by β at each iteration. The error reduction is squared for quadratic convergence, i.e., each iteration roughly doubles the number of significant digits. The methods that will be studied hereafter have convergence rates varying between linear and quadratic.

Example 1.67. Consider the problem to minimize $f(x) = x^2$, subject to $x \geq 1$.

Let the (point-to-point) algorithmic map $\mathcal{M}_1$ be defined defined as $\mathcal{M}_1(x) = \frac{1}{2}(x+1)$. It is easily verified that the sequence obtained by applying the map $\mathcal{M}_1$, with any starting point, converges to the optimal solution $x^\star = 1$, i.e., $\mathcal{M}_1$ is globally convergent. For examples, with $x^0 = 4$, the algorithm generates the sequence $\{4, 2.5, 1.75, 1.375, 1.1875, \ldots\}$. We have $(x^{k+1} - 1) = \frac{1}{2}(x^k - 1)$, so that the limit in Definition 1.66 is $\beta = \frac{1}{2}$ with $p = 1$; moreover, for $p > 1$, this limit is infinity. Consequently, $x^k \to 1$ linearly with convergence ratio $\frac{1}{2}$.

On the other hand, consider the (point-to-point) algorithmic map $\mathcal{M}_2$ be defined defined as $\mathcal{M}_2(x; k) = 1 + \frac{1}{2^{k+1}}(x-1)$. Again, the sequence obtained by applying $\mathcal{M}_2$ converges to $x^\star = 1$, from any starting point. However, we now have $\frac{|x^{k+1}-1|}{|x^k-1|} = \frac{1}{2^k}$, which approaches 0 as $k \to \infty$. Hence, $x^k \to 1$ superlinearly in this case. With $x^0 = 4$, the algorithm generates the sequence $\{4, 2.5, 1.375, 1.046875, \ldots\}$.

The algorithmic maps $\mathcal{M}_1$ and $\mathcal{M}_2$ are illustrated on the left and right plots in Fig 1.14., respectively.

It should also be noted that for most algorithms, the user must set initial values for certain parameters, such as the starting point and the initial step size, as well as parameters for terminating the algorithm. Optimization procedures are often quite sensitive to these parameters, and may produce different results, or even stop prematurely, depending on their values. Therefore, it is crucial for the user to understand the principles of the algorithms used, so that he or she can select adequate values for the parameters and diagnose the reasons of a premature termination (failure).

1.8.2 Newton-like Algorithms for nonlinear Systems

The fundamental approach to most iterative schemes was suggested over 300 years ago by Newton. In fact, Newton's method is the basis for nearly all the algorithms that are described herein.

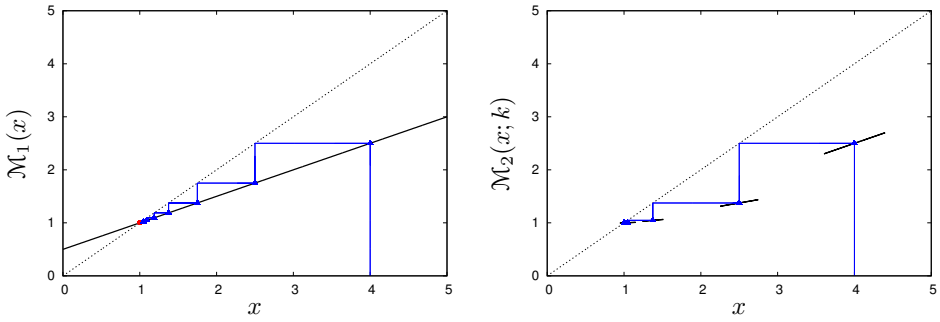


Figure 1.14. Illustration of algorithmic maps $\mathcal{M}_1$ and $\mathcal{M}_2$ in Example 1.67, with $x^0 = 4$.

Suppose one wants to find the value of the variable $\mathbf{x} \in \mathbb{R}^{n_x}$ such that

$$\boldsymbol{\phi}(\mathbf{x}) \;=\; \mathbf{0},$$

where $\boldsymbol{\phi} : \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$ is continuously differentiable. Let us assume that one such solution exists, and denote it by $\mathbf{x}^\star$. Let also $\mathbf{x}$ be a guess for the solution. The basic idea of Newton's method is to approximate the nonlinear function $\boldsymbol{\phi}$ by the first two terms in its Taylor series expansion about the current point $\mathbf{x}$. This yields a linear approximation for the vector function $\boldsymbol{\phi}$ at the new point $\bar{\mathbf{x}}$,

$$\boldsymbol{\phi}(\bar{\mathbf{x}}) \;=\; \boldsymbol{\phi}(\mathbf{x}) + \boldsymbol{\nabla}\boldsymbol{\phi}(\mathbf{x})\,[\bar{\mathbf{x}} - \mathbf{x}]\,. \tag{1.26}$$

Using this linear approximation, and provided that the Jacobian matrix $\boldsymbol{\nabla}\boldsymbol{\phi}(\mathbf{x})$ is non-singular, a new estimate for the solution $\mathbf{x}^\star$ can be computed by solving (1.26) such that $\boldsymbol{\phi}(\bar{\mathbf{x}}) = \mathbf{0}$, i.e.,

$$\bar{\mathbf{x}} \;=\; \mathbf{x} - \left[\boldsymbol{\nabla}\boldsymbol{\phi}(\mathbf{x})\right]^{-1} \boldsymbol{\phi}(\mathbf{x}).$$

Letting $\mathbf{d} := -\left[\boldsymbol{\nabla}\boldsymbol{\phi}(\mathbf{x})\right]^{-1} \boldsymbol{\phi}(\mathbf{x})$, we get the update $\bar{\mathbf{x}} = \mathbf{x} + \mathbf{d}$.

Of course, $\boldsymbol{\phi}$ being a nonlinear function, one cannot expect that $\boldsymbol{\phi}(\bar{\mathbf{x}}) = \mathbf{0}$, but there is much hope that $\bar{\mathbf{x}}$ be a better estimate for the root $\mathbf{x}^\star$ than the original guess $\mathbf{x}$. In other words, we might expect that

$$|\bar{\mathbf{x}} - \mathbf{x}^\star| \le |\mathbf{x} - \mathbf{x}^\star| \quad \text{and} \quad |\boldsymbol{\phi}(\bar{\mathbf{x}})| \le |\boldsymbol{\phi}(\mathbf{x})|.$$

If the new point is an improvement, then it makes sense to repeat the process, thereby defining a sequence of points $\mathbf{x}^0, \mathbf{x}^1, \ldots$. An algorithm implementing Newton's method is as follows:

Initialization Step

Let $\varepsilon > 0$ be a termination scalar, and choose an initial point $\mathbf{x}^0$. Let $k = 0$ and go to the main step.

Main Step

1. Solve the linear system $\boldsymbol{\nabla}\boldsymbol{\phi}(\mathbf{x}^k)\mathbf{d}^k = -\boldsymbol{\phi}(\mathbf{x}^k)$ for $\mathbf{d}^k$.
2. Compute the new estimate $\mathbf{x}^{k+1} = \mathbf{x}^k + \mathbf{d}^k$.
3. If $\|\boldsymbol{\phi}(\mathbf{x}^{k+1})\| < \varepsilon$, stop; otherwise, replace $k \leftarrow k + 1$, and go to step 1.

It can be shown that the rate of convergence for Newton's method is quadratic (see Definition 1.66). Loosely speaking, it implies that each successive estimate of the solution *doubles* the number significant digits, which is a very desirable property for an algorithm to possess.

Theorem 1.68. *Let $\boldsymbol{\phi} : \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$ be continuously differentiable, and consider Newton's algorithm defined by the map $\mathcal{M}(\mathbf{x}) := \mathbf{x} - \boldsymbol{\nabla}\boldsymbol{\phi}(\mathbf{x})^{-1}\boldsymbol{\phi}(\mathbf{x})$. Let $\mathbf{x}^\star$ be such that $\boldsymbol{\phi}(\mathbf{x}^\star) = \mathbf{0}$, and suppose that $\boldsymbol{\nabla}\boldsymbol{\phi}(\mathbf{x}^\star)$ is nonsingular. Let the starting point $\mathbf{x}^0$ be sufficiently close to $\mathbf{x}^\star$, so that there exist $c_1, c_2 > 0$ with $c_1 c_2 \|\mathbf{x}^0 - \mathbf{x}^\star\| < 1$, and*

$$\begin{aligned} \|\boldsymbol{\nabla}\boldsymbol{\phi}(\mathbf{x})^{-1}\| &\leq c_1 \\ \|\boldsymbol{\phi}(\mathbf{x}^\star) - \boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\nabla}\boldsymbol{\phi}(\mathbf{x})\,[\mathbf{x}^\star - \mathbf{x}]\,\| &\leq c_2 \|\mathbf{x}^\star - \mathbf{x}\|^2, \end{aligned} \tag{1.27}$$

for each $\mathbf{x}$ satisfying $\|\mathbf{x}^\star - \mathbf{x}\| \leq \|\mathbf{x}^\star - \mathbf{x}^0\|$. Then, Newton's algorithm converges with a quadratic rate of convergence.

Proof. See [6, Theorem 8.6.5] for a proof. □

But can anything go wrong with Newton's method?

Lack of Global Convergence First and foremost, if the initial guess is not sufficiently close to the solution, i.e., within the *region of convergence*, Newton's method may diverge. Said differently, Newton's method as presented above does *not* have the global convergence property (see Definition 1.65 and Example 1.69 hereafter). This is because $\mathbf{d}^k := \boldsymbol{\nabla}\boldsymbol{\phi}(\mathbf{x}^k)^{-1}\boldsymbol{\phi}(\mathbf{x}^k)$ may not be a valid descent direction far from the solution, and even if $\boldsymbol{\nabla}\boldsymbol{\phi}(\mathbf{x}^k)\mathbf{d}^k < 0$, a unit step size might not give a descent in $\boldsymbol{\phi}$. Globalization strategies, which aim at correcting this latter deficiency, will be presented in §1.8.3.1 in the context of unconstrained optimization.

Singular Jacobian Matrix A second difficulty occurs when the Jacobian matrix $\boldsymbol{\nabla}\boldsymbol{\phi}(\mathbf{x}^k)$ becomes singular during the iteration process, since the correction defined by (1.8.2) is not defined in this case. Note that the assumption (1.27) in Theorem 1.68 guarantees that $\boldsymbol{\nabla}\boldsymbol{\phi}(\mathbf{x}^k)$ cannot be singular. But when the Jacobian matrix is singular at the solution point $\mathbf{x}^\star$, then Newton's method looses its quadratic convergence property.

Computational Efficiency Finally, at each iteration, Newton's method requires (i) that the Jacobian matrix $\boldsymbol{\nabla}\boldsymbol{\phi}(\mathbf{x}^k)$ be computed, which may be difficult and/or costly especially when the expression of $\boldsymbol{\phi}(\mathbf{x})$ is complicated, and (ii) that a linear system be solved. The analytic Jacobian can be replaced by a finite-difference approximation, yet this is costly as n_x additional evaluations of $\boldsymbol{\phi}$ are required at each iterations. With the objective of reducing the computational effort, *quasi-Newton* methods generate an approximation of the Jacobian matrix, based on the information gathered from the iteration progress. To avoid solving a linear system for the search direction, variants of quasi-Newton methods also exist that generate an approximation of the inverse of the Jacobian matrix. Such methods will be described in §1.8.3.2 in the context of unconstrained optimization.

Example 1.69. Consider the problem to find a solution to the nonlinear equation

$$f(x) \;=\; \arctan(x) \;=\; 0.$$

The Newton iteration sequence obtained by starting from $x^0 = 1$ is as follows:

k	x^k	$\lvert f(x^k)\rvert$
0	1	0.785398
1	-0.570796	0.518669
2	0.116860	0.116332
3	-1.061022×10^{-3}	1.061022×10^{-3}
4	7.963096×10^{-10}	7.963096×10^{-10}

Notice the very fast convergence to the solution $x^\star = 0$, as could be expected from Theorem 1.68. The first three iterations are represented in Fig. 1.15., on the left plot.

However, convergence is not guaranteed for any initial guess. More precisely, it can be shown that Newton's method actually diverges when the initial guess is chosen such that $|x^0| > \alpha$, with $\alpha \approx 1.3917452002707$ being a solution of $\arctan(z) = \frac{2z}{1+z^2}$; further, the method cycles indefinitely for $|x^0| = \alpha$. Both these situations are illustrated in the right plot and the bottom plot of Fig. 1.15., respectively.

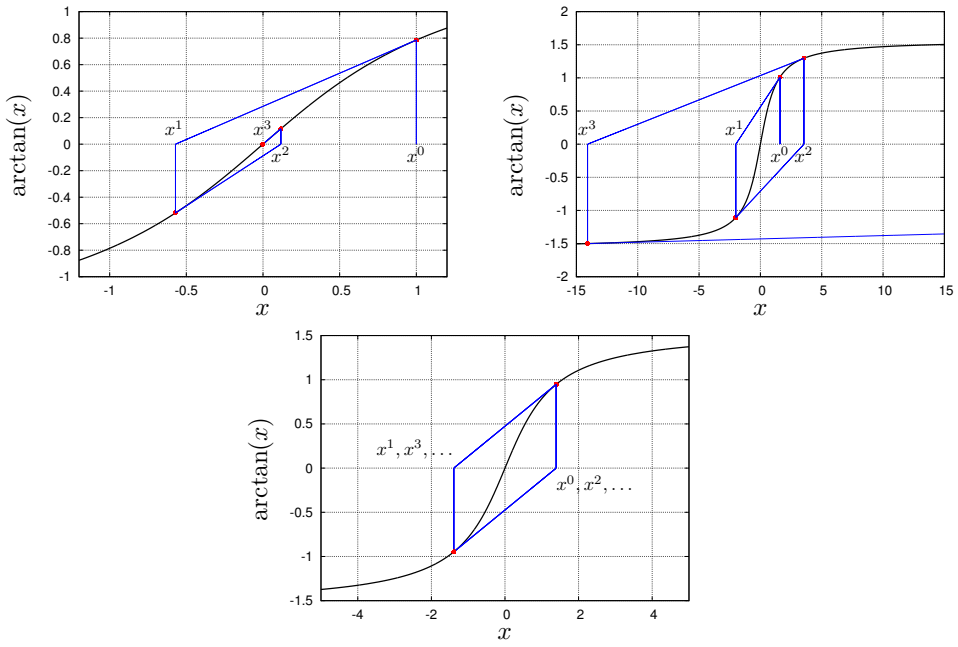


Figure 1.15. Illustration of Newton's algorithm in Example 1.69. Left plot: convergent iterates; right plot: divergent iterates; bottom plot: iterates cycle indefinitely.

1.8.3 Unconstrained Optimization

We now turn to a description of basic techniques used for iteratively solving unconstrained problems of the form:

$$\text{minimize: } f(\mathbf{x}); \quad \mathbf{x} \in \mathbb{R}^{n_x}.$$

Many unconstrained optimization algorithms work along the same lines. Starting from an initial point, a direction of movement is determined according to a fixed rule, and then a move is made in that direction so that the objective function value is reduced; at the new point, a new direction is determined and the process is repeated. The main difference between these algorithms rest with the rule by which successive directions of movement are selected. A distinction is usually made between those algorithms which determine the search direction without using gradient information (*gradient-free* methods), and those using gradient (and higher-order derivatives) information (*gradient-based* methods). Here, we shall focus our attention on the latter class of methods, and more specifically on Newton-like algorithms.

Throughout this subsection, we shall assume that the objective function f is twice continuously differentiable. By Theorem 1.22, a necessary condition for $\mathbf{x}^\star$ to be a local minimum of f is $\boldsymbol{\nabla} f(\mathbf{x}^\star) = \mathbf{0}$. Hence, the idea is to devise an iterative scheme that finds a point satisfying the foregoing condition. Following the techniques discussed earlier in §1.8.2, this can be done by using a Newton-like algorithm, with $\boldsymbol{\phi}$ corresponding to the gradient $\boldsymbol{\nabla} f$ of f, and $\boldsymbol{\nabla}\boldsymbol{\phi}$ to its Hessian matrix H.

At each iteration, a new iterate $\mathbf{x}^{k+1}$ is obtained such that the linear approximation to the gradient at that point is zero,

$$\boldsymbol{\nabla}\mathbf{f}(\mathbf{x}^{k+1}) \;=\; \boldsymbol{\nabla}\mathbf{f}(\mathbf{x}^k) + \mathsf{H}(\mathbf{x}^k)\left[\mathbf{x}^{k+1} - \mathbf{x}^k\right] \;=\; \mathbf{0}.$$

The linear approximation yields the Newton search direction

$$\mathbf{d}^k \;:=\; \mathbf{x}^{k+1} - \mathbf{x}^k \;=\; -[\mathsf{H}(\mathbf{x}^k)]^{-1}\boldsymbol{\nabla} f(\mathbf{x}^k). \tag{1.28}$$

As discussed in §1.8.2, if it converges, Newton's method exhibits a quadratic rate of convergence when the Hessian matrix H is nonsingular at the solution point. However, since the Newton iteration is based on finding a zero of the gradient vector, there is no guarantee that the step will move towards a local minimum, rather than a saddle point or even a maximum. To preclude this, the Newton steps should be taken *downhill*, i.e., the following descent condition should be satisfied at each iteration,

$$\boldsymbol{\nabla} f(\mathbf{x}^k)^{\mathsf{T}}\mathbf{d}^k \;<\; 0.$$

Interestingly enough, with the Newton direction (1.28), the descent condition becomes

$$\boldsymbol{\nabla} f(\mathbf{x}^k)^{\mathsf{T}}\mathsf{H}(\mathbf{x}^k)^{-1}\boldsymbol{\nabla} f(\mathbf{x}^k) \;>\; 0.$$

That is, a sufficient condition to obtain a descent direction at $\mathbf{x}^k$ is that the Hessian matrix $\mathsf{H}(\mathbf{x}^k)$ be positive definite. Moreover, if $\mathsf{H}(\mathbf{x}^\star)$ is positive definite at a local minimizer $\mathbf{x}^\star$ of f, then the Newton iteration converges to $\mathbf{x}^\star$ when started sufficiently close to $\mathbf{x}^\star$. (Recall that, by Theorem 1.28, positive definiteness of $\mathsf{H}(\mathbf{x}^\star)$ is a sufficient condition for a local minimum of f to be a strict local minimum.)

We now discuss two important improvements to Newton's method, which are directly related to the issues discussed in §1.8.2, namely (i) the lack of global convergence, and (ii) computational efficiency.

1.8.3.1 Globalization Strategies Up to this point, the development has focused on the application of Newton's method. However, even in the simplest one-dimensional applications, Newton's method has deficiencies (see, e.g., Example 1.69). Methods for correcting global convergence deficiencies are referred to as *globalization* strategies. It should

be stressed than an efficient globalization strategy should only alter the iterates when a problem is encountered, but it should *not* impede the ultimate behavior of the method, i.e., the quadratic convergence of a Newton's method should be retained.

In unconstrained optimization, one can detect problems in a very simple fashion, by monitoring whether the next iterate $\mathbf{x}^{k+1}$ satisfies a descent condition with respect to the actual iterate $\mathbf{x}^k$, e.g., $f(\mathbf{x}^{k+1}) < f(\mathbf{x}^k)$. Then, either one of two globalization strategies can be used to correct the Newton step. The first strategy, known as *line search* method, is to alter the *magnitude* of the step; the second one, known as *trust region* method, is to modify both the step magnitude and *direction*. We shall only concentrate on the former class of globalization strategies subsequently.

A line search method proceeds by replacing the *full* Newton step $\mathbf{x}^{k+1} = \mathbf{x}^k + \mathbf{d}^k$ with

$$\mathbf{x}^{k+1} \;=\; \mathbf{x}^k + \alpha \mathbf{d}^k,$$

where the step-length $\alpha \geq 0$ is chosen such that the objective function is reduced,

$$f(\mathbf{x}^k + \alpha \mathbf{d}^k) \;<\; f(\mathbf{x}^k).$$

Clearly, the resulting minimization problem can be solved by any one-dimensional *exact* minimization technique (e.g., Newton's method itself). However, such techniques are costly in the sense that they often require many iterations to converge and, therefore, many function (or even gradient) evaluations.

In response to this, most modern algorithms implement so-called *inexact line search criteria*, which aim to find a step-length α giving an "acceptable" decrease in the objective function. Note that sacrificing accuracy, we might impair the convergence of the overall algorithm that iteratively employs such a line search. However, by adopting a line search that guarantees a sufficient degree of descent in the objective function, the convergence of the overall algorithm can still be established.

We now describe one popular definition of an acceptable step-length know as *Armijo's rule*; other popular approaches are the *quadratic* and *cubic fit* techniques, as well as Wolfe's and Glodstein's tests. Armijo's rule is driven by two parameters $0 < \kappa_1 < 1$ and $\kappa_2 > 1$, which respectively manage the acceptable step-length from being too large or too small. (Typical vales are $\kappa_1 = 0.2$ and $\kappa_2 = 2$.) Define the line search function $\ell(\alpha) := f(\mathbf{x}^k + \alpha \mathbf{d}^k)$, for $\alpha \geq 0$, and consider the modified first-order approximation $\hat{\ell}(\alpha) := \ell(0) + \kappa_1 \alpha \ell'(0)$. A step-length $\bar{\alpha} \in (0, 1)$ is deemed acceptable if the following conditions hold:

$$\ell(\bar{\alpha}) \;\leq \hat{\ell}(\bar{\alpha}) \tag{1.29}$$

$$\ell(\kappa_2 \bar{\alpha}) \;\geq \hat{\ell}(\kappa_2 \bar{\alpha}). \tag{1.30}$$

The condition (1.29) prevents the step-length $\bar{\alpha}$ from being too large, whereas the condition (1.30) prevents $\bar{\alpha}$ from being too small. The acceptable region defined by the Armijo's rule is shown in Fig. 1.16. below.

1.8.3.2 Recursive Updates Another limitation of Newton's method when applied to unconstrained optimization problems is that the Hessian matrix of the objective function is needed at each iteration, then a linear system must be solved for obtaining the search direction. For many applications, this can be a costly computational burden. In response to this, *quasi-Newton* methods attempt to construct this information recursively. However, by so doing, the quadratic rate of convergence is lost.

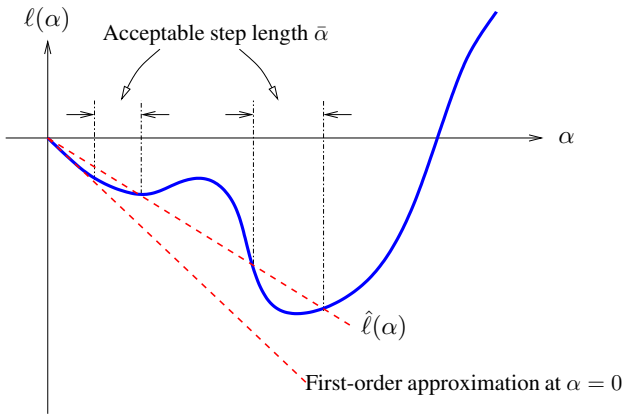


Figure 1.16. Illustration of Armijo's rule

The basic idea for many quasi-Newton methods is that two successive iterates $\mathbf{x}^k$, $\mathbf{x}^{k+1}$, together with the corresponding gradients $\boldsymbol{\nabla} f(\mathbf{x}^k)$, $\boldsymbol{\nabla} f(\mathbf{x}^{k+1})$, yield curvature information by means of the first-order approximation relation

$$\boldsymbol{\nabla} f(\mathbf{x}^{k+1}) \;=\; \boldsymbol{\nabla} f(\mathbf{x}^k) + \mathsf{H}(\mathbf{x}^k)\boldsymbol{\delta}^{\mathbf{k}} + \text{h.o.t.},$$

with $\boldsymbol{\delta}^k := \mathbf{x}^{k+1} - \mathbf{x}^k$. In particular, given n_x linearly independent iteration increments $\boldsymbol{\delta}^0, \ldots, \boldsymbol{\delta}^{n_x-1}$, an approximation of the Hessian matrix can be obtained as

$$\mathsf{H}(\mathbf{x}^{n_x}) \;\approx\; \left[\begin{array}{ccc} \boldsymbol{\gamma}^0 & \cdots & \boldsymbol{\gamma}^{n_x-1} \end{array}\right] \left[\begin{array}{ccc} \boldsymbol{\delta}^{\mathbf{0}} & \cdots & \boldsymbol{\delta}^{n_x-1} \end{array}\right]^{-1},$$

or for the inverse Hessian matrix as

$$\mathsf{H}(\mathbf{x}^{n_x})^{-1} \;\approx\; \left[\begin{array}{ccc} \boldsymbol{\delta}^{\mathbf{0}} & \cdots & \boldsymbol{\delta}^{n_x-1} \end{array}\right] \left[\begin{array}{ccc} \boldsymbol{\gamma}^0 & \cdots & \boldsymbol{\gamma}^{n_x-1} \end{array}\right]^{-1},$$

where $\boldsymbol{\gamma}^k := \boldsymbol{\nabla} f(\mathbf{x}^{k+1}) - \boldsymbol{\nabla} f(\mathbf{x}^k)$.

Note that when the objective function is quadratic, the previous relations are exact. Many interesting quasi-Newton methods use similar ways, although more sophisticated, to construct an approximate Hessian matrix B^k that progressively approaches the inverse Hessian. One of the most popular class of quasi-Newton methods (known as the *Broyden family*) proceeds as follows:

$$\begin{aligned} \mathsf{B}^{k+1} \;:=\; & \mathsf{B}^k + \frac{\boldsymbol{\delta}^k \boldsymbol{\delta}^{k\mathsf{T}}}{\boldsymbol{\delta}^{k\mathsf{T}} \boldsymbol{\gamma}^k} - \frac{\mathsf{B}^k \boldsymbol{\gamma}^k \boldsymbol{\gamma}^{k\mathsf{T}} \mathsf{B}^k}{\boldsymbol{\gamma}^{k\mathsf{T}} \mathsf{B}^k \boldsymbol{\gamma}^k} \\ & + \xi \boldsymbol{\gamma}^{k\mathsf{T}} \mathsf{B}^k \boldsymbol{\gamma}^k \left(\frac{\boldsymbol{\delta}^k}{\boldsymbol{\delta}^{k\mathsf{T}} \boldsymbol{\gamma}^k} - \frac{\mathsf{B}^k \boldsymbol{\gamma}^k}{\boldsymbol{\gamma}^{k\mathsf{T}} \mathsf{B}^k \boldsymbol{\gamma}^k} \right) \left(\frac{\boldsymbol{\delta}^k}{\boldsymbol{\delta}^{k\mathsf{T}} \boldsymbol{\gamma}^k} - \frac{\mathsf{B}^k \boldsymbol{\gamma}^k}{\boldsymbol{\gamma}^{k\mathsf{T}} \mathsf{B}^k \boldsymbol{\gamma}^k} \right)^{\mathsf{T}}, \end{aligned} \tag{1.31}$$

where $0 \leq \xi \leq 1$. It is easily seen that when supplemented with a line search strategy, $\mathbf{d}^{k\mathsf{T}} \boldsymbol{\gamma}^k < 0$ at each k, and hence the Hessian matrix approximations are guaranteed to exist. Moreover, it can be shown that the successive approximates remain are positive-definite provided that B^0 is itself positive-definite.

By setting $\xi = 0$, (1.31) yields the *Davidon-Fletcher-Powell (DFP) method*, which is historically the first quasi-Newton method, while setting $\xi = 1$ gives the *Broyden-Fletcher-Goldfard-Shanno (BFGS) method*, for which there is substantial evidence that it is the best general purpose quasi-Newton method currently known.

1.8.3.3 Summary A Newton-like algorithm including both a line search method (Armijo's rule) and Hessian recursive update (DFP update) is as follows:

Initialization Step

Let $\varepsilon > 0$ be a termination scalar, and choose an initial point $\mathbf{x}^0 \in \mathbb{R}^{n_x}$ and a symmetric, definite positive matrix $\mathsf{B}^0 \in \mathbb{R}^{n_x \times n_x}$. Let $k = 0$, and go to the main step.

Main Step

1. *Search Direction* – Obtain the search direction from $\mathbf{d}^k = -\mathsf{B}^k \boldsymbol{\nabla} f(\mathbf{x}^k)$.
2. *Line Search* – Find a step α^k satisfying Armijo's conditions (1.29,1.30).
3. *Update* – Compute the new estimates

$$\begin{aligned} \mathbf{x}^{k+1} &:= \mathbf{x}^k + \alpha^k \mathbf{d}^k \\ \mathsf{B}^{k+1} &:= \mathsf{B}^k + \frac{\boldsymbol{\delta}^k \boldsymbol{\delta}^{k\mathsf{T}}}{\boldsymbol{\delta}^{k\mathsf{T}} \boldsymbol{\gamma}^k} - \frac{\mathsf{B}^k \boldsymbol{\gamma}^k \boldsymbol{\gamma}^{k\mathsf{T}} \mathsf{B}^k}{\boldsymbol{\gamma}^{k\mathsf{T}} \mathsf{B}^k \boldsymbol{\gamma}^k}, \end{aligned}$$

   with $\boldsymbol{\delta}^k := \mathbf{x}^{k+1} - \mathbf{x}^k$ and $\boldsymbol{\gamma}^k := \boldsymbol{\nabla} f(\mathbf{x}^{k+1}) - \boldsymbol{\nabla} f(\mathbf{x}^k)$.
4. If $\|\boldsymbol{\nabla} f(\mathbf{x}^{k+1})\| < \varepsilon$, stop; otherwise, replace $k \leftarrow k + 1$, and go to step 1.

The standard unconstrained optimization algorithm in the version 3.0.2 of the Optimization Toolbox in MATLAB® is an implementation of quasi-Newton's method, with DFP or BFGS update, and a line search strategy. (See MATLAB® help pages for more information about the algorithm and the function `fminunc`.)

Example 1.70. Consider the problem to find a minimum to Rosenbrock's function

$$f(x) \;=\; (1 - x_1)^2 + c\left(x_2 - x_1^2\right)^2,$$

for $\mathbf{x} \in \mathbb{R}^2$, with $c := 105$. We solved this problem using the function `fminunc` of the Optimization Toolbox in MATLAB®. The M-files are as follows:

```
clear all;
x0 = [ 5; 5 ];
options = optimset('GradObj', 'on', 'Display', 'iter', ...
                   'DerivativeCheck', 'on', 'LargeScale', 'off', ...
                   'HessUpdate', 'bfgs', 'Diagnostics', 'on', ...
                   'LineSearchType', 'cubicpoly', 'tolX', 1e-10, ...
                   'tolFun', 1e-10)
c = 105;
[xopt, fopt, iout] = fminunc( @(x) exm1(x,c), x0, options );
```

```
%%%%%%%%%%%%%% FUNCTION TO BE MINIMIZED %%%%%%%%%%%%%%
% ROSENBROCK FUNCTION: f(x,y) := (1-x)^2+c*(y-x^2)^2
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [f,g] = exm1(x,c)
  f = (1-x(1))^2 + c*(x(2)-x(1)^2)^2; % function
  if nargout > 1
    g = [ -2*(1-x(1)) + 2*c*(x(2)-x(1)^2)*(-2*x(1)) % gradient
          2*c*(x(2)-x(1)^2) ];
  end
end
```

The results are shown in Fig. 1.17. Observe the slow convergence of the iterates far from the optimal solution $\mathbf{x}^\star = (1, 1)$, but the very fast convergence in the vicinity of $\mathbf{x}^\star$.

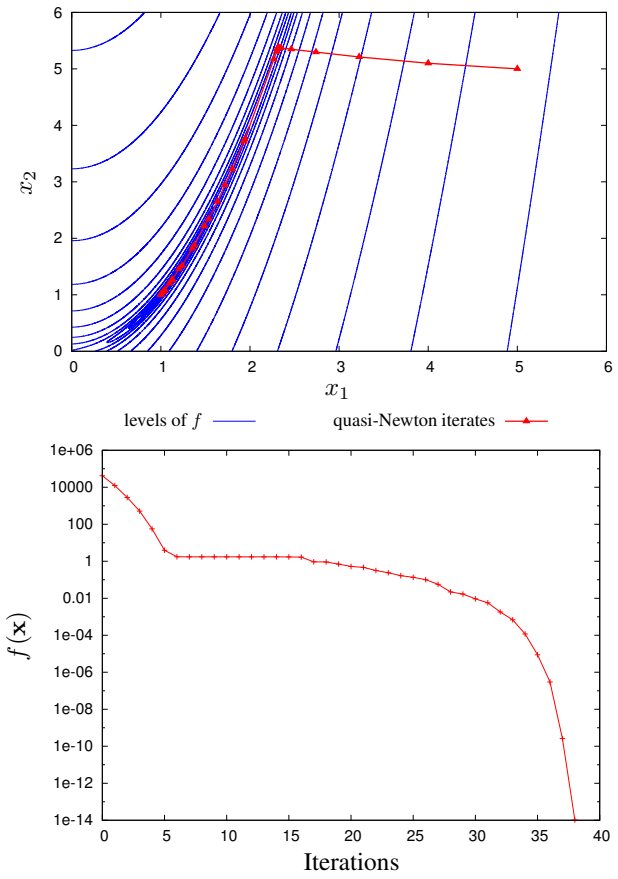


Figure 1.17. Illustration of quasi-Newton's algorithm for Rosenbrock's function in Example 1.70.

1.8.4 Constrained Nonlinear Optimization

In this subsection, we turn our attention to algorithms for iteratively solving constrained problems of the form:

$$\text{minimize: } f(\mathbf{x}); \quad \mathbf{x} \in X \subset \mathbb{R}^{n_x}.$$

Many modern deterministic algorithms for constrained NLP problems are based on the (rather natural) principle that, instead of solving a difficult problem directly, one had better solve a sequence of simpler, but related, subproblems, which converges to a solution of the original problem either in a finite number of steps or in the limit. Working along these lines, two classes of algorithms can be distinguished for solution of NLP problems with equality and/or inequality constraints. On the one hand, *penalty function* and *interior-point methods* consist of solving the problem as a sequence of unconstrained problems (or problems with simple constraints), so that algorithms for unconstrained optimization can be used. These methods, which do not rely on the KKT theory described earlier in §1.5 through §1.7, shall be briefly presented in §1.8.4.1 and §1.8.4.2. On the other hand, *Newton-like methods* solve NLP problems by attempting to find a point satisfying the necessary conditions of optimality (KKT conditions in general). Successive quadratic programming (SQP), which shall be presented in §1.8.4.3, represents one such class of methods.

1.8.4.1 Penalty Function Methods Methods using penalty functions transform a constrained problem into a single unconstrained problem or a sequence of unconstrained problems. This is done by placing the constraints into the objective function via a penalty parameter in a way that penalizes any violation of the constraints. To illustrate it, consider the NLP problem

$$\begin{array}{ll} \min\limits_{\mathbf{x}} & f(\mathbf{x}) \\ \text{s.t.} & \mathbf{g}(\mathbf{x}) \leq \mathbf{0} \\ & \mathbf{h}(\mathbf{x}) = \mathbf{0} \\ & \mathbf{x} \in X \end{array} \tag{1.32}$$

where X is a subset of $\mathbb{R}^{n_x}$, $\mathbf{x}$ is a vector of n_x components $x_1, \ldots, x_{n_x}$, and $f : X \to \mathbb{R}$, $\mathbf{g} : X \to \mathbb{R}^{n_g}$ and $\mathbf{h} : X \to \mathbb{R}^{n_h}$ are defined on X.

In general, a suitable penalty function $\alpha(\mathbf{x})$ for problem (1.32) is defined by

$$\alpha(\mathbf{x}) \; = \; \sum_{k=1}^{n_g} \phi[g_k(\mathbf{x})] + \sum_{k=1}^{n_h} \psi[h_k(\mathbf{x})], \tag{1.33}$$

where ϕ and ψ are continuous functions satisfying the conditions

$$\left\{ \begin{array}{ll} \phi(z) = 0 & \text{if } z \leq 0 \\ \phi(z) > 0 & \text{otherwise} \end{array} \right. \quad \text{and} \quad \left\{ \begin{array}{ll} \psi(z) = 0 & \text{if } z = 0 \\ \psi(z) > 0 & \text{otherwise} \end{array} \right. \tag{1.34}$$

Typically, ϕ and ψ are of the forms

$$\phi(z) \; = \; (\max\{0, z\})^p \quad \text{and} \quad \psi(z) \; = \; |z|^p,$$

with p a positive integer (taking $p \geq 2$ provides continuously differentiable penalty functions). The function $f(\mathbf{x}) + \mu\alpha(\mathbf{x})$ is referred to as the *auxiliary function*.

Example 1.71. Consider the problem to minimize $f(x) = x$, subject to $g(x) = -x+2 \leq 0$. It is immediately evident that the optimal solution lies at the point $x^\star = 2$, and has objective value $f(x^\star) = 2$.

Now, consider the penalty problem to minimize $f(x)+\mu\alpha(x) = x+\mu\max\{0, 2-x\}^2$ in $\mathbb{R}$, where μ is a large number. Note first that for any μ, the auxiliary function is convex. Thus, a necessary and sufficient condition for optimality is that the gradient of $f(x)+\mu\alpha(x)$

be equal to zero, yielding $x^\mu = 2 - \frac{1}{2\mu}$. Thus, the solution of the penalty problem can be made arbitrarily close to the solution of the original problem by choosing μ sufficiently large. Moreover, $f(x^\mu) + \mu\alpha(x^\mu) = 2 - \frac{1}{4\mu}$, which can also be made arbitrarily close to $f(x^\star)$ by taking μ sufficiently large. These considerations are illustrated in Fig. 1.18. below.

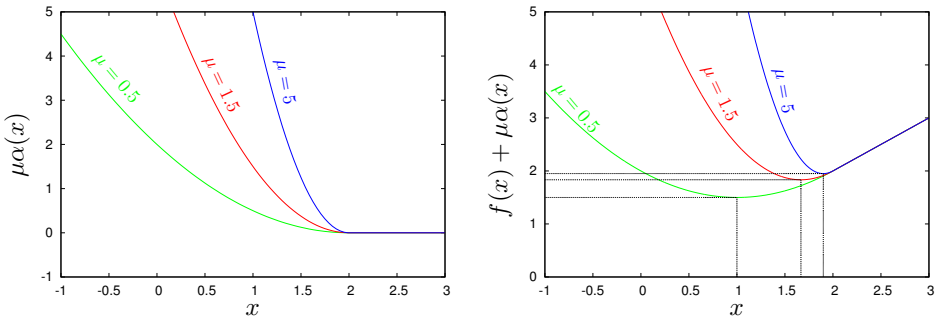


Figure 1.18. Illustration of the penalty (left plot) and auxiliary (right plot) functions in Example 1.71.

The conclusions of Example 1.71 that the solution of the penalty problem can be made arbitrarily close to the solution of the original problem, and the optimal auxiliary function value arbitrarily close to the optimal objective value, by choosing μ sufficiently large, is formalized in the following:

Theorem 1.72. *Consider the NLP problem* (1.32)*, where* f, $\mathbf{g}$ *and* $\mathbf{h}$ *are continuous functions on* $\mathbb{R}^{n_x}$ *and* X *is a nonempty convex set in* $\mathbb{R}^{n_x}$*. Suppose that* (1.32) *has a feasible solution, and let* α *be a continuous function given by (1.33,1.34). Suppose further that for each* μ*, there exists a solution* $\mathbf{x}^\mu \in X$ *to the problem* $\min\{f(\mathbf{x}) + \mu\alpha(\mathbf{x}) : \mathbf{x} \in X\}$*, and that* $\{\mathbf{x}^\mu\}$ *is contained in a compact subset of* X*. Then,*

$$\min\{f(\mathbf{x}) : \mathbf{g}(\mathbf{x}) \leq 0, \mathbf{h}(\mathbf{x}) = \mathbf{0}, \mathbf{x} \in X\} \;=\; \sup_{\mu \geq 0} \theta(\mu) \;=\; \lim_{k \to \infty} \theta(\mu),$$

with $\theta(\mu) := f(\mathbf{x}^\mu) + \mu\alpha(\mathbf{x}^\mu)$*. Furthermore, the limit* $\bar{\mathbf{x}}$ *of any convergent subsequence of* $\{\mathbf{x}^\mu\}$ *is an optimal solution to the original problem and* $\mu\alpha(\mathbf{x}^\mu) \to 0$ *as* $\mu \to \infty$*.*

Proof. See [6, Theorem 9.2.2] for a proof. □

Note that the assumption that X is compact is necessary, for it possible that the optimal objective values of the original and penalty problems are not equal otherwise. Yet, this assumption is not very restrictive in most practical cases as the variables usually lie between finite upper and lower bounds. Note also that no restriction is imposed on f, $\mathbf{g}$ and $\mathbf{h}$ other than continuity. However, the application of an efficient solution procedure for the (unconstrained) auxiliary problems may impose additional restriction on these functions (see §1.8.3).

Under the conditions that (i) f, $\mathbf{g}$, $\mathbf{h}$ in (1.32) and ϕ, ψ in (1.33,1.34) are continuously differentiable, and (ii) $\bar{x}$ is a regular point (see Definitions 1.36 and 1.47), the solution to

the penalty problem can be used to recover the Lagrange multipliers associated with the constraints at optimality. In the particular case where $X = \mathbb{R}^{n_x}$, we get

$$\nu_i^\mu = \mu\phi' [g_i(\mathbf{x}^\mu)] \quad \forall i \in \mathcal{A}(\bar{\mathbf{x}}) \tag{1.35}$$
$$\lambda_i^\mu = \mu\psi' [h_i(\mathbf{x}^\mu)] \quad \forall i = 1, \ldots, n_h. \tag{1.36}$$

The larger μ, the better the approximation of the Lagrange multipliers,

$$\boldsymbol{\nu}^\mu \to \boldsymbol{\nu}^\star \text{ and } \boldsymbol{\lambda}^\mu \to \boldsymbol{\lambda}^\star \text{ as } \mu \to \infty.$$

Example 1.73. Consider the same problem as in Example 1.71. The auxiliary function $f(x) + \mu\alpha(x) = x + \mu \max\{0, 2 - x\}^2$ being continuously differentiable, the Lagrange multiplier associated to the inequality constraint $g(x) = -x + 2 \leq 0$ can be recovered as $\nu^\mu = 2\mu \max\{0, 2 - x^\mu\} = 1$ (assuming $\mu > 0$). Note that the exact value of the Lagrange multiplier is obtained for each $\mu > 0$ here, because g is a linear constraint.

From a computational viewpoint, superlinear convergence rates might be achievable, in principle, by applying Newton's method (or its variants such as quasi-Newton methods). Yet, one can expect ill-conditioning problems when μ is taken very large in the penalty problem. With a large μ, more emphasis is placed on feasibility, and most procedures for unconstrained optimization will move quickly towards a feasible point. Even though this point may be far from the optimum, both slow convergence and premature termination can occur due to very small step size and finite precision computations (round-off errors).

As a result of the above mentioned difficulties associated with large penalty parameters, most algorithms using penalty functions employ a sequence of increasing penalty parameters. With each new value of the penalty parameter, an optimization technique is employed, starting with the optimal solution corresponding to the previously chosen parameters value. such an approach is often referred to as *sequential unconstrained minimization* (SUM) technique. This way, a sequence of infeasible points is typically generated, whose limit is an optimal solution to the original problem (hence the term *exterior penalty function approach*).

To conclude our discussion on the penalty function approach, we give an algorithm to solve problem (1.32), where the penalty function used is of the form specified in (1.33,1.34).

Initialization Step

Let $\varepsilon > 0$ be a termination scalar, and choose an initial point $\mathbf{x}^0$, a penalty parameter $\mu^0 > 0$, and a scalar $\beta > 1$. Let $k = 0$ and go to the main step.

Main Step

1. Starting with $\mathbf{x}^k$, get a solution to the problem

$$\mathbf{x}^{k+1} \in \arg\min\{f(\mathbf{x}) + \mu^k\alpha(\mathbf{x}) : \mathbf{x} \in X\}$$

2. If $\mu^k\alpha(\mathbf{x}^{k+1}) < \varepsilon$, stop; otherwise, let $\mu^{k+1} = \beta\mu^k$, replace $k \leftarrow k + 1$, and go to step 1.

1.8.4.2 Interior-Point Methods Similar to penalty functions, *barrier functions* can also be used to transform a constrained problem into an unconstrained problem (or into a sequence of unconstrained problems). These functions act as a barrier and prevent the iterates from leaving the feasible region. If the optimal solution occurs at the boundary of the feasible domain, the procedure moves from the interior to the boundary of the domain, hence the name *interior-point methods*. To illustrate these methods, consider the NLP problem

$$\begin{array}{ll} \min\limits_{\mathbf{x}} & f(\mathbf{x}) \\ \text{s.t.} & \mathbf{g}(\mathbf{x}) \leq \mathbf{0} \\ & \mathbf{x} \in X \end{array} \tag{1.37}$$

where X is a subset of $\mathbb{R}^{n_x}$, and $f : X \to \mathbb{R}$, $\mathbf{g} : X \to \mathbb{R}^{n_g}$ are continuous on $\mathbb{R}^{n_x}$. Note that equality constraints, if any, should be accommodated within the set X. (In the case of affine equality constraints, one can possibly eliminate them after solving for some variables in terms of the others, thereby reducing the dimension of the problem.) The reason why this treatment is necessary is because barrier function methods require the set $\{\mathbf{x} \in \mathbb{R}^{n_x} : \mathbf{g}(\mathbf{x}) < 0\}$ to be **nonempty**; this would obviously be not possible if the equality constraints $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ were accommodated within the set of inequalities as $\mathbf{h}(\mathbf{x}) \leq \mathbf{0}$ and $\mathbf{h}(\mathbf{x}) \geq \mathbf{0}$.

A barrier problem formulates as

$$\begin{array}{ll} \min\limits_{\mathbf{x}} & \theta(\mu) \\ \text{s.t.} & \mu \geq 0, \end{array} \tag{1.38}$$

where $\theta(\mu) := \inf\{f(\mathbf{x}) + \mu b(\mathbf{x}) : \mathbf{g}(\mathbf{x}) < \mathbf{0}, \mathbf{x} \in X\}$. Ideally, the barrier function b should take value zero on the region $\{\mathbf{x} : \mathbf{g}(\mathbf{x}) \leq 0\}$, and value ∞ on its boundary. This would guarantee that the iterates do not leave the domain $\{\mathbf{x} : \mathbf{g}(\mathbf{x}) \leq 0\}$ provided the minimization problem started at an interior point. However, this discontinuity poses serious difficulties for any computational procedure. Therefore, this ideal construction of b is replaced by the more realistic requirement that b be nonnegative and continuous over the region $\{\mathbf{x} : \mathbf{g}(\mathbf{x}) \leq 0\}$ and approach infinity as the boundary is approached from the interior:

$$b(\mathbf{x}) = \sum_{k=1}^{n_g} \phi[g_k(\mathbf{x})], \tag{1.39}$$

where ϕ is a continuous function over $\{z : z < 0\}$ that satisfies the conditions

$$\begin{cases} \phi(z) \geq 0 & \text{if } z < 0 \\ \lim_{z \to 0^-} \phi(z) = +\infty \ . \end{cases} \tag{1.40}$$

In particular, μb approaches the ideal barrier function described above as μ approaches zero.

Typically barrier functions are

$$b(\mathbf{x}) \ = \ -\sum_{k=1}^{n_g} \frac{1}{g_k(\mathbf{x})} \quad \text{or} \quad b(\mathbf{x}) \ = \ -\sum_{k=1}^{n_g} \ln[\min\{1, -g_k(\mathbf{x})\}].$$

The following barrier function, known as *Frisch's logarithmic barrier function*, is also widely used

$$b(\mathbf{x}) \ = \ -\sum_{k=1}^{n_g} \ln[-g_k(\mathbf{x})].$$

The function $f(\mathbf{x}) + \mu b(\mathbf{x})$ is referred to as the *auxiliary function*.

Given $\mu > 0$, evaluating $\theta(\mu) = \inf\{f(\mathbf{x}) + \mu b(\mathbf{x}) : \mathbf{g}(\mathbf{x}) < 0, \mathbf{x} \in X\}$ seems no simpler than solving the original problem because of the constraint $\mathbf{g}(\mathbf{x}) < 0$. However, starting the optimization from a point in the region $S := \{\mathbf{x} : \mathbf{g}(\mathbf{x}) < 0\} \cap X$ yields an optimal point in S, even when the constraint $\mathbf{g}(\mathbf{x}) < 0$ is ignored. This is because b approaches infinity as the iterates approach the boundary of $\{\mathbf{x} : \mathbf{g}(\mathbf{x}) \leq 0\}$ from within S, hence preventing them from leaving the set S. This is formalized in the following:

Theorem 1.74. *Consider the NLP problem* (1.37)*, where f and $\mathbf{g}$ are continuous functions on $\mathbb{R}^{n_x}$ and X is a nonempty convex set in $\mathbb{R}^{n_x}$. Suppose that* (1.37) *has an optimal solution $\mathbf{x}^\star$ with the property that, given any neighborhood $\mathcal{B}_\eta(\mathbf{x}^\star)$ around $\mathbf{x}^\star$, there exists an $\mathbf{x} \in X \cap \mathcal{B}_\eta(\mathbf{x}^\star)$ such that $\mathbf{g}(\mathbf{x}) < \mathbf{0}$. Suppose further that for each μ, there exists a solution $\mathbf{x}^\mu \in X$ to the problem $\min\{f(\mathbf{x}) + \mu b(\mathbf{x}) : \mathbf{x} \in X\}$. Then,*

$$\min\{f(\mathbf{x}) : \mathbf{g}(\mathbf{x}) \leq 0, \mathbf{x} \in X\} \;=\; \lim_{\mu \downarrow 0} \theta(\mu) \;=\; \inf_{\mu > 0} \theta(\mu),$$

with $\theta(\mu) := f(\mathbf{x}^\mu) + \mu b(\mathbf{x}^\mu)$. Furthermore, the limit of any convergent subsequence of $\{\mathbf{x}^\mu\}$ is an optimal solution to the original problem, and $\mu b(\mathbf{x}^\mu) \to 0$ as $\mu \to 0$.

Proof. See [6, Theorem 9.4.3] for a proof. □

Under the conditions that (i) f, $\mathbf{g}$ in (1.32) and ϕ in (1.33,1.34) are continuously differentiable, and (ii) $\bar{x}$ is a regular point (see Definitions 1.36 and 1.47), the solution to the barrier function problem can be used to recover the Lagrange multipliers associated with the constraints at optimality. In the particular case where $X = \mathbb{R}^{n_x}$, we get

$$\nu_i^\mu \;= \mu \phi' \left[g_i(\mathbf{x}^\mu) \right] \quad \forall i \in \mathcal{A}(\bar{\mathbf{x}}). \tag{1.41}$$

The approximation of the Lagrange multipliers, gets better as μ gets closer to 0,

$$\boldsymbol{\nu}^\mu \to \boldsymbol{\nu}^\star \text{ as } \mu \to 0^+.$$

Example 1.75. Consider the problem to minimize $f(x) = x$, subject to $g(x) = -x+2 \leq 0$, the solution of which lies at the point $x^\star = 2$ with objective value $f(x^\star) = 2$.

Now, consider the barrier function problem to minimize $f(x) + \mu b(x) = x - \frac{\mu}{2-x}$ in $\mathbb{R}$, where μ is a large number. Note first that for any μ, the auxiliary function is convex. Thus, a necessary and sufficient condition for optimality is that the gradient of $f(x) + \mu b(x)$ be equal to zero, yielding $x^\mu = 2 + \sqrt{\mu}$ (assuming $\mu > 0$). Thus, the solution of the penalty problem can be made arbitrarily close to the solution of the original problem by choosing μ sufficiently close to zero. Moreover, $f(x^\mu) + \mu b(x^\mu) = 2 - 2\sqrt{\mu}$, which can also be made arbitrarily close to $f(x^\star)$ by taking μ sufficiently close to zero. These considerations are illustrated in Fig. 1.19. below.

Regarding the Lagrange multiplier associated to the inequality constraint $g(x) = -x + 2 \leq 0$, the objective and constraint functions being continuously differentiable, an approximate value can be obtained as $\nu^\mu = \frac{\mu}{2-x^\mu}^2 = 1$. Here again, the exact value of the Lagrange multiplier is obtained for each $\mu > 0$ because g is a linear constraint.

The use of barrier functions for solving constrained NLP problems also faces several computational difficulties. First, the search must start with a point $\mathbf{x} \in X$ such that

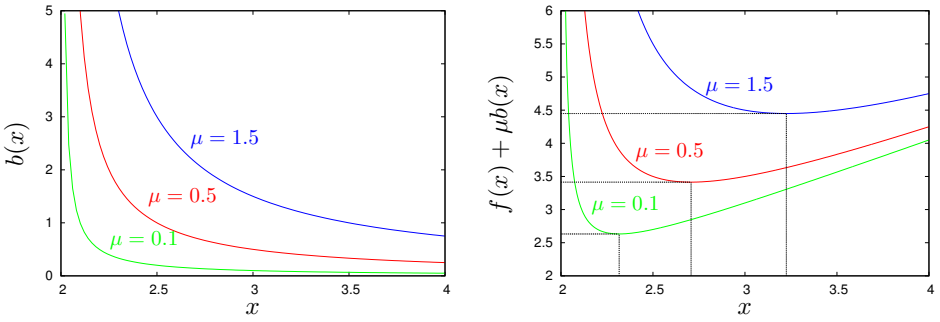


Figure 1.19. Illustration of the barrier (left plot) and auxiliary (right plot) functions in Example 1.75.

$\mathbf{g}(\mathbf{x}) < \mathbf{0}$, and finding such a point may not be an easy task for some problems. Also, because of the structure of the barrier function b, and for small values of the parameter μ, most search techniques may face serious ill-conditioning and difficulties with round-off errors while solving the problem to minimize $f(\mathbf{x}) + \mu b(\mathbf{x})$ over $\mathbf{x} \in X$, especially as the boundary of the region $\{\mathbf{x} : \mathbf{g}(\mathbf{x}) \leq 0\}$ is approached. Accordingly, interior-point algorithms employ a sequence of decreasing penalty parameters $\{\mu^k\} \to 0$ as $k \to \infty$; with each new value μ^k, an optimal solution to the barrier problem is sought by starting from the previous optimal solution. As in the exterior penalty function approach, it is highly recommended to use suitable second-order Newton or quasi-Newton methods for solving the successive barrier problems.

We describe below a scheme using barrier functions of the form (1.39,1.40) for optimizing a nonlinear programming problem such as (1.37).

Initialization Step

Let $\varepsilon > 0$ be a termination scalar, and choose an initial point $\mathbf{x}^0$ with $\mathbf{g}(\mathbf{x}^0) < 0$. Let $\mu^0 > 0$, $\beta \in (0, 1)$, $k = 0$, and go to the main step.

Main Step

1. Starting with $\mathbf{x}^k$, get a solution to the problem
$$\mathbf{x}^{k+1} \in \arg\min\{f(\mathbf{x}) + \mu^k b(\mathbf{x}) : \mathbf{x} \in X\}$$
2. If $\mu^k b(\mathbf{x}^{k+1}) < \varepsilon$, stop; otherwise, let $\mu^{k+1} = \beta\mu^k$, replace $k \leftarrow k + 1$, and go to step 1.

Note that although the constraint $\mathbf{g}(\mathbf{x}) < 0$ may be ignored, it is considered in the problem formulation as most line search methods use discrete steps, and a step could lead to a point outside the feasible region (where the value of the barrier function is a large negative number), when close to the boundary. Therefore, the problem can effectively be treated as an unconstrained optimization problem only if an explicit check for feasibility is made.

In recent years, there has been much excitement because some variants of the interior-point algorithm can be shown to be polynomial in time for many classes of convex programs. Moreover, interior-point codes are now proving to be highly competitive with codes based on other algorithms, such as SQP algorithms presented subsequently. A number of free and commercial interior-point solvers are given in Tab. 1.1. below.

Table 1.1. A number of open-source and commercial codes implementing interior-point techniques for NLP problems.

Solver	Website	Licensing	Characteristics
IPOPT	`projects.coin-or.org/Ipopt`	free, open source	line search, filter, preconditioned CG
LOQO	`www.princeton.edu/~rvdb/`	commercial	primal-dual, direct factorization
KNITRO	`www.ziena.com/knitro.htm`	commercial	trust-region, primal barrier/SQP with primal-dual scaling, direct factorization/preconditioned CG

Note first that there is currently no function implementing interior-point methods for NLP problems in the version 3.0.2 of MATLAB®'s Optimization Toolbox. The solvers listed in Tab. 1.2. are stand-alone (either in C/C++ or fortran77 programming language). However, all can be used through the modeling language AMPL (`http://www.ampl.com/`); KNITRO can also be used in MATLAB® through both the TOMLAB optimization environment (`http://tomopt.com/tomlab/`) and the modeling language GAMS (`http://www.gams.com/`).

1.8.4.3 Successive Quadratic Programming

Successive quadratic programming (SQP) methods, also known as *sequential*, or *recursive*, quadratic programming, employ Newton's method (or quasi-Newton methods) to directly solve the KKT conditions for the original problem. As a result, the accompanying subproblem turns out to be the minimization of a quadratic approximation to the Lagrangian function subject to a linear approximation to the constraints. Hence, this type of process is also known as a *projected Lagrangian*, or the *Newton-Lagrange*, approach. By its nature, this method produces both primal and dual (Lagrange multiplier) solutions.

Equality Constrained Case. To present the concept of SQP, consider first the nonlinear problem P to

$$\begin{aligned} &\text{minimize: } f(\mathbf{x}) \\ &\text{subject to: } h_i(\mathbf{x}) = 0, \quad i = 1, \ldots, n_h, \end{aligned}$$

where $\mathbf{x} \in \mathbb{R}^{n_x}$, and f, $\mathbf{h}$ are twice continuously differentiable. We shall also assume throughout that the equality constraints are linearly independent at a solution of P. (The extension for including inequality constraints is considered subsequently.)

By Theorem 1.50, the first-order necessary conditions of optimality for Problem P require a primal solution $\mathbf{x}^\star \in \mathbb{R}^{n_x}$ and a Lagrange multiplier vector $\boldsymbol{\lambda}^\star \in \mathbb{R}^{n_h}$ such that

$$\mathbf{0} = \boldsymbol{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star) = \boldsymbol{\nabla} f(\mathbf{x}^\star) + \boldsymbol{\nabla} \mathbf{h}(\mathbf{x}^\star)^{\mathsf{T}} \boldsymbol{\lambda}^\star \tag{1.42}$$

$$\mathbf{0} = \boldsymbol{\nabla}_{\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}^\star, \boldsymbol{\lambda}^\star) = \mathbf{h}(\mathbf{x}^\star), \tag{1.43}$$

where $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) := f(\mathbf{x}) + \boldsymbol{\lambda}^{\mathsf{T}} \mathbf{h}(\mathbf{x})$. Now, consider a Newton-like method to solve (1.42,1.43). Given an iterate $(\mathbf{x}^k, \boldsymbol{\lambda}^k)$, a new iterate $(\mathbf{x}^{k+1}, \boldsymbol{\lambda}^{k+1})$ is obtained by solving the first-order approximation

$$\mathbf{0} = \begin{pmatrix} \boldsymbol{\nabla}_{\mathbf{x}} \mathcal{L}(\mathbf{x}^k, \boldsymbol{\lambda}^k) \\ \boldsymbol{\nabla}_{\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}^k, \boldsymbol{\lambda}^k) \end{pmatrix} + \begin{pmatrix} \boldsymbol{\nabla}^2_{\mathbf{xx}} \mathcal{L}(\mathbf{x}^k, \boldsymbol{\lambda}^k) & \boldsymbol{\nabla} \mathbf{h}(\mathbf{x}^k)^{\mathsf{T}} \\ \boldsymbol{\nabla} \mathbf{h}(\mathbf{x}^k) & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x}^{k+1} - \mathbf{x}^k \\ \boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k \end{pmatrix}$$

to (1.42,1.43). Denoting $\mathbf{d}^k := \mathbf{x}^{k+1} - \mathbf{x}^k$, the above linear system can be rewritten as

$$\begin{pmatrix} \nabla^2_{\mathbf{xx}}\mathcal{L}(\mathbf{x}^k, \boldsymbol{\lambda}^k) & \nabla \mathbf{h}(\mathbf{x}^k)^\mathsf{T} \\ \nabla \mathbf{h}(\mathbf{x}^k) & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{d}^k \\ \boldsymbol{\lambda}^{k+1} \end{pmatrix} = - \begin{pmatrix} \nabla f(\mathbf{x}^k) \\ \mathbf{h}(\mathbf{x}^k) \end{pmatrix}, \tag{1.44}$$

which can be solved for $(\mathbf{d}^k, \boldsymbol{\lambda}^{k+1})$, if a solution exists. Setting $\mathbf{x}^{k+1} := \mathbf{x}^k + \mathbf{d}^k$, and incrementing k by 1, we can then repeat the process until $\mathbf{d}^k = \mathbf{0}$ happens to solve (1.44). When this occurs, if at all, noting (1.42,1.43), we shall have found a stationary point to Problem P.

Interestingly enough, a *quadratic programming* (QP) minimization subproblem can be employed in lieu of the foregoing linear system to find *any* optimal solution for P,

$$\begin{aligned} \min_{\mathbf{d}^k} \; & f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^\mathsf{T} \mathbf{d}^k + \frac{1}{2} \mathbf{d}^{k\mathsf{T}} \nabla^2_{\mathbf{xx}} \mathcal{L}(\mathbf{x}^k, \boldsymbol{\lambda}^k) \mathbf{d}^k \qquad \text{(QP}(\mathbf{x}^k, \boldsymbol{\lambda}^k)\text{)} \\ \text{s.t. } & h_i(\mathbf{x}) + \nabla h_i(\mathbf{x}^k)^\mathsf{T} \mathbf{d}^k = 0 \quad i = 1, \ldots, n_h. \end{aligned}$$

Note in particular that an optimum $\mathbf{d}^k$ to QP$(\mathbf{x}^k, \boldsymbol{\lambda}^k)$, if it exists, together with the set of Lagrange multipliers $\boldsymbol{\lambda}^{k+1}$ associated with the linearized constraints, is a stationary point for QP$(\mathbf{x}^k, \boldsymbol{\lambda}^k)$ and satisfies equations (1.42,1.43). That is, solving QP$(\mathbf{x}^k, \boldsymbol{\lambda}^k)$ is attractive because it tends to drive the solution towards a desirable stationary point satisfying (1.42,1.43) whenever alternatives exist. Assuming a well-behaved QP, a rudimentary SQP algorithm is as follows:

Initialization Step

Choose an initial primal/dual point $(\mathbf{x}^0, \boldsymbol{\lambda}^0)$, let $k = 0$, and go to the main step.

Main Step

1. Solve the quadratic subproblem QP$(\mathbf{x}^k, \boldsymbol{\lambda}^k)$ to obtain a solution $\mathbf{d}^k$ along with a set of Lagrange multipliers $\boldsymbol{\lambda}^{k+1}$.
2. If $\mathbf{d}^k = \mathbf{0}$, then $(\mathbf{d}^k, \boldsymbol{\lambda}^{k+1})$ satisfies the stationarity conditions (1.42,1.43) for problem P; stop. Otherwise, let $\mathbf{x}^{k+1} := \mathbf{x}^k + \mathbf{d}^k$, replace $k \leftarrow k + 1$, and go to step 1.

In the case $\mathbf{x}^\star$ is a regular stationary solution for Problem P which, together with a set of Lagrange multipliers $\boldsymbol{\lambda}^\star$, satisfies the second-order sufficiency conditions of Theorem 1.59, then the matrix

$$\mathsf{W} := \begin{pmatrix} \nabla^2_{\mathbf{xx}}\mathcal{L}(\mathbf{x}^k, \boldsymbol{\lambda}^k) & \nabla \mathbf{h}(\mathbf{x}^k)^\mathsf{T} \\ \nabla \mathbf{h}(\mathbf{x}^k) & \mathbf{0} \end{pmatrix},$$

can be shown to be nonsingular. Hence, the above rudimentary SQP algorithm exhibits a quadratic rate of convergence by Theorem 1.68.

Extension to Inequality Constrained Case. We now consider the inclusion of inequality constraints $g_i(\mathbf{x}) \leq 0$, $i = 1, \ldots, n_g$, in Problem P,

$$\begin{aligned} &\text{minimize: } f(\mathbf{x}) \\ &\text{subject to: } g_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, n_g \\ &\qquad\qquad\quad h_i(\mathbf{x}) = 0, \quad i = 1, \ldots, n_h, \end{aligned}$$

where $\mathbf{g}$ is twice continuously differentiable.

Given an iterate $(\mathbf{x}^k, \boldsymbol{\lambda}^k, \boldsymbol{\nu}^k)$, where $\boldsymbol{\lambda}^k$ and $\boldsymbol{\nu}^k \geq \mathbf{0}$ are the Lagrange multiplier estimates for the equality and inequality constraints, respectively, consider the following QP subproblem as a direct extension of QP$(\mathbf{x}^k, \boldsymbol{\lambda}^k)$:

$$\begin{aligned} \min_{\mathbf{d}^k} \; & f(\mathbf{x}^k) + \boldsymbol{\nabla} f(\mathbf{x}^k)^{\mathsf{T}} \mathbf{d}^k + \frac{1}{2} {\mathbf{d}^k}^{\mathsf{T}} \boldsymbol{\nabla}^2_{\mathbf{xx}} \mathcal{L}(\mathbf{x}^k, \boldsymbol{\lambda}^k, \boldsymbol{\nu}^k) \mathbf{d}^k \qquad (\text{QP}(\mathbf{x}^k, \boldsymbol{\lambda}^k, \boldsymbol{\nu}^k)) \\ \text{s.t. } & g_i(\mathbf{x}) + \boldsymbol{\nabla} g_i(\mathbf{x}^k)^{\mathsf{T}} \mathbf{d}^k = 0 \quad i = 1, \ldots, n_g \\ & h_i(\mathbf{x}) + \boldsymbol{\nabla} h_i(\mathbf{x}^k)^{\mathsf{T}} \mathbf{d}^k = 0 \quad i = 1, \ldots, n_h, \end{aligned}$$

where $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) := f(\mathbf{x}) + \boldsymbol{\nu}^{\mathsf{T}} \mathbf{g}(\mathbf{x}) + \boldsymbol{\lambda}^{\mathsf{T}} \mathbf{h}(\mathbf{x})$. Note that the KKT conditions for QP$(\mathbf{x}^k, \boldsymbol{\lambda}^k, \boldsymbol{\nu}^k)$ require that, in addition to primal feasibility, Lagrange multipliers $\boldsymbol{\lambda}^{k+1}$, $\boldsymbol{\nu}^{k+1}$ be found such that

$$\begin{aligned} \boldsymbol{\nabla} f(\mathbf{x}^k) + \boldsymbol{\nabla}^2_{\mathbf{xx}} \mathcal{L}(\mathbf{x}^k, \boldsymbol{\lambda}^k, \boldsymbol{\nu}^k) \mathbf{d}^k + \boldsymbol{\nabla} \mathbf{g}(\mathbf{x}^k)^{\mathsf{T}} \boldsymbol{\nu}^{k+1} + \boldsymbol{\nabla} \mathbf{h}(\mathbf{x}^k)^{\mathsf{T}} \boldsymbol{\lambda}^{k+1} &= \mathbf{0} \\ \left[\mathbf{g}(\mathbf{x}^k) + \boldsymbol{\nabla} \mathbf{g}(\mathbf{x}^k)^{\mathsf{T}} \mathbf{d}^k \right]^{\mathsf{T}} \boldsymbol{\nu}^{k+1} &= \mathbf{0}, \end{aligned}$$

with $\boldsymbol{\nu}^{k+1} \geq \mathbf{0}$ and $\boldsymbol{\lambda}^{k+1}$ unrestricted in sign. Clearly, if $\mathbf{d}^k = \mathbf{0}$, then $\mathbf{x}^k$ together with $\boldsymbol{\lambda}^{k+1}$, $\boldsymbol{\nu}^{k+1}$ yields a KKT solution to the original problem P. Otherwise, we set $\mathbf{x}^{k+1} := \mathbf{x}^k + \mathbf{d}^k$ as before, increment k by 1, and repeat the process. Regarding convergence rate, it can be shown that if $\mathbf{x}^\star$ is a regular KKT solution which, together with $\boldsymbol{\lambda}^\star$, $\boldsymbol{\nu}^\star$ satisfies the second-order sufficient conditions of Theorem 1.63, and if $(\mathbf{x}^0, \boldsymbol{\lambda}^0, \boldsymbol{\nu}^0)$ is initialized sufficiently close to $(\mathbf{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star)$, then the foregoing iterative procedure shall exhibit a *quadratic* convergence rate.

An Improved SQP Algorithm. The SQP method, as presented thus far, obviously shares the disadvantages of Newton's method: (i) it requires second-order derivatives $\boldsymbol{\nabla}^2_{\mathbf{xx}} \mathcal{L}$ to be calculated, which in addition might not be positive definite, and (ii) it lacks the global convergence property.

(i) Regarding second-order derivatives, a quasi-Newton positive definite approximation can be used for $\boldsymbol{\nabla}^2_{\mathbf{xx}} \mathcal{L}$. For example, given a positive definite approximation B^k for $\boldsymbol{\nabla}^2_{\mathbf{xx}} \mathcal{L}(\mathbf{x}^k, \boldsymbol{\lambda}^k, \boldsymbol{\nu}^k)$, the quadratic problem QP$(\mathbf{x}^k, \boldsymbol{\lambda}^k, \boldsymbol{\nu}^k)$ can be solved with $\boldsymbol{\nabla}^2_{\mathbf{xx}} \mathcal{L}(\mathbf{x}^k, \boldsymbol{\lambda}^k, \boldsymbol{\nu}^k)$ replaced by B^k. For example, an approximation of the inverse Hessian matrix can be obtained via a Broyden-like procedure (1.31), with $\boldsymbol{\gamma}^k$ given by

$$\boldsymbol{\gamma}^k := \boldsymbol{\nabla} \mathcal{L}(\mathbf{x}^{k+1}, \boldsymbol{\lambda}^{k+1}, \boldsymbol{\nu}^{k+1}) - \boldsymbol{\nabla} \mathcal{L}(\mathbf{x}^k, \boldsymbol{\lambda}^{k+1}, \boldsymbol{\nu}^{k+1}),$$

as explained earlier in §1.8.3.2. It can be shown that this modification to the rudimentary SQP algorithm, similar to the quasi-Newton modification of Newton' algorithm, looses the quadratic convergence rate property. Instead, it can be shown that the convergence is superlinear when initialized sufficiently close to a solution $(\mathbf{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star)$ that satisfies both regularity and second-order sufficiency conditions. However, this superlinear convergence rate is strongly based on the use of unit step sizes (see point (ii) below).

(ii) In order to remedy the global convergence deficiency, a globalization strategy can be used, e.g., a line search procedure (see §1.8.3.1). Unlike unconstrained optimization problems, however, the choice of a suitable line search (or *merit*) function providing

a measure of progress is not obvious in the presence of constraints. Two such popular choices of a line search function are

- The ℓ_1 *Merit Function:*

$$\ell_1(\mathbf{x};\mu) := f(\mathbf{x}) + \mu \left[\sum_{i=1}^{n_h} |h_i(\mathbf{x})| + \sum_{i=1}^{n_g} \max\{0, g_i(\mathbf{x})\} \right], \tag{1.45}$$

  which satisfies the important property that $\mathbf{x}^\star$ is a local minimizer of ℓ_1, provided $(\mathbf{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\nu}^\star)$ satisfies the second-order sufficient conditions (see Theorem 1.63) and the penalty parameter μ is so chosen that $\mu > |\lambda_i^\star|$, $i = 1, \ldots, n_h$, and $\mu > \nu_i^\star$, $i = 1, \ldots, n_g$. Yet, the ℓ_1 merit function is not differentiable at those $\mathbf{x}$ with either $g_i(\mathbf{x}) = 0$ or $h_i(\mathbf{x}) = 0$, and it can be unbounded below even though $\mathbf{x}^\star$ is a local minimizer.

- The *Augmented Lagrangian (ALAG) Merit Function:*

$$\ell_2(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}; \mu) := f(\mathbf{x}) + \sum_{i=1}^{n_h} \lambda_i h_i(\mathbf{x}) + \frac{\mu}{2} \sum_{i=1}^{n_h} [h_i(\mathbf{x})]^2 + \frac{1}{2} \sum_{i=1}^{n_g} \psi_i(\mathbf{x}, \boldsymbol{\nu}; \mu) \tag{1.46}$$

  with $\psi_i(\mathbf{x}, \boldsymbol{\nu}; \mu) := \frac{1}{\mu} \left(\max\{0, \nu_i + \mu g_i(\mathbf{x})\}^2 - \nu_i^2 \right)$, has similar properties to the ℓ_1 merit function, provided μ is chosen large enough, and is continuously differentiable (although its Hessian matrix is discontinuous). Yet, for $\mathbf{x}^\star$ to be a (local) minimizer of $\ell_2(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}; \mu)$, it is necessary that $\boldsymbol{\lambda} = \boldsymbol{\lambda}^\star$ and $\boldsymbol{\nu} = \boldsymbol{\nu}^\star$.

An SQP algorithm including the modifications discussed in (i) and (ii) is as follows:

Initialization Step

Choose an initial primal/dual point $(\mathbf{x}^0, \boldsymbol{\lambda}^0, \boldsymbol{\nu}^0)$, with $\boldsymbol{\nu}^0 \geq \mathbf{0}$, and a positive definite matrix B^0. Let $k = 0$, and go to the main step.

Main Step

1. Solve the quadratic subproblem QP$(\mathbf{x}^k, \boldsymbol{\lambda}^k, \boldsymbol{\nu}^k)$, with $\boldsymbol{\nabla}^2_{\mathbf{xx}} \mathcal{L}(\mathbf{x}^k, \boldsymbol{\lambda}^k, \boldsymbol{\nu}^k)$ replaced by B^k, to obtain a direction $\mathbf{d}^k$ along with a set of Lagrange multipliers $(\boldsymbol{\lambda}^{k+1}, \boldsymbol{\nu}^{k+1})$.
2. If $\mathbf{d}^k = \mathbf{0}$, then $(\mathbf{x}^k, \boldsymbol{\lambda}^{k+1}, \boldsymbol{\nu}^{k+1})$ satisfies the KKT conditions for problem P; stop.
3. Find $\mathbf{x}^{k+1} := \mathbf{x}^k + \alpha^k \mathbf{d}^k$, where α^k improves $\ell_1(\mathbf{x}^k + \alpha \mathbf{d}^k)$ over $\{\alpha \in \mathbb{R} : \alpha > 0\}$ [or any other suitable merit functions]. Update B^k to a positive definite matrix B^{k+1} [e.g., according to the quasi-Newton update scheme (1.31)]. Replace $k \leftarrow k + 1$, and go to step 1.

A number of free and commercial interior-point solvers is given in Tab 1.2. below.

Note first that `fmincon` is the function implementing SQP in the version 3.0.2 of MATLAB®'s Optimization Toolbox. The other solvers listed in Tab. 1.2. are stand-alone (either in C/C++ or fortran77 programming language). However, NLPQL, SNOPT and filterSQP can be used in MATLAB® through the TOMLAB optimization environment

Table 1.2. A number of open-source and commercial codes implementing SQP techniques.

Solver	Website	Licensing	Characteristics
fmincon	`http://www.mathworks.com/access/helpdesk/help/toolbox/optim/`	commercial	line search, active set, dense problems
NLPQL	`http://www.uni-bayreuth.de/departments/math/~kschittkowski/nlpqlp22.htm`	commercial	line search, active set, dense problems
RFSQP	`http://www.aemdesign.com/RFSQPwhatis.htm`	free for acad.	line search, active set, feasible SQP, dense problem
SNOPT	`http://www.sbsi-sol-optimize.com/asp/sol_products_snopt_desc.htm`	commercial	line search, active set, reduced Hessian, sparse/large-scale problems
filterSQP	`http://www-unix.mcs.anl.gov/~leyffer/solvers.html`	commercial	trust region, exact Hessian, dense/sparse problems

(`http://tomopt.com/tomlab/`); RFSQP, SNOPT and filterSQP can be used through the AMPL modeling language (`http://www.ampl.com/`); and, finally, SNOPT can be used through the modeling language GAMS (`http://www.gams.com/`).

Example 1.76. Consider the problem to find a solution to the problem

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^2} \; & f(\mathbf{x}) := x_1^2 + x_2^2 + \log(x_1 x_2) \\ \text{s.t.} \; & g(\mathbf{x}) := 1 - x_1 x_2 \leq 0 \\ & 0 \leq x_1, x_2 \leq 10. \end{aligned} \tag{1.47}$$

We solved this problem using the function `fmincon` of the Optimization Toolbox in MATLAB®. The M-files are as follows:

```
clear all
x0 = [ 2; 1 ];
xL = [ 0.1; 0.1 ];
xU = [ 10; 10 ];
options = optimset('Display', 'iter', 'GradObj', 'on', ...
                   'GradConstr', 'on', 'DerivativeCheck', 'on', ...
                   'LargeScale', 'off', 'HessUpdate', 'bfgs', ...
                   'Diagnostics', 'on',  'TolX', 1e-7, ...
                   'TolFun', 1e-7, 'TolCon', 1e-7, ...
                   'MaxFunEval', 100, 'MaxIter', 100 )
[xopt, fopt, iout] = fmincon( @SQP_fun, x0, [], [], [], [], xL, xU, ...
                              @SQP_ctr, options );
```

```
%%%%%%%%%%%%%%% FUNCTION TO BE MINIMIZED %%%%%%%%%%%%%%%
% Objective: f(x,y) := x^2+y^2+log(x*y)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [f,df] = SQP_fun(x)
  f = x(1)^2+x(2)^2+log(x(1)*x(2)); % function
  if nargout > 1
```

```
    df = [ 2*x(1)+1/x(1)     % gradient
           2*x(2)+1/x(2) ];
  end
end
```

```
%%%%%%%%%%%%%%%%%%%%% CONSTRAINTS %%%%%%%%%%%%%%%%%%%%%
% inequality constraint: g(x,y) := x*y
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [g,h,dg,dh] = SQP_ctr(x)
  g = [ 1-x(1)*x(2) ]; % inequality constraints
  h = [];              % equality constraints
  if nargout > 2
    dg = [ -x(2); -x(1) ]; % gradient of inequality constraints
    dh = [];               % gradient of equality constraints
  end
end
```

The results are shown in Fig. 1.20. Notice, the rather fast convergence to the optimal solution $\mathbf{x}^\star = (1, 1)$. Note also that the SQP algorithm does not necessarily take a feasible path to reach an optimal solution.

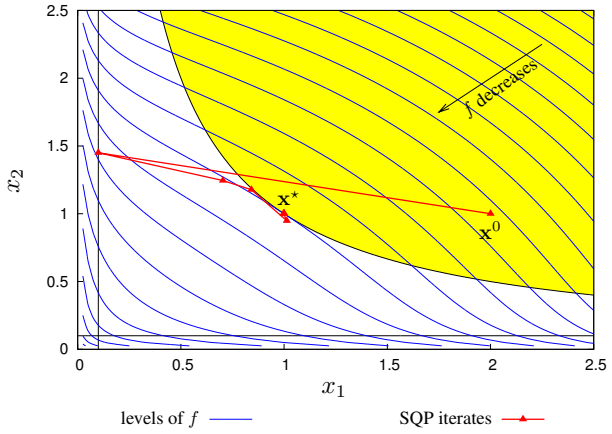


Figure 1.20. SQP iterates for Problem (1.47).

What Can Go Wrong?. The material presented up to this point was intended to give the reader an understanding of how SQP methods should work. Things do not go so smoothly in practice though. We now discuss a number of common difficulties that can be encountered, and suggest remedial actions to correct the deficiencies. Because real applications may involve more than a single difficulty, the user must be prepared to correct *all* problems before obtaining satisfactory performance from an optimization software.

Infeasible Constraints One of the most common difficulties occurs when the NLP problem has infeasible constraints, i.e., the constraints taken all together have *no* solution. Applying general-purpose SQP software to such problems typically produces one or more of the following symptoms:

- one of the QP subproblem happen to be infeasible, which occurs when the linearized constraints have no solution;
- many NLP iterations produce very little progress;
- the penalty parameters μ in (1.45) or (1.46) grows very large; or
- the Lagrange multipliers become very large.

Although robust SQP software attempt to diagnose this situation, ultimately the only remedy is to reformulate the NLP.

Rank-Deficient Constraints In contrast to the previous situation, it is possible that the constraints be consistent, but the Jacobian matrix of the active constraints, at the solution point, be either ill-conditioned or rank deficient. This situation was illustrated in Examples 1.43 and 1.55. The application of general-purpose SQP software is likely to produce the following symptoms:

- many NLP iterations produce very little progress;
- the penalty parameters μ in (1.45) or (1.46) grows very large;
- the Lagrange multipliers become very large; or
- the rank deficiency in the Jacobian of the active constraints is detected.

Note that many symptoms of rank-deficient constraints are the same as those of inconsistent constraints. It is therefore quite common to confuse this deficiency with inconsistent constraints. Again, the remedy is to reformulate the problem.

Redundant Constraints A third type of difficulty occurs when the NLP problem contains redundant constraints. Two types of redundancy may be distinguished. In the first type, some of the constraints are unnecessary to the problem formulation, which typically results in the following symptoms:

- the Lagrange multipliers are close to zero; and
- the solver has difficulty in detecting the active set.

In the second type, the redundant constraints give rise to rank deficiency, and the problem then exhibits symptoms similar to the rank-deficient case discussed previously. Obviously, the remedy is to reformulate the problem by eliminating the redundant constraints.

Discontinuities Perhaps the biggest obstacle encountered in the practical application of SQP methods (as well as many other NLP methods including SUM techniques) is the presence of discontinuous behavior. All of the numerical methods described herein assume continuous and differentiable objective function and constraints. Yet, there are many common examples of discontinuous functions in practice, including: `IF` tests in codes; absolute value, $\min$, and $\max$ functions; linear interpolation of data; internal iterations such as root finding; etc.

Regarding SQP methods, the standard QP subproblems are no longer appropriate when discontinuities are present. In fact, the KKT necessary conditions simply do not apply! The most common symptoms of discontinuous functions are:

- the iterates converge slowly or, even, diverge;
- the line search takes very small steps ($\alpha \approx 0$); and
- the Hessian matrix becomes badly ill-conditioned.

The remedy consists in reformulating discontinuous problems into smooth problems: for absolute value, $\min$, and $\max$ functions, tricks can be used that introduce slack variables and additional constraints; linear data interpolation can be replaced by higher order interpolation schemes that are continuous through second derivatives; internal iterations can also be handled via additional NLP constraints; etc.

Inaccurate Gradient Estimation Any SQP code requires that the user supply the objective function and constraint values, as well as their gradient (and possibly their Hessian too). In general, the user is proposed the option to calculate the gradients via finite differences, e.g.,

$$\boldsymbol{\nabla} f(\mathbf{x}) \;\approx\; \frac{f(\mathbf{x}+\boldsymbol{\delta}\mathbf{x}) - f(\mathbf{x})}{\boldsymbol{\delta}\mathbf{x}}.$$

However, this may cause the problem to stop prematurely. First of all, the choice of the perturbation vector $\boldsymbol{\delta}\mathbf{x}$ is highly non trivial. If too large a value clearly provides inaccurate estimates, too small a value may also result in very bad estimates due to finite arithmetic precision computations. Therefore, one must try to find a trade-off between these two extreme situations. The difficulty stems from the fact that a trade-off may not necessarily exist if the requested accuracy for the gradient is too high. In other word, the error made in the finite-difference approximation of a gradient cannot be made as small as desired. Further, the maximum accuracy that can be achieved with finite difference is both problem dependent (e.g., badly-scaled functions are more problematic than well-scaled functions) *and* machine dependent (e.g., `double precision` computations provides more accurate estimates than `single precision` computations). Typical symptoms of inaccurate gradient estimates in an SQP code are:

- the iterates converge slowly, and the solver may stop prematurely at a suboptimal point (jamming); and
- the line search takes very small steps ($\alpha \approx 0$).

The situation can be understood as follows. Assume that the gradient estimate is contaminated with noise. Then, instead of computing the true value $\nabla\mathcal{L}(\mathbf{x})$, we get $\nabla\mathcal{L}(\mathbf{x}) + \varepsilon$. But since the iteration seeks a point such that $\nabla\mathcal{L}(\mathbf{x}) = \mathbf{0}$, we can expect either a degraded rate of convergence or, worse, no convergence at all, because ultimately the gradient will be dominated by noise.

To avoid these problems, the user should always consider providing the gradients explicitely to the SQP solver, instead of relying on finite-difference estimates. For large-scale problems, this is obviously a time-consuming and error-prone task. In response to this, efficient *algorithmic differentiation* tools (also called *automatic differentiation*) have been developed within the last fifteen years. The idea behind it is that, given a piece of program calculating a number of function values (e.g., in fortran77 or C language), an auxiliary program is generated that calculates the derivatives of these functions. (See, e.g., the book by A. Griewank [24] for a general introduction on the topic.)

Scaling Scaling affects everything! Poor scaling can make a good algorithm behave badly. Scaling changes the convergence rate, termination tests, and numerical conditioning.

The most common way of scaling a problem is by introducing scaled variables of the form

$$\tilde{x}_k \;:=\; u_k x_k + r_k,$$

for $k = 1, \ldots, n_x$, with u_k and r_k being scale weights and shifts, respectively. Likewise, the objective function and constraints are commonly scaled using

$$\begin{aligned} \tilde{f} \;&:=\; \omega_0 f \\ \tilde{g}_k \;&:=\; \omega_k g_k, \end{aligned}$$

for $k = 1, \ldots, n_g$. The idea is to let the optimization algorithm work with the well-scaled quantities in order to improve performance. However, what *well-scaled quantities* mean is hard to define, although conventional wisdom suggests the following hints

- normalize the independent variables to have the same range, e.g., $0 \leq \tilde{x}_k \leq 1$;
- normalize the dependent functions to have the same magnitude, e.g., $\tilde{f} \approx \tilde{g}_1 \approx \ldots \approx \tilde{g}_{n_g} \approx 1$;
- normalize the rows and columns of the Jacobian to be of the same magnitude;
- scale the dependent functions so that the Lagrange multipliers are close to one, e.g., $|\lambda_1| \approx \ldots \approx |\lambda_{n_g}| \approx 1$; etc.

1.9 NOTES AND REFERENCES

The material on convex optimization (§1.3) is taken from the book by Boyd and Vandenberghe [10, Chapters 2 and 3]. Many more details and properties of convex programs can be found in this book, together with algorithms specifically tailored to such problems.

Regarding conditions of optimality, the material presented in §1.4 and §1.5 is mostly a summary of the material in Bazaraa, Sherali and Shetty's book [6, Chap. 4 and 5]. The derivation of necessary and sufficient conditions of optimality for equality constrained problems in §1.6 is inspired from the material in Luenberger's book [36, Chap. 10].

Additional information on the concept of algorithm (§1.8.1) and, more particularly, on the convergence aspects can be found ,e.g., in [6, Chap. 7] and [36, Chap. 6]. Regarding, unconstrained minimization techniques (§1.8.3), many additional details are given in Bertsekas' book [7, Chap. 1], as well as in [6, Chap. 8] and [36, Chap. 7 to 9]. More information on sequential unconstrained minimization algorithms (§1.8.4.1 and 1.8.4.2) can be found in [6, Chap. 9] and [36, Chap. 12]; and on sequential quadratic programming algorithms (§1.8.4.3), in [6, Chap. 10]. Many practical details on the numerical algorithms were also found in Betts' book [8, Chap. 1] and Herskovits' overview article [26].

Finally, material on Lagrangian duality theory and saddle point optimality conditions has been omitted from this chapter. The interested reader in referred to [6, Chap. 6].

Appendix: Technical Lemmas and Alternative Proofs

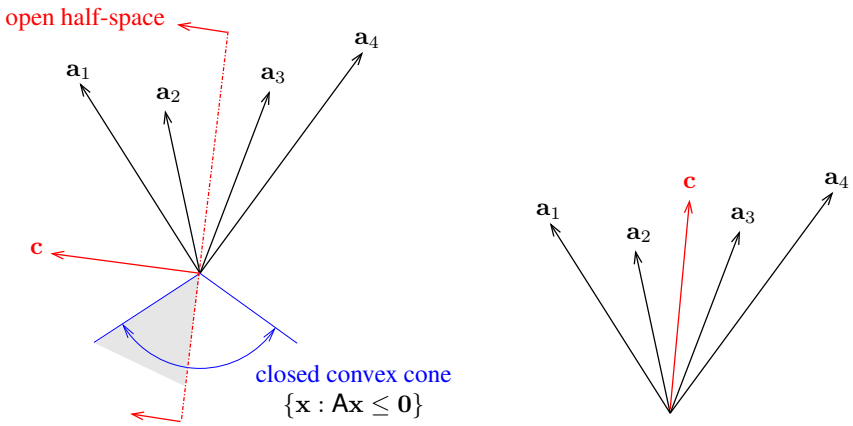
Theorem 1.A.77 (Farkas' Theorem). *Let* A *be an* $m \times n$ *matrix and* $\mathbf{c}$ *be an* n *vector. Then, exactly one of the following two statements holds:*

System 1. $\exists \mathbf{x} \in \mathbb{R}^n$ *such that* $\mathsf{A}\mathbf{x} \leq \mathbf{0}$ *and* $\mathbf{c}^\mathsf{T}\mathbf{x} > 0$,

System 2. $\exists \mathbf{y} \in \mathbb{R}^m$ *such that* $\mathsf{A}^\mathsf{T}\mathbf{y} = \mathbf{c}$ *and* $\mathbf{y} \geq \mathbf{0}$.

Proof. See, e.g., [6, Theorem 2.4.5] for a proof. □

Farkas' Theorem is used extensively in the derivation of optimality conditions of (linear and) nonlinear programming problems. A geometrical interpretation of Farkas' Theorem is shown in Fig. 1.A.1.. If $\mathbf{a}_1, \ldots, \mathbf{a}_m$ denote the rows of A, then system 2 has a solution if $\mathbf{c}$ lies in the convex cone generated by $\mathbf{a}_1, \ldots, \mathbf{a}_m$; On the other hand, system 1 has a solution if the closed convex cone $\{\mathbf{x} : \mathsf{A}\mathbf{x} \leq \mathbf{0}\}$ and the open half-space $\{\mathbf{x} : \mathbf{c}^\mathsf{T}\mathbf{x} > \mathbf{0}\}$ have a nonempty intersection.



System 1 has a solution System 2 has a solution

Figure 1.A.1. Illustration of Farkas' Theorem (with $n = 2$ and $m = 4$).

Corollary 1.A.78 (Gordan's Theorem). *Let* A *be an* $m \times n$ *matrix. Then, exactly one of the following two statements holds:*

System 1. $\exists \mathbf{x} \in \mathbb{R}^n$ *such that* $\mathsf{A}\mathbf{x} < \mathbf{0}$,

System 2. $\exists \mathbf{y} \in \mathbb{R}^m, \mathbf{y} \neq \mathbf{0}$ *such that* $\mathsf{A}^\mathsf{T}\mathbf{y} = \mathbf{0}$ *and* $\mathbf{y} \geq \mathbf{0}$.

Proof. System 1 can be equivalently written as $\mathsf{A}\mathbf{x} + \varrho\mathbf{e}$ where $\varrho > 0$ is a scalar and $\mathbf{e}$ is a vector of m ones. Rewriting this in the form of System 1 in Farkas' Theorem 1.A.77, we get $(\mathsf{A} \quad \mathbf{e})\mathbf{p}$ and $(0, \ldots, 0, 1)\mathbf{p} > 0$ where $\mathbf{p} := (\mathbf{x} \quad \varrho)$. The associated System 2 by Farkas' Theorem 1.A.77 states that $(\mathsf{A} \quad \mathbf{e})^\mathsf{T}(0, \ldots, 0, 1)^\mathsf{T}$ and $\mathbf{y} \geq \mathbf{0}$ for some $\mathbf{y} \in \mathbb{R}^m$, i.e., $\mathsf{A}^\mathsf{T}\mathbf{y} = \mathbf{0}$, $\mathbf{e}^\mathsf{T}\mathbf{y} = 1$ and $\mathbf{y} \geq \mathbf{0}$, which is equivalent to the System 2 of the corollary. □

Below is an alternative proof for Theorem 1.50 on p. 24 that does not use the concept of tangent sets.

Alternative Proof for Theorem 1.50. For $k = 1, 2, \ldots$, let $\varphi^k : \mathbb{R}^{n_x} \to \mathbb{R}$ be the (continuously differentiable) functions defined as:

$$\varphi^k(\mathbf{x}) = f(\mathbf{x}) + \frac{k}{2}\|\mathbf{h}(\mathbf{x})\|^2 + \frac{\alpha}{2}\|\mathbf{x} - \mathbf{x}^\star\|^2,$$

where $\alpha > 0$. Let also $\varepsilon > 0$ be chosen such that:

$$f(\mathbf{x}^\star) \leq f(\mathbf{x}) \quad \forall \mathbf{x} \in \bar{\mathcal{B}}_\varepsilon(\mathbf{x}^\star),$$

with $\bar{\mathcal{B}}_\varepsilon(\mathbf{x}^\star) := \{\mathbf{x} \in D : \|\mathbf{x} - \mathbf{x}^\star\| \leq \varepsilon\}$, and denote $\mathbf{x}^k \in \arg\min\{\varphi^k(\mathbf{x}) : \mathbf{x} \in \bar{\mathcal{B}}_\varepsilon(\mathbf{x}^\star)\}$. [6]

Since $\mathbf{x}^\star \in \bar{\mathcal{B}}_\varepsilon(\mathbf{x}^\star)$, we have

$$f(\mathbf{x}^k) + \frac{k}{2}\|\mathbf{h}(\mathbf{x}^k)\|^2 + \frac{\alpha}{2}\|\mathbf{x} - \mathbf{x}^\star\|^2 = \varphi^k(\mathbf{x}^k) \leq \varphi^k(\mathbf{x}^\star) = f(\mathbf{x}^\star) \quad \forall k \geq 1.$$

Hence, $\lim_{k\to\infty} \|\mathbf{h}(\mathbf{x}^k)\| = 0$, so for every limit point $\bar{\mathbf{x}} \in S$ of $\{\mathbf{x}^k\}$, we have $\mathbf{h}(\bar{\mathbf{x}}) = 0$. Moreover, $\mathbf{x}^\star$ being a local solution and $\bar{\mathbf{x}} \in \bar{\mathcal{B}}_\varepsilon(\mathbf{x}^\star)$ being a feasible point,

$$f(\mathbf{x}^\star) \leq f(\bar{\mathbf{x}}). \tag{1.A.1}$$

On the other hand, noting that $f(\mathbf{x}^k) + \frac{\alpha}{2}\|\mathbf{x} - \mathbf{x}^\star\|^2 \leq f(\mathbf{x}^\star)$ for each $k \geq 1$, and taking the limit as $k \to \infty$, we have

$$f(\bar{\mathbf{x}}) + \frac{\alpha}{2}\|\bar{\mathbf{x}} - \mathbf{x}^\star\|^2 \leq f(\mathbf{x}^\star). \tag{1.A.2}$$

Combining (1.A.1) and (1.A.2), we obtain $\|\bar{\mathbf{x}} - \mathbf{x}^\star\| = 0$, so that $\bar{\mathbf{x}} = \mathbf{x}^\star$ and $\lim_{k\to\infty} \|\bar{\mathbf{x}}\| = \mathbf{x}^\star$.

Since $\lim_{k\to\infty} \|\bar{\mathbf{x}}\| = \mathbf{x}^\star$ and $\mathbf{x}^\star \in \text{int}\left(\bar{\mathcal{B}}_\varepsilon(\mathbf{x}^\star)\right)$,

$$\exists K_1 \text{ such that } \mathbf{x}^k \in \text{int}\left(\bar{\mathcal{B}}_\varepsilon(\mathbf{x}^\star)\right) \quad \forall k > K_1,$$

i.e., $\mathbf{x}^k$ is an *unconstrained* minimum of φ^k. By Theorem 1.22 (page 11), we get

$$\mathbf{0} = \boldsymbol{\nabla}\varphi^k(\mathbf{x}^k) = \boldsymbol{\nabla} f(\mathbf{x}^k) + k\boldsymbol{\nabla}\mathbf{h}(\mathbf{x}^k)^\mathsf{T}\mathbf{h}(\mathbf{x}^k) + \alpha(\mathbf{x}^k - \mathbf{x}^\star) \quad \forall k > K_1. \tag{1.A.3}$$

$\mathbf{x}^\star$ being a regular point for the equality constraints, $\text{rank}\,(\boldsymbol{\nabla}\mathbf{h}(\mathbf{x}^\star)) = n_h$. By continuity of $\mathbf{h}$,

$$\exists K_2 \text{ such that } \text{rank}\left(\boldsymbol{\nabla}\mathbf{h}(\mathbf{x}^k)\right) = n_h \quad \forall k > K_2.$$

Therefore, $\boldsymbol{\nabla}\mathbf{h}(\mathbf{x}^k)\boldsymbol{\nabla}\mathbf{h}(\mathbf{x}^k)^\mathsf{T}$ is invertible for $k > K_2$, and we have

$$k\mathbf{h}(\mathbf{x}^k) = -\left[\boldsymbol{\nabla}\mathbf{h}(\mathbf{x}^k)\boldsymbol{\nabla}\mathbf{h}(\mathbf{x}^k)^\mathsf{T}\right]^{-1}\boldsymbol{\nabla}\mathbf{h}(\mathbf{x}^k)\left[\boldsymbol{\nabla} f(\mathbf{x}^k) + \alpha(\mathbf{x}^k - \mathbf{x}^\star)\right] \quad \forall k > K,$$

where $K = \max\{K_1, K_2\}$. Taking the limit as $k \to \infty$,

$$\lim_{k\to\infty} k\mathbf{h}(\mathbf{x}^k) = \boldsymbol{\lambda}^\star,$$

with $\boldsymbol{\lambda}^\star := -[\boldsymbol{\nabla}\mathbf{h}(\mathbf{x}^\star)\boldsymbol{\nabla}\mathbf{h}(\mathbf{x}^\star)^\mathsf{T}]^{-1}\boldsymbol{\nabla}\mathbf{h}(\mathbf{x}^\star)\boldsymbol{\nabla} f(\mathbf{x}^\star)$. Finally, taking the limit as $k \to \infty$ in (1.A.3), we obtain

$$\boldsymbol{\nabla} f(\mathbf{x}^\star) + \boldsymbol{\nabla}\mathbf{h}(\mathbf{x}^\star)^\mathsf{T}\boldsymbol{\lambda}^\star = \mathbf{0}.$$

□

Lemma 1.A.79. *Let* P *and* Q *be two symmetric matrices, such that* $\mathsf{P} \succeq \mathbf{0}$ *and* $\mathsf{P} \succ \mathbf{0}$ *on the null space of* Q *(i.e.,* $\mathbf{y}^\mathsf{T}\mathsf{P}\mathbf{y} > 0, \forall \mathbf{y} \neq \mathbf{0}$ *with* $\mathsf{Q}\mathbf{y} = 0$*). Then,*

$$\exists \bar{c} > 0 \text{ such that } \mathsf{P} + c\mathsf{Q} \succ \mathbf{0} \quad \forall c > \bar{c}.$$

[6] The minimum $\mathbf{x}^k$ exists because $\bar{\mathcal{B}}_\varepsilon(\mathbf{x}^\star)$ is nonempty, closed and bounded, and φ^k is continuous on $\bar{\mathcal{B}}_\varepsilon(\mathbf{x}^\star)$ — see Theorem 1.14 (page 7).

Proof. Assume the contrary. Then,

$$\forall k > 0, \exists \mathbf{x}^k, \|\mathbf{x}^k\| = 1 \text{ such that } {\mathbf{x}^k}^\mathsf{T} \mathsf{P} \mathbf{x}^k + k {\mathbf{x}^k}^\mathsf{T} \mathsf{Q} \mathbf{x}^k \preceq \mathbf{0}. \tag{1.A.4}$$

Consider a subsequence $\{\mathbf{x}^k\}_\mathcal{K}$ converging to some $\bar{\mathbf{x}}$ with $\|\bar{\mathbf{x}}\| = 1$. Dividing (1.A.4) by k, and taking the limit as $k \in \mathcal{K} \to \infty$, we obtain

$$\bar{\mathbf{x}}^\mathsf{T} \mathsf{Q} \bar{\mathbf{x}} \preceq \mathbf{0}.$$

On the other hand, Q being semidefinite positive, we must have

$$\bar{\mathbf{x}}^\mathsf{T} \mathsf{Q} \bar{\mathbf{x}} \succeq \mathbf{0},$$

hence $\bar{\mathbf{x}}^\mathsf{T} \mathsf{Q} \bar{\mathbf{x}} = \mathbf{0}$. That is, using the hypothesis, $\bar{\mathbf{x}}^\mathsf{T} \mathsf{P} \bar{\mathbf{x}} \succ \mathbf{0}$. This contradicts the fact that

$$\bar{\mathbf{x}}^\mathsf{T} \mathsf{P} \bar{\mathbf{x}} + \limsup_{k \to \infty, k \in \mathcal{K}} k {\mathbf{x}^k}^\mathsf{T} \mathsf{Q} \mathbf{x}^k \preceq \mathbf{0}.$$

□

CHAPTER 2

CALCULUS OF VARIATIONS

"I, Johann Bernoulli, greet the most clever mathematicians in the world. Nothing is more attractive to intelligent people than an honest, challenging problem whose possible solution will bestow fame and remain as a lasting monument. Following the example set by Pascal, Fermat, etc., I hope to earn the gratitude of the entire scientific community by placing before the finest mathematicians of our time a problem which will test their methods and the strength of their intellect. If someone communicates to me the solution of the proposed problem, I shall then publicly declare him worthy of praise."

—The Brachistochrone Challenge, Groningen, January 1, 1697.

2.1 INTRODUCTION

In an NLP problem, one seeks a real variable or real vector variable that minimizes (or maximizes) some objective function, while satisfying a set of constraints,

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in X \subset \mathbb{R}^{n_x}. \end{aligned} \tag{2.1}$$

We have seen in Chapter 1 that, when the function f and the set X have a particular functional form, a solution $\mathbf{x}^\star$ to this problem can be determined precisely. For example, if f is continuously differentiable and $X = \mathbb{R}^{n_x}$, then $\mathbf{x}^\star$ should satisfy the first-order necessary condition $\boldsymbol{\nabla} f(\mathbf{x}^\star) = \mathbf{0}$.

A multistage discrete-time generalization of (2.1) involves choosing the decision variables $\mathbf{x}_k$ in each stage $k = 1, 2, \ldots, N$, so that

$$\begin{aligned} \min_{\mathbf{x}_1,\ldots,\mathbf{x}_N} \quad & \sum_{k=1}^{N} f(k, \mathbf{x}_k) \\ \text{s.t.} \quad & \mathbf{x}_k \in X_k \subset \mathbb{R}^{n_x}, \quad k = 1, \ldots, N. \end{aligned} \tag{2.2}$$

But, since the output in a given stage affects the result in that particular stage only, (2.2) reduces to a sequence of NLP problems, namely, to select the decision variables $\mathbf{x}_k$ in each stage so as to minimize $f(k, \mathbf{x}_k)$ on X. That is, the N first-order necessary conditions satisfied by the N decision variables yield N separate conditions, each in a separate decision variable $\mathbf{x}_k$.

The problem becomes truly dynamic if the decision variables in a given stage affect not only that particular stage, but also the following stage as

$$\begin{aligned} \min_{\mathbf{x}_1,\ldots,\mathbf{x}_N} \quad & \sum_{k=1}^{N} f(k, \mathbf{x}_k, \mathbf{x}_{k-1}) \\ \text{s.t.} \quad & \mathbf{x}_0 \text{ given} \\ & \mathbf{x}_k \in X_k \subset \mathbb{R}^{n_x}, \quad k = 1, \ldots, N. \end{aligned} \tag{2.3}$$

Note that $\mathbf{x}_0$ must be specified since it affects the result in the first stage. In this case, the N first-order conditions do not separate, and must be solved simultaneously.

We now turn to continuous-time formulations. The continuous-time analog of (2.2) formulates as

$$\begin{aligned} \min_{\mathbf{x}(t)} \quad & \int_{t_1}^{t_2} f(t, \mathbf{x}(t))\, \mathrm{d}t \\ \text{s.t.} \quad & \mathbf{x}(t) \in X(t) \subset \mathbb{R}^{n_x}, \quad t_1 \le t \le t_2. \end{aligned} \tag{2.4}$$

The solution to that problem shall be a real-vector-valued function $\mathbf{x}(t)$, $t_1 \le t \le t_2$, giving the minimum value of $f(t, \mathbf{x}(t)$ at each point in time over the optimization horizon $[t_1, t_2]$. Similar to (2.2), this is not really a dynamic problem since the output at any time only affects the current function value at that time.

The continuous-time analog of (2.3) is less immediate. Time being continuous, the meaning of "previous period" relates to the idea that the objective function value depends on the decision variable $\mathbf{x}(t)$ and its rate of change $\dot{\mathbf{x}}(t)$, both at t. Thus, the problem may be written

$$\min_{\mathbf{x}(t)} \quad \int_{t_1}^{t_2} f(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\, \mathrm{d}t \tag{2.5}$$

$$\text{s.t.} \quad \mathbf{x}(t) \in X(t) \subset \mathbb{R}^{n_x}, \quad t_1 \le t \le t_2. \tag{2.6}$$

The calculus of variations refers to the latter class of problems. It is an old branch of optimization theory that has had many applications both in physics and geometry. Apart from a few examples known since ancient times such as *Queen Dido's problem* (reported in *The Aeneid* by Virgil), the problem of finding optimal curves and surfaces has been posed first by physicists such as Newton, Huygens, and Galileo. Their contemporary mathematicians, starting with the Bernoulli brothers and Leibnitz, followed by Euler and Lagrange, then invented the calculus of variations of a functional in order to solve these problems.

The calculus of variations offers many connections with the concepts introduced in Chapter 1 for nonlinear optimization problems, through the necessary and sufficient conditions of optimality and the Lagrange multipliers. The calculus of variations also constitutes an excellent introduction to the theory of optimal control, which will be presented in subsequent chapters. This is because the general form of the problem (e.g., *"Find the curve such that..."*) requires that the optimization be performed in a real function space, i.e., an infinite dimensional space, in contrast to nonlinear optimization, which is performed in a finite dimensional Euclidean space; that is, we shall consider that the variable is an element of some normed linear space of real-valued or real-vector-valued functions.

Another major difference with the nonlinear optimization problems encountered in Chapter 1 is that, in general, it is very hard to know with certainty whether a given problem of the calculus of variations has a solution (before one such solution has actually been identified). In other words, general enough, easy-to-check conditions guaranteeing the existence of a solution are lacking. Hence, the necessary conditions of optimality that we shall derive throughout this chapter are only instrumental to detect candidate solutions, and start from the implicit assumption that such solutions actually exist.

This chapter is organized as follows. We start by describing the general problem of the calculus of variations in §2.2, and describe optimality criteria in §2.3. A brief discussion of existence conditions is given in §2.4. Then, optimality conditions are presented for problems without and with constraints in §2.5 through §2.7.

Throughout this chapter, as well as the following chapter, we shall make use of basic concepts and results from functional analysis and, in particular, real function spaces. A summary is given in Appendix A.4; also refer to §A.6 for general references to textbooks on functional analysis.

2.2 PROBLEM STATEMENT

We are concerned with the problem of finding minima (or maxima) of a *functional* $\mathcal{J} : \mathcal{D} \to \mathbb{R}$, where $\mathcal{D}$ is a subset of a (normed) linear space $\mathcal{X}$ of real-valued (or real-vector-valued) functions. The formulation of a problem of the calculus of variations requires two steps: The specification of a performance criterion is discussed in §2.2.1; then, the statement of physical constraints that should be satisfied is described in §2.2.2.

2.2.1 Performance Criterion

A *performance criterion* $\mathcal{J}$, also called *cost functional* or simply *cost* must be specified for evaluating the performance of a system quantitatively. The typical form of $\mathcal{J}$ is

$$\mathcal{J}(\mathbf{x}) := \int_{t_1}^{t_2} \ell(t, \mathbf{x}(t), \dot{\mathbf{x}}(t)) \mathrm{d}t, \tag{2.7}$$

where $t \in \mathbb{R}$ is the real or independent variable, usually called *time*; $\mathbf{x}(t) \in \mathbb{R}^{n_x}$, $n_x \geq 1$, is a real vector variable, usually called the *phase* variable; the functions $\mathbf{x}(t) = (x_1, \ldots, x_{n_x})$, $t_1 \leq t \leq t_2$ are generally called *trajectories* or *curves*; $\dot{\mathbf{x}}(t) \in \mathbb{R}^{n_x}$ stands for the derivative of $\mathbf{x}(t)$ with respect to time; and $\ell : \mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}$ is a real-valued function, called a *Lagrangian* function or, briefly, a *Lagrangian*.[1] Overall, we may thus think of $\mathcal{J}(\mathbf{x})$ as dependent on an real-vector-valued continuous function $\mathbf{x}(t) \in \mathcal{X}$.

[1] The function ℓ has nothing to do with the Lagrangian $\mathcal{L}$ defined in the context of constrained optimization in Remark 1.53.

Instead of the *Lagrange problem of the calculus of variations* (2.7), we may as well consider the problem of finding a minimum (or a maximum) to the functional

$$\mathcal{J}(\mathbf{x}) := \varphi(t_1, \mathbf{x}(t_1), t_2, \mathbf{x}(t_2)) + \int_{t_1}^{t_2} \ell(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\mathrm{d}t, \tag{2.8}$$

where $\varphi : \mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R} \times \mathbb{R}^{n_x} \to \mathbb{R}$ is a real-valued function, often called a *terminal cost*. Such problems are referred to as *Bolza problems of the calculus of variations* in the literature.

Example 2.1 (Brachistochrone Problem). Consider the problem of finding the curve $x(\xi), \xi_A \leq \xi \leq \xi_B$, in the vertical plane (ξ, x), joining given points $A = (\xi_A, x_A)$ and $B = (\xi_B, x_B)$, $\xi_A < \xi_B$, $x_A < x_B$, and such that a material point sliding along $x(\xi)$ without friction from A to B, under gravity and with initial speed $v_A \geq 0$, reaches B in a minimal time (see Fig. 2.1.). This problem was first formulated then solved by Johann Bernoulli, more than 300 years ago!

The objective function $\mathcal{J}(x)$ is the time required for the point to travel from A to B along the curve $x(\xi)$,

$$\mathcal{J}(x) = \int_{\xi_A}^{\xi_B} \mathrm{d}t = \int_{\xi_A}^{\xi_B} \frac{\mathrm{d}s}{v(\xi)},$$

where s denotes the Jordan length of $x(\xi)$, defined by $\mathrm{d}s = \sqrt{1 + \dot{x}(\xi)^2}\mathrm{d}\xi$, and v, the velocity along $x(\xi)$. Since the point is sliding along $x(\xi)$ without friction, energy is conserved,

$$\frac{1}{2}m\left(v(\xi)^2 - v_A^2\right) + mg\left(x(\xi) - x_A\right) = 0,$$

with m being the mass of the point, and g, the gravity acceleration. That is, $v(\xi) = \sqrt{v_A^2 - 2g(x(\xi) - x_A)}$, and

$$\mathcal{J}(x) = \int_{\xi_A}^{\xi_B} \sqrt{\frac{1 + \dot{x}(\xi)^2}{v_A^2 - 2g(x(\xi) - x_A)}}\mathrm{d}\xi.$$

The Brachistochrone problem thus formulates as a Lagrange problem of the calculus of variations.

2.2.2 Physical Constraints

Enforcing constraints in the optimization problem reduces the set of candidate functions, i.e., not all functions in $\mathcal{X}$ are allowed. This leads to the following:

Definition 2.2 (Admissible Trajectory). *A trajectory* $\mathbf{x}$ *in a real linear function space* $\mathcal{X}$*, is said to be an* admissible trajectory *provided that it satisfies all of the physical constraints (if any) along the interval* $[t_1, t_2]$*. The set* $\mathcal{D}$ *of admissible trajectories is defined as*

$$\mathcal{D} := \{\mathbf{x} \in \mathcal{X} : \mathbf{x} \textit{ admissible}\}.$$

Typically, the functions $\mathbf{x}(t)$ are required to satisfy conditions at their end-points. Problems of the calculus of variations having end-point constraints only, are often referred to

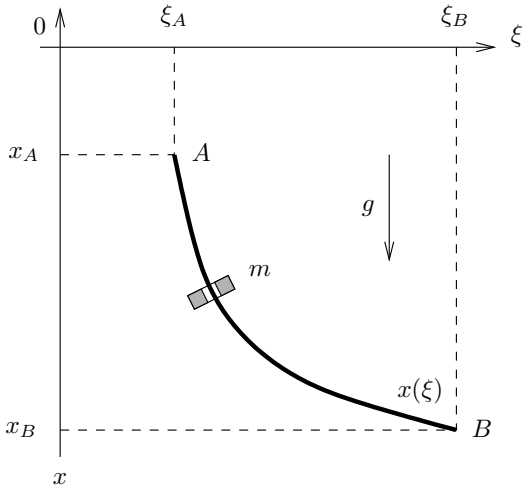


Figure 2.1. Brachistochrone problem.

as *free problems* of the calculus of variations. A great variety of boundary conditions is of interest. The simplest one is to enforce both end-points fixed, e.g., $\mathbf{x}(t_1) = \mathbf{x}_1$ and $\mathbf{x}(t_2) = \mathbf{x}_2$. Then, the set of admissible trajectories can be defined as

$$\mathcal{D} := \{\mathbf{x} \in \mathcal{X} : \mathbf{x}(t_1) = \mathbf{x}_1, \mathbf{x}(t_2) = \mathbf{x}_2\}.$$

In this case, we may say that we seek for trajectories $\mathbf{x}(t) \in \mathcal{X}$ joining the fixed points $(t_1, \mathbf{x}_1)$ and $(t_2, \mathbf{x}_2)$. Such problems will be addressed in §2.5.2.

Alternatively, we may require that the trajectory $\mathbf{x}(t) \in \mathcal{X}$ join a fixed point $(t_1, \mathbf{x}_1)$ to a specified curve $\Gamma : \mathbf{x} = \mathbf{g}(t), t_1 \leq t \leq T$. Because the final time t_2 is now free, not only the optimal trajectory $\mathbf{x}(t)$ shall be determined, but also the optimal value of t_2. That is, the set of admissible trajectories is now defined as

$$\mathcal{D} := \{(\mathbf{x}, t_2) \in \mathcal{X} \times [t_1, T] : \mathbf{x}(t_1) = \mathbf{x}_1, \mathbf{x}(t_2) = \mathbf{g}(t_2)\}.$$

Such problems will be addressed in §2.7.3.

Besides bound constraints, another type of constraints is often required,

$$\mathcal{J}_k(\mathbf{x}) := \int_{t_1}^{t_2} \ell_k(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\mathrm{d}t = C_k \quad k = 1, \ldots, m, \tag{2.9}$$

for $m \geq 1$ functionals $\ell_1, \ldots, \ell_m$. These constraints are often referred to as *isoperimetric constraints*. Similar constraints with $\leq$ sign are sometimes called *comparison functionals*.

Finally, further restriction may be necessary in practice, such as requiring that $\mathbf{x}(t) \geq \mathbf{0}$ for all or part of the optimization horizon $[t_1, t_2]$. More generally, constraints of the form

$$\begin{aligned} \Psi\left(t, \mathbf{x}(t), \dot{\mathbf{x}}(t)\right) &\leq 0 \\ \Phi\left(t, \mathbf{x}(t), \dot{\mathbf{x}}(t)\right) &= 0, \end{aligned}$$

for t in some interval $I \subset [t_1, t_2]$ are called *constraints of the Lagrangian form*, or *path constraints*. A discussion of problems having path constraints is deferred until the following chapter on optimal control.

2.3 CLASS OF FUNCTIONS AND OPTIMALITY CRITERIA

Having defined an objective functional $\mathcal{J}(\mathbf{x})$ and the set of admissible trajectories $\mathbf{x}(t) \in \mathcal{D} \subset \mathcal{X}$, one must then decide about the class of functions with respect to which the optimization shall be performed. The traditional choice in the calculus of variations is to consider the class of continuously differentiable functions, e.g., $\mathcal{C}^1[t_1, t_2]$. Yet, as shall be seen later on, an optimal solution may not exist in this class. In response to this, a more general class of functions shall be considered, such as the class $\hat{\mathcal{C}}^1[t_1, t_2]$ of *piecewise continuously differentiable functions* (see §2.6).

At this point, we need to define what is meant by a minimum (or a maximum) of $\mathcal{J}(\mathbf{x})$ on $\mathcal{D}$. Similar to finite-dimensional optimization problems (see §1.2), we shall say that $\mathcal{J}$ assumes its minimum value at $\mathbf{x}^\star$ provided that

$$\mathcal{J}(\mathbf{x}^\star) \leq \mathcal{J}(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{D}.$$

This assignment is *global* in nature and may be made without consideration of a norm (or, more generally, a distance). Yet, the specification of a norm permits an analogous description of the *local* behavior of $\mathcal{J}$ at a point $\mathbf{x}^\star \in \mathcal{D}$. In particular, $\mathbf{x}^\star$ is said to be a *local minimum* for $\mathcal{J}(\mathbf{x})$ in $\mathcal{D}$, relative to the norm $\|\cdot\|$, if

$$\exists \delta > 0 \text{ such that } \mathcal{J}(\mathbf{x}^\star) \leq \mathcal{J}(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{B}_\delta(\mathbf{x}^\star) \cap \mathcal{D},$$

with $\mathcal{B}_\delta(\mathbf{x}^\star) := \{\mathbf{x} \in \mathcal{X} : \|\mathbf{x} - \mathbf{x}^\star\| < \delta\}$. Unlike finite-dimensional linear spaces, different norms are **not** necessarily equivalent in infinite-dimensional linear spaces, in the sense that $\mathbf{x}^\star$ may be a local minimum with respect to one norm but *not* with respect to another. (See Example A.30 in Appendix A.4 for an illustration.)

Having chosen the class of functions of interest as $\mathcal{C}^1[t_1, t_2]$, several norms can be used. Maybe the most natural choice for a norm on $\mathcal{C}^1[t_1, t_2]$ is

$$\|x\|_{1,\infty} := \max_{a \leq t \leq b} |x(t)| + \max_{a \leq t \leq b} |\dot{x}(t)|,$$

since $\mathcal{C}^1[t_1, t_2]$ endowed with $\|\cdot\|_{1,\infty}$ is a Banach space. Another choice is to endow $\mathcal{C}^1[t_1, t_2]$ with the maximum norm of continuous functions,

$$\|x\|_\infty := \max_{a \leq t \leq b} |x(t)|.$$

(See Appendix A.4 for more information on norms and completeness in real function spaces.) The maximum norms, $\|\cdot\|_\infty$ and $\|\cdot\|_{1,\infty}$, are called the *strong* norm and the *weak* norm, respectively. Similarly, we shall endow $\mathcal{C}^1[t_1, t_2]^{n_x}$ with the strong norm $\|\cdot\|_\infty$ and the weak norm $\|\cdot\|_{1,\infty}$,

$$\|\mathbf{x}\|_\infty := \max_{a \leq t \leq b} \|\mathbf{x}(t)\|,$$
$$\|\mathbf{x}\|_{1,\infty} := \max_{a \leq t \leq b} \|\mathbf{x}(t)\| + \max_{a \leq t \leq b} \|\dot{\mathbf{x}}(t)\|,$$

where $\|\mathbf{x}(t)\|$ stands for any norm in $\mathbb{R}^{n_x}$.

The strong and weak norms lead to the following definitions for a local minimum:

Definition 2.3 (Strong Local Minimum, Weak Local Minimum). $\mathbf{x}^\star \in \mathcal{D}$ *is said to be a* strong local minimum *for* $\mathcal{J}(\mathbf{x})$ *in* $\mathcal{D}$ *if*

$$\exists \delta > 0 \textit{ such that } \mathcal{J}(\mathbf{x}^\star) \leq \mathcal{J}(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{B}^\infty_\delta(\mathbf{x}^\star) \cap \mathcal{D}.$$

Likewise, $\mathbf{x}^\star \in \mathcal{D}$ *is said to be a* weak local minimum *for* $\mathcal{J}(\mathbf{x})$ *in* $\mathcal{D}$ *if*

$$\exists \delta > 0 \text{ such that } \mathcal{J}(\mathbf{x}^\star) \leq \mathcal{J}(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{B}_\delta^{1,\infty}(\mathbf{x}^\star) \cap \mathcal{D}.$$

Note that for strong (local) minima, we completely disregard the values of the derivatives of the comparison elements $\mathbf{x} \in \mathcal{D}$. That is, a neighborhood associated to the topology induced by $\|\cdot\|_\infty$ has many more curves than in the topology induced by $\|\cdot\|_{1,\infty}$. In other words, **a weak minimum may not necessarily be a strong minimum**. These important considerations are illustrated in the following:

Example 2.4 (A Weak Minimum that is Not a Strong One). Consider the problem P to minimize the functional

$$\mathcal{J}(x) := \int_0^1 [\dot{x}(t)^2 - \dot{x}(t)^4] \mathrm{d}t,$$

on $\mathcal{D} := \{x \in \hat{\mathcal{C}}^1[0,1] : x(0) = x(1) = 0\}$.

We first show that the function $\bar{x}(t) = 0$, $0 \leq t \leq 1$, is a weak local minimum for P. In the topology induced by $\|\cdot\|_{1,\infty}$, consider the open ball of radius 1 centered at $\bar{x}$, i.e., $\mathcal{B}_1^{1,\infty}(\bar{x})$. For every $x \in \mathcal{B}_1^{1,\infty}(\bar{x})$, we have

$$\dot{x}(t) \leq 1, \quad \forall t \in [0,1],$$

hence $\mathcal{J}(x) \geq 0$. This proves that $\bar{x}$ is a local minimum for P since $\mathcal{J}(\bar{x}) = 0$.

In the topology induced by $\|\cdot\|_\infty$, on the other hand, the admissible trajectories $x \in \mathcal{B}_\delta^\infty(\bar{x})$ are allowed to take arbitrarily large values $\dot{x}(t)$, $0 \leq t \leq 1$. Consider the sequence of functions defined by

$$x_k(t) := \begin{cases} \frac{1}{k} + 2t - 1 & \text{if } \frac{1}{2} - \frac{1}{2k} \leq t \leq \frac{1}{2} \\ \frac{1}{k} - 2t + 1 & \text{if } \frac{1}{2} \leq t \leq \frac{1}{2} + \frac{1}{2k} \\ 0 & \text{otherwise.} \end{cases}$$

and illustrated in Fig. 2.2. below. Clearly, $x_k \in \hat{\mathcal{C}}^1[0,1]$ and $x_k(0) = x_k(1) = 0$ for each $k \geq 1$, i.e., $x_k \in \mathcal{D}$. Moreover,

$$\|x_k\| = \max_{0 \leq t \leq 1} |x_k(t)| = \frac{1}{k}.$$

meaning that for every $\delta > 0$, there is a $k \geq 1$ such that $x_k \in \mathcal{B}_\delta^\infty(\bar{x})$. (E.g., by taking $k = \mathsf{E}(\frac{2}{\delta})$.) Finally,

$$\mathcal{J}(x_k) = \int_0^1 [\dot{x}_k(t)^2 - \dot{x}_k(t)^4] \, \mathrm{d}t = [-12t]_{\frac{1}{2} - \frac{1}{2k}}^{\frac{1}{2} + \frac{1}{2k}} = -\frac{12}{k} < 0,$$

for each $k \geq 1$. Therefore, the trajectory $\bar{x}$ cannot be a strong local minimum for P (see Definition 2.3).

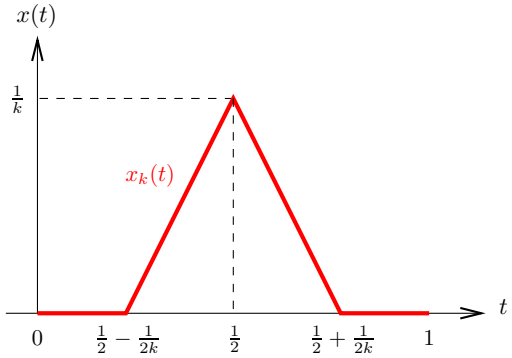


Figure 2.2. Perturbed trajectories around $\bar{x} = 0$ in Example 2.4.

2.4 EXISTENCE OF AN OPTIMAL SOLUTION

Prior to deriving conditions that must be satisfied for a function to be a minimum (or a maximum) of a problem of the calculus of variations, one must ask the question whether such solutions actually exist for that problem.

In the case of optimization in finite-dimensional Euclidean spaces, it has been shown that a continuous function on a nonempty compact set assumes its minimum (and its maximum) value in that set (see Theorem 1.14, p. 7). Theorem A.44 in Appendix A.4 extends this result to continuous functionals on a nonempty compact subset of a normed linear space. But as attractive as this solution to the problem of establishing the existence of maxima and minima may appear, it is of little help because most of the sets of interest are too "large" to be compact.

The principal reason is that most of the sets of interest are **not** bounded with respect to the norms of interest. As just an example, the set $\mathcal{D} := \{x \in \mathcal{C}^1[a,b] : x(a) = x_a, x(b) = x_b\}$ is clearly not compact with respect to the strong norm $\|\cdot\|_\infty$ as well as the weak norm $\|\cdot\|_{1,\infty}$, for we can construct a sequence of curves in $\mathcal{D}$ (e.g., parabolic functions satisfying the boundary conditions) which have maximum values as large as desired. That is, the problem of minimizing, say, $\mathcal{J}(x) = -\|x\|$ or $\mathcal{J}(x) = \int_a^b x(t)\,\mathrm{d}t$, on $\mathcal{D}$, does not have a solution.

A problem of the calculus of variations that does not have a minimum is addressed in the following:

Example 2.5 (A Problem with No Minimum). Consider the problem P to minimize

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

on $\mathcal{D} := \{x \in \mathcal{C}^1[0,1] : x(0) = 0, x(1) = 1\}$.

Observe first that for any admissible trajectory $x(t)$ joining the two end-points, we have

$$\mathcal{J}(x) = \int_0^1 \sqrt{x^2 + \dot{x}^2}\,\mathrm{d}t > \int_0^1 |\dot{x}|\,\mathrm{d}t \geq \int_0^1 \dot{x}\,\mathrm{d}t = x(1) - x(0) = 1. \tag{2.10}$$

That is,

$$\mathcal{J}(x) > 1, \quad \forall x \in \mathcal{D}, \qquad \text{and} \qquad \inf\{\mathcal{J}(x) : x \in \mathcal{D}\} \geq 1.$$

Now, consider the sequence of functions $\{x_k\}$ in $\mathcal{C}^1[0,1]$ defined by $x_k(t) := t^k$. Then,

$$\begin{aligned}\mathcal{J}(x_k) &= \int_0^1 \sqrt{t^{2k} + k^2 t^{2k-2}}\mathrm{d}t = \int_0^1 t^{k-1}\sqrt{t^2 + k^2}\mathrm{d}t \\ &\leq \int_0^1 t^{k-1}\,(t+k)\,\mathrm{d}t = 1 + \frac{1}{k+1},\end{aligned}$$

so we have $\mathcal{J}(x_k) \xrightarrow{k\to\infty} 1$. Overall, we have thus shown that $\inf \mathcal{J}(x) = 1$. But since $\mathcal{J}(x) > 1$ for any x in $\mathcal{D}$, we know with certainty that $\mathcal{J}$ has no global minimizer on $\mathcal{D}$.

General conditions can be obtained under which a minimum is guaranteed to exist, e.g., by restricting the class of functions of interest. Yet, these conditions are very restrictive and, hence, often useless in practice. Therefore, we shall proceed with the theoretically unattractive task of seeking maxima and minima of functions which need **not** have them, as the above example shows.

2.5 FREE PROBLEMS OF THE CALCULUS OF VARIATIONS

2.5.1 Geometric Optimality Conditions

The concept of variation of a functional is central to solution of problems of the calculus of variations. It is defined as follows:

Definition 2.6 (Variation of a Functional, Gâteaux Derivative). *Let $\mathcal{J}$ be a functional defined on a linear space $\mathcal{X}$. Then, the* first variation of $\mathcal{J}$ at $\mathbf{x} \in \mathcal{X}$ in the direction $\boldsymbol{\xi} \in \mathcal{X}$, *also called* Gâteaux derivative with respect to $\boldsymbol{\xi}$ at $\mathbf{x}$, *is defined as*

$$\delta\mathcal{J}(\mathbf{x};\boldsymbol{\xi}) := \lim_{\eta\to 0} \frac{\mathcal{J}(\mathbf{x}+\eta\boldsymbol{\xi}) - \mathcal{J}(\mathbf{x})}{\eta} = \frac{\partial}{\partial\eta}\mathcal{J}(\mathbf{x}+\eta\boldsymbol{\xi})\bigg|_{\eta=0}$$

(provided it exists). If the limit exists for all $\boldsymbol{\xi} \in \mathcal{X}$, then $\mathcal{J}$ is said to be Gâteaux differentiable *at* $\mathbf{x}$.

Note that the Gâteaux derivative $\delta\mathcal{J}(\mathbf{x};\boldsymbol{\xi})$ need not exist in any direction $\boldsymbol{\xi} \neq \mathbf{0}$, or it may exist is some directions and not in others. Its existence presupposes that:

(i) $\mathcal{J}(\mathbf{x})$ is defined;

(ii) $\mathcal{J}(\mathbf{x}+\eta\boldsymbol{\xi})$ is defined for all sufficiently small η.

Then,

$$\delta\mathcal{J}(\mathbf{x};\boldsymbol{\xi}) = \frac{\partial}{\partial\eta}\mathcal{J}(\mathbf{x}+\eta\boldsymbol{\xi})\bigg|_{\eta=0},$$

only if this "ordinary" derivative with respect to the real variable η exists at $\eta = 0$. Obviously, if the Gâteaux derivative exists, then it is unique.

Example 2.7 (Calculation of a Gâteaux Derivative). Consider the functional $\mathcal{J}(x) := \int_a^b [x(t)]^2\,\mathrm{d}t$. For each $x \in \mathcal{C}^1[a,b]$, the integrand $[x(t)]^2$ is continuous and the integral is

finite, hence $\mathcal{J}$ is defined on all $\mathcal{C}^1[a,b]$, with $a < b$. For an arbitrary direction $\xi \in \mathcal{C}^1[a,b]$, we have

$$\begin{aligned}\frac{\mathcal{J}(x+\eta\xi) - \mathcal{J}(x)}{\eta} &= \frac{1}{\eta}\int_a^b [x(t)+\eta\xi(t)]^2 - [x(t)]^2 \,\mathrm{d}t \\ &= 2\int_a^b x(t)\,\xi(t)\,\mathrm{d}t + \eta\int_a^b [\xi(t)]^2\,\mathrm{d}t.\end{aligned}$$

Letting $\eta \to 0$ and from Definition 2.6, we get

$$\delta\mathcal{J}(x;\xi) \;=\; 2\int_a^b x(t)\,\xi(t)\,\mathrm{d}t,$$

for each $\xi \in \mathcal{C}^1[a,b]$. Hence, $\mathcal{J}$ is Gâteaux differentiable at each $x \in \mathcal{C}^1[a,b]$.

Example 2.8 (Non-Existence of a Gâteaux Derivative). Consider the functional $\mathcal{J}(x) := \int_0^1 |x(t)|\mathrm{d}t$. Clearly, $\mathcal{J}$ is defined on all $\mathcal{C}^1[0,1]$, for each continuous function $x \in \mathcal{C}^1[0,1]$ results in a continuous integrand $|x(t)|$, whose integral is finite. For $x_0(t) := 0$ and $\xi_0(t) := t$, we have,

$$\mathcal{J}(x_0+\eta\xi_0) \;=\; \int_0^1 |\eta t|\,\mathrm{d}t.$$

Therefore,

$$\frac{\mathcal{J}(x_0+\eta\xi_0) - \mathcal{J}(x_0)}{\eta} \;=\; \begin{cases} \frac{1}{2} & \text{if } \eta > 0 \\ -\frac{1}{2} & \text{if } \eta < 0, \end{cases}$$

and a Gâteaux derivative does not exist at x_0 in the direction ξ_0.

Observe that $\delta\mathcal{J}(\mathbf{x};\boldsymbol{\xi})$ depends only on the local behavior of $\mathcal{J}$ near $\mathbf{x}$. It can be understood as the generalization of the directional derivative of a real-value function in a finite-dimensional Euclidean space $\mathbb{R}^{n_x}$. As is to be expected from its definition, the Gâteaux derivative is a linear operation on the functional $\mathcal{J}$ (by the linearity of the ordinary derivative):

$$\delta(\mathcal{J}_1+\mathcal{J}_2)(\mathbf{x};\boldsymbol{\xi}) \;=\; \delta\mathcal{J}_1(\mathbf{x};\boldsymbol{\xi}) + \delta\mathcal{J}_2(\mathbf{x};\boldsymbol{\xi}),$$

for any two functionals $\mathcal{J}_1$, $\mathcal{J}_2$, whenever these variations exist. Moreover, for any real scalar α, we have

$$\delta\mathcal{J}(\mathbf{x};\alpha\boldsymbol{\xi}) \;=\; \alpha\delta\mathcal{J}(\mathbf{x};\boldsymbol{\xi}),$$

provided that any of these variations exist; in other words, $\delta\mathcal{J}(\mathbf{x};\cdot)$ is an homogeneous operator. However, since $\delta\mathcal{J}(\mathbf{x};\cdot)$ is *not* additive[2] in general, it may not define a linear operator on $\mathcal{X}$.

It should also be noted that the Gâteaux Derivative can be formed without consideration of a norm (or distance) on $\mathcal{X}$. That is, when non-vanishing, it precludes local minimum (and maximum) behavior with respect to *any* norm:

[2]Indeed, the sum of two Gâteaux derivatives $\delta\mathcal{J}(\mathbf{x};\boldsymbol{\xi}_1) + \delta\mathcal{J}(\mathbf{x};\boldsymbol{\xi}_2)$ need not be equal to the Gâteaux derivative $\delta\mathcal{J}(\mathbf{x};\boldsymbol{\xi}_1+\boldsymbol{\xi}_2)$

Lemma 2.9. *Let $\mathcal{J}$ be a functional defined on a normed linear space $(\mathcal{X}, \|\cdot\|)$. Suppose that $\mathcal{J}$ has a strictly negative variation $\delta\mathcal{J}(\bar{\mathbf{x}};\boldsymbol{\xi}) < 0$ at a point $\bar{\mathbf{x}} \in \mathcal{X}$ in some direction $\boldsymbol{\xi} \in \mathcal{X}$. Then, $\bar{\mathbf{x}}$ cannot be a local minimum point for $\mathcal{J}$ (in the sense of the norm $\|\cdot\|$).*

Proof. Since $\delta\mathcal{J}(\bar{\mathbf{x}};\boldsymbol{\xi}) < 0$, from Definition 2.6,

$$\exists \delta > 0 \text{ such that } \frac{\mathcal{J}(\bar{\mathbf{x}} + \eta\boldsymbol{\xi}) - \mathcal{J}(\bar{\mathbf{x}})}{\eta} < 0, \quad \forall \eta \in \mathcal{B}_\delta(0).$$

Hence,

$$\mathcal{J}(\bar{\mathbf{x}} + \eta\boldsymbol{\xi}) < \mathcal{J}(\bar{\mathbf{x}}), \quad \forall \eta \in (0, \delta).$$

Since $\|(\bar{\mathbf{x}} + \eta\boldsymbol{\xi}) - \bar{\mathbf{x}}\| = \eta\|\boldsymbol{\xi}\| \to 0$ as $\eta \to 0^+$, the points $\bar{\mathbf{x}} + \eta\boldsymbol{\xi}$ are eventually in each neighborhood of $\bar{\mathbf{x}}$, irrespective of the norm $\|\cdot\|$ considered on $\mathcal{X}$. Thus, local minimum behavior of $\mathcal{J}$ is not possible in the direction $\boldsymbol{\xi}$ at $\bar{\mathbf{x}}$ (see Definition 2.3, p. 66). □

Remark 2.10. A direction $\boldsymbol{\xi} \in \mathcal{X}$ such that $\delta\mathcal{J}(\bar{\mathbf{x}};\boldsymbol{\xi}) < 0$ defines a *descent direction* for $\mathcal{J}$ at $\bar{\mathbf{x}}$. That is, the condition $\delta\mathcal{J}(\bar{\mathbf{x}};\boldsymbol{\xi}) < 0$ can be seen as a generalization of the algebraic condition $\boldsymbol{\nabla} f(\bar{\mathbf{x}})^{\mathsf{T}}\boldsymbol{\xi}$, for $\boldsymbol{\xi}$ to be a descent direction at $\bar{\mathbf{x}}$ of a real-valued function $f : \mathbb{R}^{n_x} \to \mathbb{R}$ (see Lemma 1.21, p. 10). By analogy, we shall define the following set:

$$\mathcal{F}_0(\bar{\mathbf{x}}) := \{\boldsymbol{\xi} \in \mathcal{X} : \delta\mathcal{J}(\bar{\mathbf{x}};\boldsymbol{\xi}) < 0\}.$$

Then, by Lemma 2.9, we have that $\mathcal{F}_0(\bar{\mathbf{x}}) = \emptyset$, if $\bar{\mathbf{x}}$ is a local minimizer of $\mathcal{J}$ on $\mathcal{X}$.

Example 2.11 (Non-Minimum Point). Consider the problem to minimize the functional $\mathcal{J}(x) := \int_a^b [x(t)]^2\,\mathrm{d}t$ for $x \in \mathcal{C}^1[a,b]$, $a < b$. It was shown in Example 2.7, that $\mathcal{J}$ is defined on all $\mathcal{C}^1[a,b]$ and is Gâteaux differentiable at each $x \in \mathcal{C}^1[a,b]$, with

$$\delta\mathcal{J}(x;\xi) \;=\; 2\int_a^b x(t)\,\xi(t)\,\mathrm{d}t,$$

for each $\xi \in \mathcal{C}^1[a,b]$.

Now, let $x_0(t) := t^2$ and $\xi_0(t) := -\exp(t)$. Clearly, $\delta\mathcal{J}(x_0;\xi_0)$ is non-vanishing and negative in the direction ξ_0. Thus, ξ_0 defines a descent direction for $\mathcal{J}$ at x_0, i.e., x_0 cannot be a local minimum point for $\mathcal{J}$ on $\mathcal{C}^1[a,b]$ (endowed with any norm).

We now turn to the problem of minimizing a functional on a *subset* of a linear normed linear space.

Definition 2.12 ($\mathcal{D}$-Admissible Directions). *Let $\mathcal{J}$ be a functional defined on a subset $\mathcal{D}$ of a linear space $\mathcal{X}$, and let $\bar{\mathbf{x}} \in \mathcal{D}$. Then, a direction $\boldsymbol{\xi} \in \mathcal{X}$, $\boldsymbol{\xi} \neq \mathbf{0}$, is said to be* $\mathcal{D}$-admissible at $\bar{\mathbf{x}}$ for $\mathcal{J}$, *if*

(i) $\delta\mathcal{J}(\bar{\mathbf{x}};\boldsymbol{\xi})$ exists; and

(ii) $\bar{\mathbf{x}} + \eta\boldsymbol{\xi} \in \mathcal{D}$ for all sufficiently small η, i.e.,

$$\exists \delta > 0 \text{ such that } \bar{\mathbf{x}} + \eta\boldsymbol{\xi} \in \mathcal{D}, \quad \forall \eta \in \mathcal{B}_\delta(0).$$

Observe, in particular, that if $\boldsymbol{\xi}$ is admissible at $\bar{\mathbf{x}}$, then so is each direction $c\boldsymbol{\xi}$, for $c \in \mathbb{R}$ (not necessarily a positive scalar).

For NLP problems of the form $\min\{f(\mathbf{x}) : \mathbf{x} \in S \subset \mathbb{R}^{n_x}\}$, we saw in Chapter 1 that a (local) *geometric optimality condition* is $\mathcal{F}_0(\mathbf{x}^\star) \cap \mathcal{D}(\mathbf{x}^\star) = \emptyset$ (see Theorem 1.31, p. 16). The following theorem extends this characterization to optimization problems in normed linear spaces, based on the concept of $\mathcal{D}$-admissible directions.

Theorem 2.13 (Geometric Necessary Conditions for a Local Minimum). *Let $\mathcal{J}$ be a functional defined on a subset $\mathcal{D}$ of a normed linear space $(\mathcal{X}, \|\cdot\|)$. Suppose that $\mathbf{x}^\star \in \mathcal{D}$ is a local minimum point for $\mathcal{J}$ on $\mathcal{D}$. Then*

$$\delta\mathcal{J}(\mathbf{x}^\star; \boldsymbol{\xi}) = 0, \quad \textit{for each } \mathcal{D}\textit{-admissible direction } \boldsymbol{\xi} \textit{ at } \mathbf{x}^\star. \tag{2.11}$$

Proof. By contradiction, suppose that there exists a $\mathcal{D}$-admissible direction $\boldsymbol{\xi}$ such that $\delta\mathcal{J}(\mathbf{x}^\star, \boldsymbol{\xi}) < 0$. Then, by Lemma 2.9, $\mathbf{x}^\star$ cannot be a local minimum for $\mathcal{J}$. Likewise, there cannot be a $\mathcal{D}$-admissible direction $\boldsymbol{\xi}$ such that $\delta\mathcal{J}(\mathbf{x}^\star, \boldsymbol{\xi}) > 0$. Indeed, $-\boldsymbol{\xi}$ being $\mathcal{D}$-admissible and $\delta\mathcal{J}(\mathbf{x}^\star, -\boldsymbol{\xi}) = -\delta\mathcal{J}(\mathbf{x}^\star, \boldsymbol{\xi}) < 0$, by Lemma 2.9, $\mathbf{x}^\star$ cannot be a local minimum otherwise. Overall, we must therefore have that $\delta\mathcal{J}(\mathbf{x}^\star, \boldsymbol{\xi}) = 0$ for each $\mathcal{D}$-admissible direction $\boldsymbol{\xi}$ at $\mathbf{x}^\star$. □

Our hope is that there will be "enough" admissible directions so that the condition $\delta\mathcal{J}(\mathbf{x}^\star; \boldsymbol{\xi}) = 0$ can determine $\mathbf{x}^\star$. There may be "too many" nontrivial admissible directions to allow any $\mathbf{x} \in \mathcal{D}$ to fulfill this condition; there may as well be just one such direction, or even no admissible direction at all (see Example 2.41 on p. 90 for an illustration of this latter possibility).

Observe also that the condition $\delta\mathcal{J}(\mathbf{x}^\star; \boldsymbol{\xi}) = 0$ alone cannot distinguish between a local minimum and a local maximum point, nor can it distinguish between a local minimum and a global minimum point. As in finite-dimensional optimization, we must also admit the possibility of *stationary points* (such as saddle points), which satisfy this condition but are neither local maximum points nor local minimum points (see Remark 1.23, p. 11). Further, the Gâteaux derivative is a *weak variation* in the sense that minima and maxima obtained via Theorem 2.13 are weak minima/maxima; that is, it does not distinguish between weak and strong minima/maxima.

Despite these limitations, the condition $\delta\mathcal{J}(\mathbf{x}^\star; \boldsymbol{\xi}) = 0$ provides the most obvious approach to attacking problems of the calculus of variations (and also in optimal control problems). We apply it to solve elementary problems of the calculus of variations in subsections §2.5.2. Then, Legendre second-order necessary condition is derived in §2.5.3, and a simple first-order sufficient condition is given in §2.5.4. Finally, necessary conditions for problems with free end-points are developed in §2.5.5.

2.5.2 Euler's Necessary Condition

In this section, we present a first-order necessary condition that must be satisfied for an arc $\mathbf{x}(t)$, $t_1 \le t \le t_2$, to be a (weak) local minimum of a functional in the Lagrange form on $(\mathcal{C}^1[t_1, t_2])^{n_x}$, subject to the bound constraints $\mathbf{x}(t_1) = \mathbf{x}_1$ and $\mathbf{x}(t_2) = \mathbf{x}_2$. This problem is know as the *elementary problem of the calculus of variations*.

Theorem 2.14 (Euler's Necessary Conditions). *Consider the problem to minimize the functional*

$$\mathcal{J}(\mathbf{x}) := \int_{t_1}^{t_2} \ell(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\, \mathrm{d}t,$$

on $\mathcal{D} := \{\mathbf{x} \in \mathcal{C}^1[t_1, t_2]^{n_x} : \mathbf{x}(t_1) = \mathbf{x}_1, \mathbf{x}(t_2) = \mathbf{x}_2\}$, *with* $\ell : \mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}$ *a continuously differentiable function. Suppose that* $\mathbf{x}^\star$ *gives a (local) minimum for* $\mathcal{J}$ *on* $\mathcal{D}$. *Then,*

$$\frac{\mathrm{d}}{\mathrm{d}t}\ell_{\dot{x}_i}(t, \mathbf{x}^\star(t), \dot{\mathbf{x}}^\star(t)) = \ell_{x_i}(t, \mathbf{x}^\star(t), \dot{\mathbf{x}}^\star(t)), \tag{2.12}$$

for each $t \in [t_1, t_2]$, *and each* $i = 1, \ldots, n_x$.

Proof. Based on the differentiability properties of ℓ, and by Theorem 2.A.59, we have

$$\begin{aligned}\frac{\partial}{\partial\eta}\mathcal{J}(\mathbf{x}^\star + \eta\boldsymbol{\xi}) &= \int_{t_1}^{t_2} \frac{\partial}{\partial\eta}\ell\left(t, \mathbf{x}^\star(t) + \eta\boldsymbol{\xi}(t), \dot{\mathbf{x}}^\star(t) + \eta\dot{\boldsymbol{\xi}}(t)\right)\mathrm{d}t \\ &= \int_{t_1}^{t_2} \left(\ell_{\mathbf{x}}\left[\mathbf{x}^\star + \eta\boldsymbol{\xi}\right]^{\mathsf{T}}\boldsymbol{\xi} + \ell_{\dot{\mathbf{x}}}\left[\mathbf{x}^\star + \eta\boldsymbol{\xi}\right]^{\mathsf{T}}\dot{\boldsymbol{\xi}}\right)\mathrm{d}t,\end{aligned}$$

for each $\boldsymbol{\xi} \in \mathcal{C}^1[t_1, t_2]^{n_x}$, where the compressed notation $\ell_z[\mathbf{y}] := \ell_z(t, \mathbf{y}(t), \dot{\mathbf{y}}(t))$ is used. Taking the limit as $\eta \to 0$, we get

$$\delta\mathcal{J}(\mathbf{x}^\star; \boldsymbol{\xi}) = \int_{t_1}^{t_2} \left(\ell_{\mathbf{x}}\left[\mathbf{x}^\star\right]^{\mathsf{T}}\boldsymbol{\xi} + \ell_{\dot{\mathbf{x}}}\left[\mathbf{x}^\star\right]^{\mathsf{T}}\dot{\boldsymbol{\xi}}\right)\mathrm{d}t,$$

which is finite for each $\boldsymbol{\xi} \in \mathcal{C}^1[t_1, t_2]^{n_x}$, since the integrand is continuous on $[t_1, t_2]$. Therefore, the functional $\mathcal{J}$ is Gâteaux differentiable at each $\mathbf{x} \in \mathcal{C}^1[t_1, t_2]^{n_x}$.

Now, for fixed $i = 1, \ldots, n_x$, choose $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{n_x})^{\mathsf{T}} \in \mathcal{C}^1[t_1, t_2]^{n_x}$ such that $\xi_j = 0$ for all $j \neq i$, and $\xi_i(t_1) = \xi_i(t_2) = 0$. Clearly, $\boldsymbol{\xi}$ is $\mathcal{D}$-admissible, since $\mathbf{x}^\star + \eta\boldsymbol{\xi} \in \mathcal{D}$ for each $\eta \in \mathbb{R}$ and the Gâteaux derivative $\delta\mathcal{J}(\mathbf{x}^\star; \boldsymbol{\xi})$ exists. Then, $\mathbf{x}^\star$ being a local minimizer for $\mathcal{J}$ on $\mathcal{D}$, by Theorem 2.13, we get

$$0 = \int_{t_1}^{t_2} \left(\ell_{x_i}\left[\mathbf{x}^\star\right]\xi_i + \ell_{\dot{x}_i}\left[\mathbf{x}^\star\right]\dot{\xi}_i\right)\mathrm{d}t \tag{2.13}$$

$$= \int_{t_1}^{t_2} \ell_{\dot{x}_i}\left[\mathbf{x}^\star\right]\dot{\xi}_i\,\mathrm{d}t + \int_{t_1}^{t_2} \frac{\mathrm{d}}{\mathrm{d}t}\left[\int_{t_1}^{t} \ell_{x_i}\left[\mathbf{x}^\star\right]\mathrm{d}\tau\right]\xi_i\,\mathrm{d}t \tag{2.14}$$

$$= \int_{t_1}^{t_2} \ell_{\dot{x}_i}\left[\mathbf{x}^\star\right]\dot{\xi}_i\,\mathrm{d}t + \left[\xi_i\int_{t_1}^{t} \ell_{x_i}\left[\mathbf{x}^\star\right]\mathrm{d}\tau\right]_{t_1}^{t_2} - \int_{t_1}^{t_2}\left[\int_{t_1}^{t} \ell_{x_i}\left[\mathbf{x}^\star\right]\mathrm{d}\tau\right]\dot{\xi}_i\,\mathrm{d}t \tag{2.15}$$

$$= \int_{t_1}^{t_2}\left[\ell_{\dot{x}_i}\left[\mathbf{x}^\star\right] - \int_{t_1}^{t} \ell_{x_i}\left[\mathbf{x}^\star\right]\mathrm{d}\tau\right]\dot{\xi}_i\,\mathrm{d}t, \tag{2.16}$$

and by Lemma 2.A.57,

$$\ell_{\dot{x}_i}\left[\mathbf{x}^\star\right] - \int_{t_1}^{t} \ell_{x_i}\left[\mathbf{x}^\star\right]\mathrm{d}\tau = C_i, \quad \forall t \in [t_1, t_2], \tag{2.17}$$

for some real scalar C_i. In particular, we have $\ell_{\dot{x}_i}\left[\mathbf{x}^\star\right] \in \mathcal{C}^1[t_1, t_2]$, and differentiating (2.17) with respect to t yields (2.12). □

Definition 2.15 (Stationary Function). *Each* $\mathcal{C}^1$ *function* $\bar{\mathbf{x}}$ *that satisfies the Euler equation* (2.12) *on some interval is called a* stationary function *for the Lagrangian* ℓ.

An old and rather entrenched tradition calls such functions *extremal functions* or simply *extremals*, although they may provide neither a local minimum nor a local maximum for

the problem; here, we shall employ the term *extremal* only for those functions which are actual minima or maxima to a given problem. Observe also that stationary functions are *not* required to satisfy any particular boundary conditions, even though we might be interested only in those which meet given boundary conditions in a particular problem.

Example 2.16. Consider the functional

$$\mathcal{J}(x) := \int_0^T [\dot{x}(t)]^2 \, \mathrm{d}t,$$

for $x \in \mathcal{D} := \{x \in \mathcal{C}^1[0,T] : x(0) = 0, x(T) = A\}$. Since the Lagrangian function is independent of x, the Euler equation for this problem reads

$$\ddot{x}(t) \;=\; 0.$$

Integrating this equation twice yields,

$$x(t) \;=\; c_1 t + c_2,$$

where c_1 and c_2 are constants of integration. These two constants are determined from the end-point conditions of the problem as $c_1 = \frac{A}{T}$ and $c_2 = 0$. The resulting stationary function is

$$\bar{x}(t) := A\frac{t}{T}.$$

Note that nothing can be concluded as to whether $\bar{x}$ gives a minimum, a maximum, or neither of these, for $\mathcal{J}$ on $\mathcal{D}$, based solely on Euler equation.

Remark 2.17 (First Integrals). Solutions to the Euler equation (2.12) can be obtained more easily in the following three situations:

(i) The Lagrangian function ℓ does not depend on the independent variable t: $\ell = \ell(\mathbf{x}, \dot{\mathbf{x}})$. Defining the *Hamiltonian* as

$$\mathcal{H} := \ell(\mathbf{x}, \dot{\mathbf{x}}) - \dot{\mathbf{x}}^{\mathsf{T}} \ell_{\dot{\mathbf{x}}}(\mathbf{x}, \dot{\mathbf{x}}),$$

we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H} \;=\; \ell_{\mathbf{x}}^{\mathsf{T}}\dot{\mathbf{x}} + \ell_{\dot{\mathbf{x}}}^{\mathsf{T}}\ddot{\mathbf{x}} - \ddot{\mathbf{x}}^{\mathsf{T}}\ell_{\dot{\mathbf{x}}} - \dot{\mathbf{x}}^{\mathsf{T}}\frac{\mathrm{d}}{\mathrm{d}t}\ell_{\dot{\mathbf{x}}} \;=\; \dot{\mathbf{x}}^{\mathsf{T}}\left(\ell_{\mathbf{x}} - \frac{\mathrm{d}}{\mathrm{d}t}\ell_{\dot{\mathbf{x}}}\right) \;=\; 0.$$

Therefore, $\mathcal{H}$ is constant along a stationary trajectory. Note that for physical systems, such invariants often provide energy conservation laws; for controlled systems, $\mathcal{H}$ may correspond to a cost that depends on the trajectory, but not on the position onto that trajectory.

(ii) The Lagrangian function ℓ does not depend on x_i. Then,

$$\frac{\mathrm{d}}{\mathrm{d}t}\ell_{\dot{x}_i}(t, \dot{\mathbf{x}}) \;=\; 0,$$

meaning that $\ell_{\dot{x}_i}$ is an invariant for the system. For physical systems, such invariants typically provide momentum conservation laws.

(iii) The Lagrangian function ℓ does not depend on $\dot{x}_i$. Then, the ith component of the Euler equation becomes $\ell_{x_i} = 0$ along a stationary trajectory. Although this is a degenerate case, we should try to reduce problems into this form systematically, for this is also an easy case to solve.

Example 2.18 (Minimum Path Problem). Consider the problem to minimize the distance between two fixed points, namely $A = (x_A, y_A)$ and $B = (x_B, y_B)$, in the (x, y)-plane. That is, we want to find the curve $y(x)$, $x_A \leq x \leq x_B$ such that the functional $\mathcal{J}(y) := \int_{x_A}^{x_B} \sqrt{1 + \dot{y}(x)^2}\, \mathrm{d}x$ is minimized, subject to the bound constraints $y(x_A) = y_A$ and $y(x_B) = x_B$. Since $\ell(x, y, \dot{y}) = \sqrt{1 + \dot{y}(x)^2}$ does not depend on x and y, Euler equation reduces to $\ell_{\dot{y}} = C$, with C a constant. That is, $\dot{y}(x)$ is constant along any stationary trajectory and we have,

$$y(x) = C_1 x + C_2 = \frac{y_B - y_A}{x_B - x_A} x + \frac{y_A x_B - y_B x_A}{x_B - x_A}.$$

Quite expectedly, we get that the curve minimizing the distance between two points is a straight line ☺

2.5.3 Second-Order Necessary Conditions

In optimizing a twice continuously differentiable function $f(\mathbf{x})$ in $\mathbb{R}^{n_x}$, it was shown in Chapter 1 that if $\mathbf{x}^\star \in \mathbb{R}^{n_x}$ gives a (local) minimizer of f, then $\boldsymbol{\nabla} f(\mathbf{x}^\star) = \mathbf{0}$ and $\boldsymbol{\nabla}^2 f(\mathbf{x}^\star)$ is positive semi-definite (see, e.g., Theorem 1.25, p. 12). That is, if $\mathbf{x}^\star$ provides a local minimum, then f is both stationary and locally convex at $\mathbf{x}^\star$.

Somewhat analogous conditions can be developed for free problems of the calculus of variations. Their development relies on the concept of second variation of a functional:

Definition 2.19 (second Variation of a Functional). *Let $\mathcal{J}$ be a functional defined on a linear space $\mathcal{X}$. Then, the* second variation of $\mathcal{J}$ at $\mathbf{x} \in \mathcal{X}$ in the direction $\boldsymbol{\xi} \in \mathcal{X}$ *is defined as*

$$\delta^2 \mathcal{J}(\mathbf{x}; \boldsymbol{\xi}) := \frac{\partial^2}{\partial \eta^2} \mathcal{J}(\mathbf{x} + \eta \boldsymbol{\xi}) \bigg|_{\eta = 0}$$

(provided it exists).

We are now ready to formulate the second-order necessary conditions:

Theorem 2.20 (Second-Order Necessary Conditions (Legendre)). *Consider the problem to minimize the functional*

$$\mathcal{J}(\mathbf{x}) := \int_{t_1}^{t_2} \ell(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\, \mathrm{d}t,$$

on $\mathcal{D} := \{\mathbf{x} \in \mathcal{C}^1[t_1, t_2]^{n_x} : \mathbf{x}(t_1) = \mathbf{x}_1, \mathbf{x}(t_2) = \mathbf{x}_2\}$, with $\ell : \mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}$ a twice continuously differentiable function. Suppose that $\mathbf{x}^\star$ gives a (local) minimum for $\mathcal{J}$ on $\mathcal{D}$. Then $\mathbf{x}^\star$ satisfies the Euler equation (2.12) *along with the so-called Legendre condition,*

$$\ell_{\dot{\mathbf{x}}\dot{\mathbf{x}}}(t, \mathbf{x}^\star(t), \dot{\mathbf{x}}^\star(t)) \text{ semi-definite positive},$$

for each $t \in [t_1, t_2]$.

Proof. The conclusion that $\mathbf{x}^\star$ should satisfy the Euler equation (2.12) if it is a local minimizer for $\mathcal{J}$ on $\mathcal{D}$ directly follows from Theorem 2.14.

Based on the differentiability properties of ℓ, and by repeated application of Theorem 2.A.59, we have

$$\begin{aligned}\frac{\partial^2}{\partial \eta^2}\mathcal{J}(\mathbf{x}^\star + \eta\boldsymbol{\xi}) &= \int_{t_1}^{t_2} \frac{\partial^2}{\partial \eta^2}\ell\left[\mathbf{x}^\star + \eta\boldsymbol{\xi}\right] \mathrm{d}t \\ &= \int_{t_1}^{t_2} \left(\boldsymbol{\xi}^\mathsf{T}\ell_{\mathbf{xx}}\left[\mathbf{x}^\star + \eta\boldsymbol{\xi}\right]\boldsymbol{\xi} + 2\boldsymbol{\xi}^\mathsf{T}\ell_{\mathbf{x}\dot{\mathbf{x}}}\left[\mathbf{x}^\star + \eta\boldsymbol{\xi}\right]\dot{\boldsymbol{\xi}} + \dot{\boldsymbol{\xi}}^\mathsf{T}\ell_{\dot{\mathbf{x}}\dot{\mathbf{x}}}\left[\mathbf{x}^\star + \eta\boldsymbol{\xi}\right]\dot{\boldsymbol{\xi}}\right) \mathrm{d}t\end{aligned}$$

for each $\boldsymbol{\xi} \in \mathcal{C}^1[t_1, t_2]^{n_x}$, with the usual compressed notation. Taking the limit as $\eta \to 0$, we get

$$\delta^2\mathcal{J}(\mathbf{x}^\star; \boldsymbol{\xi}) = \int_{t_1}^{t_2} \left(\boldsymbol{\xi}^\mathsf{T}\ell_{\mathbf{xx}}\left[\mathbf{x}^\star\right]\boldsymbol{\xi} + 2\boldsymbol{\xi}^\mathsf{T}\ell_{\mathbf{x}\dot{\mathbf{x}}}\left[\mathbf{x}^\star\right]\dot{\boldsymbol{\xi}} + \dot{\boldsymbol{\xi}}^\mathsf{T}\ell_{\dot{\mathbf{x}}\dot{\mathbf{x}}}\left[\mathbf{x}^\star\right]\dot{\boldsymbol{\xi}}\right) \mathrm{d}t,$$

which is finite for each $\boldsymbol{\xi} \in \mathcal{C}^1[t_1, t_2]^{n_x}$, since the integrand is continuous on $[t_1, t_2]$. Therefore, the second variation of $\mathcal{J}$ at $\mathbf{x}^\star$ exists in any direction $\boldsymbol{\xi} \in \mathcal{C}^1[t_1, t_2]^{n_x}$ (and so does the first variation $\delta\mathcal{J}(\mathbf{x}^\star; \boldsymbol{\xi})$).

Now, let $\boldsymbol{\xi} \in \mathcal{C}^1[t_1, t_2]^{n_x}$ such that $\boldsymbol{\xi}(t_1) = \boldsymbol{\xi}(t_2) = \mathbf{0}$. Clearly, $\boldsymbol{\xi}$ is $\mathcal{D}$-admissible, since $\mathbf{x}^\star + \eta\boldsymbol{\xi} \in \mathcal{D}$ for each $\eta \in \mathbb{R}$ and the Gâteaux derivative $\delta\mathcal{J}(\mathbf{x}^\star; \boldsymbol{\xi})$ exists. Consider the function $f : \mathbb{R} \to \mathbb{R}$ be defined as

$$f(\eta) := \mathcal{J}(\mathbf{x}^\star + \eta\boldsymbol{\xi}).$$

We have, $\nabla f(0) = \delta\mathcal{J}(\mathbf{x}^\star; \boldsymbol{\xi})$ and $\nabla^2 f(0) = \delta^2\mathcal{J}(\mathbf{x}^\star; \boldsymbol{\xi})$. Moreover, $\mathbf{x}^\star$ being a local minimizer of $\mathcal{J}$ on $\mathcal{D}$, we must have that $\eta^\star = 0$ is a local (unconstrained) minimizer of f. Therefore, by Theorem 1.25, $\nabla f(0) = 0$ and $\nabla^2 f(0) \geq 0$. This latter condition makes it *necessary* that $\delta^2\mathcal{J}(\mathbf{x}^\star; \boldsymbol{\xi})$ be nonnegative:

$$\begin{aligned}0 &\leq \int_{t_1}^{t_2} \left(\boldsymbol{\xi}^\mathsf{T}\ell_{\mathbf{xx}}\left[\mathbf{x}^\star\right]\boldsymbol{\xi} + 2\boldsymbol{\xi}^\mathsf{T}\ell_{\mathbf{x}\dot{\mathbf{x}}}\left[\mathbf{x}^\star\right]\dot{\boldsymbol{\xi}} + \dot{\boldsymbol{\xi}}^\mathsf{T}\ell_{\dot{\mathbf{x}}\dot{\mathbf{x}}}\left[\mathbf{x}^\star\right]\dot{\boldsymbol{\xi}}\right) \mathrm{d}t \\ &= \int_{t_1}^{t_2} \left(\boldsymbol{\xi}^\mathsf{T}\ell_{\mathbf{xx}}\left[\mathbf{x}^\star\right]\boldsymbol{\xi} - \boldsymbol{\xi}^\mathsf{T}\frac{\mathrm{d}}{\mathrm{d}t}\ell_{\mathbf{x}\dot{\mathbf{x}}}\left[\mathbf{x}^\star\right]\boldsymbol{\xi} + \dot{\boldsymbol{\xi}}^\mathsf{T}\ell_{\dot{\mathbf{x}}\dot{\mathbf{x}}}\left[\mathbf{x}^\star\right]\dot{\boldsymbol{\xi}}\right) \mathrm{d}t + \left[\boldsymbol{\xi}^\mathsf{T}\frac{\mathrm{d}}{\mathrm{d}t}\ell_{\mathbf{x}\dot{\mathbf{x}}}\left[\mathbf{x}^\star\right]\boldsymbol{\xi}\right]_{t_1}^{t_2} \\ &= \int_{t_1}^{t_2} \left[\boldsymbol{\xi}^\mathsf{T}\left(\ell_{\mathbf{xx}}\left[\mathbf{x}^\star\right] - \frac{\mathrm{d}}{\mathrm{d}t}\ell_{\mathbf{x}\dot{\mathbf{x}}}\left[\mathbf{x}^\star\right]\right)\boldsymbol{\xi} + \dot{\boldsymbol{\xi}}^\mathsf{T}\ell_{\dot{\mathbf{x}}\dot{\mathbf{x}}}\left[\mathbf{x}^\star\right]\dot{\boldsymbol{\xi}}\right) \mathrm{d}t. \qquad (2.18)\end{aligned}$$

Finally, by Lemma 2.A.58, a necessary condition for (2.18) to be nonnegative is that $\ell_{\dot{\mathbf{x}}\dot{\mathbf{x}}}\left(t, \mathbf{x}^\star(t), \dot{\mathbf{x}}^\star(t)\right)$ be nonnegative, for each $t \in [t_1, t_2]$. □

Notice the strong link regarding first- and second-order necessary conditions of optimality between unconstrained NLP problems and free problems of the calculus of variations. It is therefore tempting to try to extend second-order sufficient conditions for unconstrained NLPs to variational problems. By Theorem 1.28 (p. 12), $\mathbf{x}^\star$ is a strict local minimum of the problem to minimize a twice-differentiable function f on $\mathbb{R}^{n_x}$, provided that f is both stationary and locally strictly convex at $\mathbf{x}^\star$. Unfortunately, the requirements that (i) $\bar{x}$ be a stationary function for ℓ, and (ii) $\ell(t, \mathbf{y}, \mathbf{z})$ be strictly convex with respect to $\mathbf{z}$, are

not sufficient for $\bar{x}$ to be a (weak) local minimum of a free problem of the the calculus of variations. Additional conditions must hold, such as the absence of points *conjugate* to the point t_1 in $[t_1, t_2]$ (Jacobi's sufficient condition). However, these considerations fall out of the scope of an introductory class on the calculus of variations (see also §2.8).

Example 2.21. Consider the same functional as in Example 2.16, namely,

$$\mathcal{J}(x) := \int_0^T [\dot{x}(t)]^2 \, \mathrm{d}t,$$

for $\mathbf{x} \in \mathcal{D} := \{x \in \mathcal{C}^1[0,T] : x(0) = 0, x(T) = A\}$. It was shown that the unique stationary function relative to $\mathcal{J}$ on $\mathcal{D}$ is given by

$$\bar{x}(t) := A\frac{t}{T}.$$

Here, the Lagrangian function is $\ell(t, y, z) := z^2$, and we have that

$$\ell_{zz}(t, \bar{x}(t), \dot{\bar{x}}(t)) \;=\; 2,$$

for each $0 \leq t \leq T$. By Theorem 2.20 above, $\bar{x}$ is therefore a candidate local minimizer for $\mathcal{J}$ on $\mathcal{D}$, but it cannot be a local maximizer.

2.5.4 Sufficient Conditions: Joint Convexity

The following sufficient condition for an extremal solution to be a global minimum extends the well-known convexity condition in finite-dimensional optimization, as given by Theorem 1.18 (p. 9). It should be noted, however, that this condition is somewhat restrictive and is rarely satisfied in most practical applications.

Theorem 2.22 (Sufficient Conditions). *Consider the problem to minimize the functional*

$$\mathcal{J}(\mathbf{x}) := \int_{t_1}^{t_2} \ell(t, \mathbf{x}(t), \dot{\mathbf{x}}(t)) \, \mathrm{d}t,$$

on $\mathcal{D} := \{\mathbf{x} \in \mathcal{C}^1[t_1, t_2]^{n_x} : \mathbf{x}(t_1) = \mathbf{x}_1, \mathbf{x}(t_2) = \mathbf{x}_2\}$. *Suppose that the Lagrangian* $\ell(t, \mathbf{y}, \mathbf{z})$ *is continuously differentiable and [strictly] jointly convex in* $(\mathbf{y}, \mathbf{z})$. *If* $\mathbf{x}^\star$ *stationary function for the Lagrangian* ℓ, *then* $\mathbf{x}^\star$ *is also a [strict] global minimizer for* $\mathcal{J}$ *on* $\mathcal{D}$.

Proof. The integrand $\ell(t, \mathbf{y}, \mathbf{z})$ being continuously differentiable and jointly convex in $(\mathbf{y}, \mathbf{z})$, we have for an arbitrary $\mathcal{D}$-admissible trajectory $\mathbf{x}$:

$$\begin{aligned} \int_{t_1}^{t_2} \ell[\mathbf{x}] - \ell[\mathbf{x}^\star] \, \mathrm{d}t \;&\geq\; \int_{t_1}^{t_2} (\mathbf{x} - \mathbf{x}^\star)^\mathsf{T} \ell_{\mathbf{x}}[\mathbf{x}^\star] + (\dot{\mathbf{x}} - \dot{\mathbf{x}}^\star)^\mathsf{T} \ell_{\dot{\mathbf{x}}}[\mathbf{x}^\star] \, \mathrm{d}t \\ &\geq\; \int_{t_1}^{t_2} (\mathbf{x} - \mathbf{x}^\star)^\mathsf{T} \left(\ell_{\mathbf{x}}[\mathbf{x}^\star] - \frac{\mathrm{d}}{\mathrm{d}t} \ell_{\dot{\mathbf{x}}}[\mathbf{x}^\star] \right) \mathrm{d}t + \Big[(\mathbf{x} - \mathbf{x}^\star)^\mathsf{T} \ell_{\dot{\mathbf{x}}}[\mathbf{x}^\star] \Big]_{t_1}^{t_2}, \end{aligned} \tag{2.19}$$

with the usual compressed notation. Clearly, the first term is equal to zero, for $\mathbf{x}^\star$ is a solution to the Euler equation (2.12); and the second term is also equal to zero, since $\mathbf{x}$ is $\mathcal{D}$-admissible, i.e., $\mathbf{x}(t_1) = \mathbf{x}^\star(t_1) = \mathbf{x}_1$ and $\mathbf{x}(t_2) = \mathbf{x}^\star(t_2) = \mathbf{x}_2$. Therefore,

$$\mathcal{J}(\mathbf{x}) \;\geq\; \mathcal{J}(\mathbf{x}^\star)$$

for each admissible trajectory. (A proof that strict joint convexity of ℓ in $(\mathbf{y}, \mathbf{z})$ is sufficient for a strict global minimizer is obtained upon replacing (2.19) by a strict inequality.) □

Example 2.23. Consider the same functional as in Example 2.16, namely,

$$\mathcal{J}(x) := \int_0^T [\dot{x}(t)]^2 \, \mathrm{d}t,$$

for $x \in \mathcal{D} := \{x \in \mathcal{C}^1[0,T] : x(0) = 0, x(T) = A\}$. Since the Lagrangian $\ell(t, y, z)$ does not depend depend on y and is convex in z, by Theorem 2.22, every stationary function for the Lagrangian gives a global minimum for $\mathcal{J}$ on $\mathcal{D}$. But since $\bar{x}(t) := A\frac{t}{T}$ is the unique stationary function for ℓ, then $\bar{x}(t)$ is a global minimizer for $\mathcal{J}$ on $\mathcal{D}$.

2.5.5 Problems with Free End-Points

So far, we have only considered those problems with fixed end-points t_1 and t_2, and fixed bound constraints $\mathbf{x}(t_1) = \mathbf{x}_1$ and $\mathbf{x}(t_2) = \mathbf{x}_2$. Yet, many problems of the calculus of variations do not fall into this category.

In this subsection, we shall consider problems having a free end-point t_2 and no constraint on $\mathbf{x}(t_2)$. Instead of specifying t_2 and $\mathbf{x}(t_2)$, we add a *terminal cost* (also called *salvage term*) to the objective function, so that the problem is now in Bolza form:

$$\min_{\mathbf{x}, t_2} \int_{t_1}^{t_2} \ell(t, \mathbf{x}, \dot{\mathbf{x}}) \mathrm{d}t + \phi(t_2, \mathbf{x}(t_2)), \tag{2.20}$$

on $\mathcal{D} := \{(\mathbf{x}, t_2) \in \mathcal{C}^1[t_1, T]^{n_x} \times (t_1, T) : \mathbf{x}(t_1) = \mathbf{x}_1\}$. Other end-point problems, such as constraining either of the end-points to be on a specified curve, shall be considered later in §2.7.3.

To characterize the proper boundary condition at the right end-point, more general variations must be considered. In this objective, we suppose that the functions $\mathbf{x}(t)$ are defined by extension on a fixed "sufficiently" large interval $[t_1, T]$, and introduce the linear space $\mathcal{C}^1[t_1, T]^{n_x} \times \mathbb{R}$, supplied with the (weak) norm $\|(\mathbf{x}, t)\|_{1,\infty} := \|\mathbf{x}\|_{1,\infty} + |t|$. In particular, the geometric optimality conditions given in Theorem 2.13 can be used in the normed linear space $(\mathcal{C}^1[t_1, T]^{n_x} \times \mathbb{R}, \|(\cdot, \cdot)\|_{1,\infty})$, with the Gâteaux derivative defined as:

$$\delta\mathcal{J}(\mathbf{x}, t_2; \boldsymbol{\xi}, \tau) := \lim_{\eta \to 0} \frac{\mathcal{J}(\mathbf{x} + \eta\boldsymbol{\xi}, t_2 + \eta\tau) - \mathcal{J}(\mathbf{x})}{\eta} = \frac{\partial}{\partial \eta} \mathcal{J}(\mathbf{x} + \eta\boldsymbol{\xi}, t_2 + \eta\tau)\bigg|_{\eta=0}.$$

The idea is to derive stationarity conditions for $\mathcal{J}$ with respect to *both* $\mathbf{x}$ and the additional variable t_2. These conditions are established in the following:

Theorem 2.24 (Free Terminal Conditions Subject to a Terminal Cost). *Consider the problem to minimize the functional* $\mathcal{J}(\mathbf{x}, t_2) := \int_{t_1}^{t_2} \ell(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\mathrm{d}t + \phi(t_2, \mathbf{x}(t_2))$*, on* $\mathcal{D} := \{(\mathbf{x}, t_2) \in \mathcal{C}^1[t_1, T]^{n_x} \times (t_1, T) : \mathbf{x}(t_1) = \mathbf{x}_1\}$*, with* $\ell : \mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}$*, and* $\phi : \mathbb{R} \times \mathbb{R}^{n_x} \to \mathbb{R}$ *being continuously differentiable. Suppose that* $(\mathbf{x}^\star, t_2^\star)$ *gives a (local)*

minimum for $\mathcal{J}$ on $\mathcal{D}$. Then, $\mathbf{x}^\star$ is the solution to the Euler equation (2.12) *on the interval $[t_1, t_2^\star]$, and satisfies both the initial condition $\mathbf{x}^\star(t_1) = \mathbf{x}_1$ and the transversal conditions*

$$[\ell_{\dot{\mathbf{x}}} + \phi_{\mathbf{x}}]_{\mathbf{x}=\mathbf{x}^\star, t=t_2^\star} = \mathbf{0} \tag{2.21}$$

$$\left[\ell - \dot{\mathbf{x}}^{\mathsf{T}} \ell_{\dot{\mathbf{x}}} + \phi_t\right]_{\mathbf{x}=\mathbf{x}^\star, t=t_2^\star} = 0. \tag{2.22}$$

Proof. By fixing $t_2 = t_2^\star$ and varying $\mathbf{x}^\star$ in the $\mathcal{D}$-admissible direction $\boldsymbol{\xi} \in \mathcal{C}^1[t_1, T]^{n_x}$ such that $\boldsymbol{\xi}(t_1) = \boldsymbol{\xi}(t_2^\star) = \mathbf{0}$, it can be shown, as in the proof of Theorem 2.14, that $\mathbf{x}^\star$ must be a solution to the Euler equation (2.12) on $[t_1, t_2^\star]$.

Based on the differentiability assumptions for ℓ and ϕ, and by Theorem 2.A.60, we have

$$\begin{aligned}
\frac{\partial}{\partial \eta} \mathcal{J}(\mathbf{x}^\star + \eta\boldsymbol{\xi}, t_2^\star + \eta\tau) &= \int_{t_1}^{t_2^\star + \eta\tau} \frac{\partial}{\partial \eta} \ell\left[(\mathbf{x}^\star + \eta\boldsymbol{\xi})(t)\right] \mathrm{d}t + \ell\left[(\mathbf{x}^\star + \eta\boldsymbol{\xi})(t_2^\star + \eta\tau)\right] \tau \\
&\quad + \frac{\partial}{\partial \eta} \phi\left[(\mathbf{x}^\star + \eta\boldsymbol{\xi})(t_2^\star + \eta\tau)\right] \\
&= \int_{t_1}^{t_2^\star + \eta\tau} \ell_{\mathbf{x}}\left[(\mathbf{x}^\star + \eta\boldsymbol{\xi})(t)\right]^{\mathsf{T}} \boldsymbol{\xi} + \ell_{\dot{\mathbf{x}}}\left[(\mathbf{x}^\star + \eta\boldsymbol{\xi})(t)\right]^{\mathsf{T}} \dot{\boldsymbol{\xi}} \, \mathrm{d}t \\
&\quad + \ell\left[(\mathbf{x}^\star + \eta\boldsymbol{\xi})(t_2^\star + \eta\tau)\right] \tau + \phi_t\left[(\mathbf{x}^\star + \eta\boldsymbol{\xi})(t_2^\star + \eta\tau)\right] \tau \\
&\quad + \phi_{\mathbf{x}}\left[(\mathbf{x}^\star + \eta\boldsymbol{\xi})(t_2^\star + \eta\tau)\right]^{\mathsf{T}} \left(\boldsymbol{\xi}(t_2^\star + \eta\tau) + (\dot{\mathbf{x}} + \eta\dot{\boldsymbol{\xi}})(t_2^\star + \eta\tau)\tau\right),
\end{aligned}$$

where the usual compressed notation is used. Taking the limit as $\eta \to 0$, we get

$$\begin{aligned}
\delta\mathcal{J}(\mathbf{x}^\star, t_2^\star; \boldsymbol{\xi}, \tau) &= \int_{t_1}^{t_2^\star} \ell_{\mathbf{x}}\left[\mathbf{x}^\star\right]^{\mathsf{T}} \boldsymbol{\xi} + \ell_{\dot{\mathbf{x}}}\left[\mathbf{x}^\star\right]^{\mathsf{T}} \dot{\boldsymbol{\xi}} \, \mathrm{d}t + \ell\left[\mathbf{x}^\star(t_2^\star)\right] \tau + \phi_t\left[\mathbf{x}^\star(t_2^\star)\right] \tau \\
&\quad + \phi_{\mathbf{x}}\left[\mathbf{x}^\star(t_2^\star)\right]^{\mathsf{T}} \left(\boldsymbol{\xi}(t_2^\star) + \dot{\mathbf{x}}(t_2^\star)\tau\right),
\end{aligned}$$

which exists for each $(\boldsymbol{\xi}, \tau) \in \mathcal{C}^1[t_1, T]^{n_x} \times \mathbb{R}$. Hence, the functional $\mathcal{J}$ is Gâteaux differentiable at $(\mathbf{x}^\star, t_2^\star)$. Moreover, $\mathbf{x}^\star$ satisfying the Euler equation (2.12) on $[t_1, t_2^\star]$, we have that $\ell_{\dot{\mathbf{x}}}[\mathbf{x}^\star] \in \mathcal{C}^1[t_1, t_2^\star]^{n_x}$, and

$$\begin{aligned}
\delta\mathcal{J}(\mathbf{x}^\star, t_2^\star; \boldsymbol{\xi}, \tau) &= \int_{t_1}^{t_2^\star} \left(\ell_{\mathbf{x}}\left[\mathbf{x}^\star\right] - \frac{\mathrm{d}}{\mathrm{d}t} \ell_{\dot{\mathbf{x}}}\left[\mathbf{x}^\star\right]\right)^{\mathsf{T}} \boldsymbol{\xi} \, \mathrm{d}t + \ell_{\dot{\mathbf{x}}}\left[\mathbf{x}^\star(t_2^\star)\right]^{\mathsf{T}} \boldsymbol{\xi}(t_2^\star) \\
&\quad + \ell\left[\mathbf{x}^\star(t_2^\star)\right] \tau + \phi_t\left[\mathbf{x}^\star(t_2^\star)\right] \tau + \phi_{\mathbf{x}}\left[\mathbf{x}^\star(t_2^\star)\right]^{\mathsf{T}} \left(\boldsymbol{\xi}(t_2^\star) + \dot{\mathbf{x}}(t_2^\star)\tau\right) \\
&= \left(\phi_t\left[\mathbf{x}^\star(t_2^\star)\right] + \ell\left[\mathbf{x}^\star(t_2^\star)\right] - \ell_{\dot{\mathbf{x}}}\left[\mathbf{x}^\star(t_2^\star)\right]^{\mathsf{T}} \dot{\mathbf{x}}(t_2^\star)\right) \tau \\
&\quad + \left(\ell_{\dot{\mathbf{x}}}\left[\mathbf{x}^\star(t_2^\star)\right] + \phi_{\mathbf{x}}\left[\mathbf{x}^\star(t_2^\star)\right]\right)^{\mathsf{T}} \left(\boldsymbol{\xi}(t_2^\star) + \dot{\mathbf{x}}(t_2^\star)\tau\right).
\end{aligned}$$

By Theorem 2.13, a necessary condition of optimality is

$$\begin{aligned}
&\left(\phi_t\left[\mathbf{x}^\star(t_2^\star)\right] + \ell\left[\mathbf{x}^\star(t_2^\star)\right] - \ell_{\dot{\mathbf{x}}}\left[\mathbf{x}^\star(t_2^\star)\right]^{\mathsf{T}} \dot{\mathbf{x}}(t_2^\star)\right) \tau \\
&\qquad + \left(\ell_{\dot{\mathbf{x}}}\left[\mathbf{x}^\star(t_2^\star)\right] + \phi_{\mathbf{x}}\left[\mathbf{x}^\star(t_2^\star)\right]\right)^{\mathsf{T}} \left(\boldsymbol{\xi}(t_2^\star) + \dot{\mathbf{x}}(t_2^\star)\tau\right) = 0,
\end{aligned}$$

for each $\mathcal{D}$-admissible direction $(\boldsymbol{\xi}, \tau) \in \mathcal{C}^1[t_1, T]^{n_x} \times \mathbb{R}$.

In particular, restricting attention to those $\mathcal{D}$-admissible directions $(\boldsymbol{\xi}, \tau) \in \Xi := \{(\boldsymbol{\xi}, \tau) \in \mathcal{C}^1[t_1, T]^{n_x} \times \mathbb{R} : \boldsymbol{\xi}(t_1) = \mathbf{0}, \boldsymbol{\xi}(t_2^\star) = -\dot{\mathbf{x}}(t_2^\star)\tau\}$, we get

$$0 = \left(\phi_t\left[\mathbf{x}^\star(t_2^\star)\right] + \ell\left[\mathbf{x}^\star(t_2^\star)\right] - \ell_{\dot{\mathbf{x}}}\left[\mathbf{x}^\star(t_2^\star)\right]^{\mathsf{T}} \dot{\mathbf{x}}(t_2^\star)\right) \tau,$$

for every τ sufficiently small. Analogously, for each $i = 1, \ldots, n_x$, considering variation those $\mathcal{D}$-admissible directions $(\boldsymbol{\xi}, \tau) \in \Xi := \{(\boldsymbol{\xi}, \tau) \in \mathcal{C}^1[t_1, T]^{n_x} \times \mathbb{R} : \xi_{j \neq i}(t) = 0 \, \forall t \in [t_1, t_2], \xi_i(t_1) = 0, \tau = 0\}$, we get

$$0 \;=\; (\ell_{\dot{x}_i}\,[x_i^\star(t_2^\star)] + \phi_{x_i}\,[x_i^\star(t_2^\star)])\,\xi_i(t_2^\star),$$

for every $\xi_i(t_2^\star)$ sufficiently small. The result follows by dividing the last two equations by τ and $\xi_i(t_2^\star)$, respectively. □

Remark 2.25 (Natural Boundary Conditions). Problems wherein t_2 is left totally free can also be handled with the foregoing theorem, e.g., by setting the terminal cost to zero. Then, the transversal conditions (2.21,2.22) yield the so-called *natural boundary conditions*:

- If t_2 is free,

$$\Big[\ell - \dot{\mathbf{x}}^\mathsf{T}\ell_{\dot{\mathbf{x}}}\Big]_{\mathbf{x}=\mathbf{x}^\star, t=t_2^\star} \;=\; \mathcal{H}(t_2, \mathbf{x}^\star(t_2), \dot{\mathbf{x}}^\star(t_2)) \;=\; 0;$$

- If $\mathbf{x}(t_2)$ is free,

$$[\ell_{\dot{\mathbf{x}}}]_{\mathbf{x}=\mathbf{x}^\star, t=t_2^\star} \;=\; \mathbf{0}.$$

A summary of the necessary conditions of optimality encountered so far is given below.

Remark 2.26 (Summary of Necessary Conditions). Necessary conditions of optimality for the problem

$$\begin{aligned} &\text{minimize:} \quad \int_{t_1}^{t_2} \ell(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\mathrm{d}t + \phi(t_2, \mathbf{x}(t_2)) \\ &\text{subject to:} \quad (\mathbf{x}, t_2) \in \mathcal{D} := \{(\mathbf{x}, t_2) \in \mathcal{C}^1[t_1, t_2]^{n_x} \times [t_1, T] : \mathbf{x}(t_1) = \mathbf{x}_1\}, \end{aligned}$$

are as follow:

- Euler Equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}\ell_{\dot{\mathbf{x}}} = \ell_{\mathbf{x}}, \quad t_1 \leq t \leq t_2;$$

- Legendre Condition:

$$\ell_{\dot{\mathbf{x}}\dot{\mathbf{x}}} \text{ semi-definite positive}, \quad t_1 \leq t \leq t_2;$$

- End-Point Conditions:

$$\begin{aligned} &\mathbf{x}(t_1) = \mathbf{x}_1 \\ &\text{If } t_2 \text{ is fixed}, \quad t_2 \text{ given} \\ &\text{If } \mathbf{x}(t_2) \text{ is fixed}, \quad \mathbf{x}(t_2) = \mathbf{x}_2 \text{ given}; \end{aligned}$$

- Transversal Conditions:

$$\begin{aligned} &\text{If } t_2 \text{ is free}, \quad \Big[\ell - \dot{\mathbf{x}}^\mathsf{T}\ell_{\dot{\mathbf{x}}} + \phi_t\Big]_{t_2} = 0 \\ &\text{If } \mathbf{x}(t_2) \text{ is free}, \quad [\ell_{\dot{\mathbf{x}}} + \phi_{\mathbf{x}}]_{t_2} = \mathbf{0}. \end{aligned}$$

(Analogous conditions hold in the case where either t_1 or $\mathbf{x}(t_1)$ is free.)

Example 2.27. Consider the functional

$$\mathcal{J}(x) := \int_0^T [\dot{x}(t)]^2 \, \mathrm{d}t + W[x(T) - A]^2,$$

for $x \in \mathcal{D} := \{x \in \mathcal{C}^1[0,T] : x(0) = 0, T \text{ given}\}$. The stationary function $\bar{x}(t)$ for the Lagrangian ℓ is unique and identical to that found in Example 2.16,

$$\bar{x}(t) := c_1 t + c_2,$$

where c_1 and c_2 are constants of integration. Using the supplied end-point condition $\bar{x}(0) = 0$, we get $c_2 = 0$. Unlike Example 2.16, however, c_1 cannot be determined directly, for $\bar{x}(T)$ is given implicitly through the transversal condition (2.21) at $t = T$,

$$2\dot{\bar{x}}(T) + 2W[\bar{x}(T) - A] = 0.$$

Hence,

$$c_1 = \frac{AW}{1 + TW}.$$

Overall, the unique stationary function $\bar{x}$ is thus given by

$$\bar{x}(t) = \frac{AW}{1 + TW} t.$$

By letting the coefficient W grow large, more weight is put on the terminal term than on the integral term, which makes the end-point value $\bar{x}(T)$ closer to the value A. At the limit $W \to \infty$, we get $\bar{x}(t) \to \frac{A}{T}t$, which is precisely the stationary function found earlier in Example 2.16 (without terminal term and with end-point condition $\mathbf{x}(T) = A$).

2.6 PIECEWISE $\mathcal{C}^1$ EXTREMAL FUNCTIONS

In all the problems examined so far, the functions defining the class for optimization were required to be continuously differentiable, $\mathbf{x} \in \mathcal{C}^1[t_1, t_2]^{n_x}$. Yet, it is natural to wonder whether *cornered trajectories*, i.e., trajectories represented by *piecewise continously differentiable functions*, might not yield improved results. Besides improvement, it is also natural to wonder whether those problems of the calculus of variations which do not have extremals in the class of $\mathcal{C}^1$ functions actually have extremals in the larger class of piecewise $\mathcal{C}^1$ functions.

In the present section, we shall extend our previous investigations to include piecewise $\mathcal{C}^1$ functions as possible extremals. In §2.6.1, we start by defining the class of piecewise $\mathcal{C}^1$ functions, and discuss a number of related properties. Then, piecewise $\mathcal{C}^1$ extremals for problems of the calculus of variations are considered in §2.6.2, with emphasis on the so-called Weierstrass-Erdmann conditions which must hold at a corner point of an extremal solution. Finally, the characterization of strong extremals is addressed in §2.6.3, based on a new class of piecewise $\mathcal{C}^1$ (strong) variations.

2.6.1 The Class of Piecewise $\mathcal{C}^1$ Functions

The class of piecewise $\mathcal{C}^1$ functions is defined as follows:

Definition 2.28 (Piecewise $\mathcal{C}^1$ Functions). *A real-valued function $\hat{x} \in \mathcal{C}[a,b]$ is said to be* piecewise $\mathcal{C}^1$*, denoted $\hat{x} \in \hat{\mathcal{C}}^1[a,b]$, if there is a finite (irreducible) partition $a = c_0 < c_1 < \cdots < c_{N+1} = b$ such that $\hat{x}$ may be regarded as a function in $\mathcal{C}^1[c_k, c_{k+1}]$ for each $k = 0, 1, \ldots, N$. When present, the interior points $c_1, \ldots, c_N$ are called* corner points *of x.*

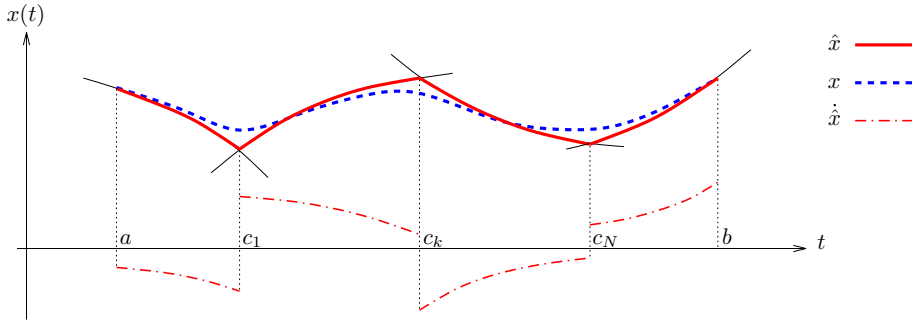


Figure 2.3. Illustration of a piecewise continuously differentiable function $\hat{x} \in \hat{\mathcal{C}}^1[a,b]$ (thick red line), and its derivative $\dot{\hat{x}}$ (dash-dotted red line); without corners, $\hat{x}$ may resemble the continuously differentiable function $x \in \mathcal{C}^1[a,b]$ (dashed blue curve).

Some remarks are in order. Observe first that, when there are no corners, then $\hat{x} \in \mathcal{C}^1[a,b]$. Further, for any $\hat{x} \in \hat{\mathcal{C}}^1[a,b]$, $\dot{\hat{x}}$ is defined and continuous on $[a,b]$, except at its corner point $c_1, \ldots, c_N$ where it has distinct limiting values $\dot{\hat{x}}(c_k^{\pm})$; such discontinuities are said to *simple*, and $\dot{\hat{x}}$ is said to be *piecewise continuous* on $[a,b]$, denoted $\dot{\hat{x}} \in \hat{\mathcal{C}}[a,b]$. Fig. 2.3. illustrates the effect of the discontinuities of $\dot{\hat{x}}$ in producing corners on the graph of $\hat{x}$. Without these corners, $\hat{x}$ might resemble the $\mathcal{C}^1$ function x, whose graph is presented for comparison. In particular, each piecewise $\mathcal{C}^1$ function is *"almost" $\mathcal{C}^1$*, in the sense that it is only necessary to *round out* the corners to produce the graph of a $\mathcal{C}^1$ function. These considerations are formalized by the following:

Lemma 2.29 (Smoothing of Piecewise $\mathcal{C}^1$ Functions). *Let $\hat{x} \in \hat{\mathcal{C}}^1[a,b]$. Then, for each $\delta > 0$, there exists $x \in \mathcal{C}^1[a,b]$ such that $x \equiv \hat{x}$, except in a neighborhood $\mathcal{B}_\delta(c_k)$ of each corner point of $\hat{x}$. Moreover, $\|x - \hat{x}\|_\infty \leq \hat{A}\delta$, where $\hat{A}$ is a constant determined by $\hat{x}$.*

Proof. See, e.g., [55, Lemma 7.2] for a proof. □

Likewise, we shall consider the class $\hat{\mathcal{C}}^1[a,b]^{n_x}$ of n_x-dimensional vector valued analogue of $\hat{\mathcal{C}}^1[a,b]$, consisting of those functions $\hat{\mathbf{x}} \in \hat{\mathcal{C}}^1[a,b]^{n_x}$ with components $\hat{x}_j \in \hat{\mathcal{C}}^1[a,b]$, $j = 1, \ldots, n_x$. The corners of such $\hat{\mathbf{x}}$ are by definition those of any one of its components $\hat{x}_j$. Note that the above lemma can be applied to each component of a given $\hat{\mathbf{x}}$, and shows that $\hat{\mathbf{x}}$ can be approximated by a $\mathbf{x} \in \mathcal{C}^1[a,b]^{n_x}$ which agrees with it except in prescribed neighborhoods of its corner points.

Both real valued and real vector valued classes of piecewise $\mathcal{C}^1$ functions form linear spaces of which the subsets of $\mathcal{C}^1$ functions are subspaces. Indeed, it is obvious that the constant multiple of one of these functions is another of the same kind, and the sum of two

such functions exhibits the piecewise continuous differentiability with respect to a suitable partition of the underlying interval $[a, b]$.

Since $\hat{\mathcal{C}}^1[a, b] \subset \mathcal{C}[a, b]$,

$$\|y\|_\infty := \max_{a \le t \le b} |y(t)|,$$

defines a norm on $\hat{\mathcal{C}}^1[a, b]$. Moreover,

$$\|y\|_{1,\infty} := \max_{a \le t \le b} |y(t)| + \sup_{t \in \bigcup_{k=0}^{N} (c_k, c_{k+1})} |\dot{y}(t)|,$$

can be shown to be another norm on $\hat{\mathcal{C}}^1[a, b]$, with $a = c_0 < c_1 < \cdots < c_N < c_{N+1} = b$ being a suitable partition for $\hat{x}$. (The space of vector valued piecewise $\mathcal{C}^1$ functions $\hat{\mathcal{C}}^1[a, b]^{n_x}$ can also be endowed with the norms $\|\cdot\|_\infty$ and $\|\cdot\|_{1,\infty}$.)

By analogy to the linear space of $\mathcal{C}^1$ functions, the maximum norms $\|\cdot\|_\infty$ and $\|\cdot\|_{1,\infty}$ are referred to as the *strong norm* and the *weak norm*, respectively; the functions which are locally extremal with respect to the former [latter] norm are said to be *strong* [*weak*] *extremal functions*.

2.6.2 The Weierstrass-Erdmann Corner Conditions

A natural question that arises when considering the class of piecewise $\mathcal{C}^1$ functions is whether a (local) extremal point for a functional in the class of $\mathcal{C}^1$ functions also gives a (local) extremal point for this functional in the larger class of piecewise $\mathcal{C}^1$ functions:

Theorem 2.30 (Piecewise $\mathcal{C}^1$ Extremals vs. $\mathcal{C}^1$ Extremals). *If $\mathbf{x}^\star$ gives a [local] extremal point for the functional*

$$\mathcal{J}(\mathbf{x}) := \int_{t_1}^{t_2} \ell(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\, \mathrm{d}t$$

on $\mathcal{D} := \{\mathbf{x} \in \mathcal{C}^1[t_1, t_2]^{n_x} : \mathbf{x}(t_1) = \mathbf{x}_1, \mathbf{x}(t_2) = \mathbf{x}_2\}$, with $\ell \in \mathcal{C}([t_1, t_2] \times \mathbb{R}^{2n_x})$, then $\mathbf{x}^\star$ also gives a [local] extremal point for $\mathcal{J}$ on $\hat{\mathcal{D}} := \{\hat{\mathbf{x}} \in \hat{\mathcal{C}}^1[t_1, t_2]^{n_x} : \hat{\mathbf{x}}(t_1) = \mathbf{x}_1, \hat{\mathbf{x}}(t_2) = \mathbf{x}_2\}$, with respect to the same $\|\cdot\|_\infty$ or $\|\cdot\|_{1,\infty}$ norm.

Proof. See, e.g., [55, Theorem 7.7] for a proof. □

On the other hand, a functional $\mathcal{J}$ may not have $\mathcal{C}^1$ extremals, but be extremized by a piecewise $\mathcal{C}^1$ function. Analogous to the developments in §2.5, we shall first seek for *weak* (local) extremals $\hat{\mathbf{x}}^\star \in \hat{\mathcal{C}}^1[t_1, t_2]^{n_x}$, i.e., extremal trajectories with respect to some weak neighborhood $\mathcal{B}_\delta^{1,\infty}(\hat{\mathbf{x}}^\star)$.

- Observe that $\hat{\mathbf{x}} \in \mathcal{B}_\delta^{1,\infty}(\hat{\mathbf{x}}^\star)$ *if and only if* $\hat{\mathbf{x}} = \hat{\mathbf{x}}^\star + \eta\hat{\boldsymbol{\xi}}$ for $\hat{\boldsymbol{\xi}} \in \hat{\mathcal{C}}^1[t_1, t_2]^{n_x}$, and a sufficiently small η. In characterizing (weak) local extremals for the functional

  $$\mathcal{J}(\hat{\mathbf{x}}) := \int_{t_1}^{t_2} \ell(t, \hat{\mathbf{x}}(t), \dot{\hat{\mathbf{x}}}(t))\, \mathrm{d}t$$

  on $\hat{\mathcal{D}} := \{\hat{\mathbf{x}} \in \hat{\mathcal{C}}^1[t_1, t_2]^{n_x} : \hat{\mathbf{x}}(t_1) = \mathbf{x}_1, \hat{\mathbf{x}}(t_2) = \mathbf{x}_2\}$, where ℓ and its partials $\ell_{\mathbf{x}}, \ell_{\dot{\mathbf{x}}}$ are continuous on $[t_1, t_2] \times \mathbb{R}^{2n_x}$, one can therefore duplicate the analysis of §2.5.2. This is done by splitting the integral into a *finite* sum of integrals with continuously differentiable integrands, then differentiating each under the integral sign. Overall, it can be shown that a (weak, local) extremal $\hat{\mathbf{x}}^\star \in \hat{\mathcal{C}}^1[t_1, t_2]^{n_x}$ must be stationary in

intervals excluding corner points, i.e., the Euler equation (2.12) is satisfied on $[t_1, t_2]$ except at corner points $c_1, \ldots, c_N$ of $\hat{\mathbf{x}}^\star$.

- Likewise, both Legendre second-order necessary conditions (§2.5.3) and convexity sufficient conditions (§2.5.4) can be shown to hold on intervals excluding corners points of a piecewise $\mathcal{C}^1$ extremal.

- Finally, *transversal conditions* corresponding to the various free end-point problems considered in §2.5.5 *remain the same*. To see this most easily, e.g., in the case where freedom is permitted only at the right end-point, suppose that $\hat{\mathbf{x}}^\star \in \hat{\mathcal{C}}^1[t_1, t_2]^{n_x}$ gives a local extremal for $\hat{\mathcal{D}}$, and let c_N be the right-most corner point of $\hat{\mathbf{x}}^\star$. Then, restricting comparison to those competing $\hat{\mathbf{x}}$ having their right-most corner point at c_N and satisfying $\hat{\mathbf{x}}(c_N) = \hat{\mathbf{x}}^\star(c_N)$, it is seen that the corresponding directions $(\hat{\boldsymbol{\xi}}, \tau)$ must utilize the end-point freedom exactly as in §2.5.5. Thus, the resulting boundary conditions are identical to (2.21) and (2.22).

Besides necessary conditions of optimality on intervals excluding corner points $c_1, \ldots, c_N$ of a local extremal $\hat{\mathbf{x}}^\star \in \hat{\mathcal{C}}^1[t_1, t_2]^{n_x}$, the discontinuities of $\dot{\hat{\mathbf{x}}}^\star$ which are permitted at each c_k are restricted. The so-called *first Weierstrass-Erdmann corner condition* are given subsequently.

Theorem 2.31 (First Weierstrass-Erdmann Corner Condition). *Let $\hat{\mathbf{x}}^\star(t)$ be a (weak) local extremal of the problem to minimize the functional*

$$\mathcal{J}(\hat{\mathbf{x}}) := \int_{t_1}^{t_2} \ell(t, \hat{\mathbf{x}}(t), \dot{\hat{\mathbf{x}}}(t))\, \mathrm{d}t$$

on $\hat{\mathcal{D}} := \{\hat{\mathbf{x}} \in \hat{\mathcal{C}}^1[t_1, t_2]^{n_x} : \hat{\mathbf{x}}(t_1) = \mathbf{x}_1, \hat{\mathbf{x}}(t_2) = \mathbf{x}_2\}$, where ℓ and its partials $\ell_{\mathbf{x}}$, $\ell_{\dot{\mathbf{x}}}$ are continuous on $[t_1, t_2] \times \mathbb{R}^{2n_x}$. Then, at every (possible) corner point $c \in (t_1, t_2)$ of $\hat{\mathbf{x}}^\star$, we have

$$\ell_{\dot{\mathbf{x}}}\left(c, \hat{\mathbf{x}}^\star(c), \dot{\hat{\mathbf{x}}}^\star(c^-)\right) \;=\; \ell_{\dot{\mathbf{x}}}\left(c, \hat{\mathbf{x}}^\star(c), \dot{\hat{\mathbf{x}}}^\star(c^+)\right), \tag{2.23}$$

where $\dot{\hat{\mathbf{x}}}^\star(c^-)$ and $\dot{\hat{\mathbf{x}}}^\star(c^+)$ denote the left and right time derivative of $\hat{\mathbf{x}}^\star$ at c, respectively.

Proof. By Euler equation (2.12), we have

$$\ell_{\dot{x}_i}(t, \hat{\mathbf{x}}^\star(t), \dot{\hat{\mathbf{x}}}^\star(t)) = \int_{t_1}^{t} \ell_{x_i}(\tau, \hat{\mathbf{x}}^\star(\tau), \dot{\hat{\mathbf{x}}}^\star(\tau)) + C \quad i = 1, \ldots, n_x,$$

for some real constant C. Therefore, the function defined by $\phi(t) := \ell_{\dot{x}_i}(t, \hat{\mathbf{x}}^\star(t), \dot{\hat{\mathbf{x}}}^\star(t))$ is continuous at each $t \in (t_1, t_2)$, even though $\dot{\hat{\mathbf{x}}}^\star(t))$ may be discontinuous at that point; that is, $\phi(c^-) = \phi(c^+)$. Moreover, $\ell_{\dot{x}_i}$ being continuous in its $1 + 2n_x$ arguments, $\hat{\mathbf{x}}^\star(t)$ being continuous at c, and $\dot{\hat{\mathbf{x}}}^\star(t)$ having finite limits $\dot{\hat{\mathbf{x}}}^\star(c^\pm)$ at c, we get

$$\ell_{\dot{x}_i}\left(c, \hat{\mathbf{x}}^\star(c), \dot{\hat{\mathbf{x}}}^\star(c^-)\right) \;= \ell_{\dot{x}_i}\left(c, \hat{\mathbf{x}}^\star(c), \dot{\hat{\mathbf{x}}}^\star(c^+)\right),$$

for each $i = 1, \ldots, n_x$. □

Remark 2.32. The first Weierstrass-Erdmann condition (2.23) shows that the discontinuities of $\dot{\hat{\mathbf{x}}}^\star$ which are permitted at corner points of a local extremal trajectory $\hat{\mathbf{x}}^\star \in$

$\hat{\mathcal{C}}^1[t_1,t_2]^{n_x}$ are those which preserve the continuity of $\ell_{\dot{\mathbf{x}}}$. Likewise, it can be shown that the continuity of the Hamiltonian function $\mathcal{H} := \ell - \dot{\mathbf{x}}\ell_{\dot{\mathbf{x}}}$ must be preserved at corner points of $\hat{\mathbf{x}}^\star$,

$$\ell\left[\hat{\mathbf{x}}^\star(c^-)\right] - \dot{\hat{\mathbf{x}}}^\star(c^-)\ell_{\dot{\mathbf{x}}}\left[\hat{\mathbf{x}}^\star(c^-)\right] \;= \ell\left[\hat{\mathbf{x}}^\star(c^+)\right] - \dot{\hat{\mathbf{x}}}^\star(c^+)\ell_{\dot{\mathbf{x}}}\left[\hat{\mathbf{x}}^\star(c^+)\right] \tag{2.24}$$

(using compressed notation), which yields the so-called *second Weierstrass-Erdmann corner condition*.

Example 2.33. Consider the problem to minimize the functional

$$\mathcal{J}(x) := \int_{-1}^{1} x(t)^2(1-\dot{x}(t))^2 \,\mathrm{d}t,$$

on $\mathcal{D} := \{x \in \mathcal{C}^1[-1,1] : x(-1) = 0, x(1) = 1\}$. Noting that the Lagrangian $\ell(y,z) := y^2(1-z)^2$ is independent of t, we have

$$\mathcal{H} := \ell(x,\dot{x}) - \dot{x}\ell_{\dot{x}}(x,\dot{x}) = x(t)^2(1-\dot{x}(t))^2 - \dot{x}(t)[2x(t)^2(\dot{x}(t)-1)] = c \quad \forall t \in [-1,1],$$

for some constant c. Upon simplification, we get

$$x(t)^2(1-\dot{x}(t)^2) = c \quad \forall t \in [-1,1].$$

With the substitution $u(t) := x(t)^2$ (so that $\dot{u}(t) := 2x(t)\dot{x}(t)$), we obtain the new equation $\dot{u}(t)^2 = 4(u(t)-c)$, which has general solution

$$u(t) := (t+k)^2 + c,$$

where k is a constant of integration. In turn, a stationary solution $\bar{x} \in \mathcal{C}^1[-1,1]$ for the Lagrangian ℓ satisfies the equation

$$\bar{x}(t)^2 = (t+k)^2 + c.$$

In particular, the boundary conditions $x(-1) = 0$ and $x(1) = 1$ produce constants $c = -(\frac{3}{4})^2$ and $k = \frac{1}{4}$. however, the resulting stationary function,

$$\bar{x}(t) = \sqrt{\left(t+\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2} = \sqrt{(t+1)\left(t-\frac{1}{2}\right)}$$

is defined only for $t \geq \frac{1}{2}$ of $t \leq -1$. Thus, there is *no* stationary function for the Lagrangian ℓ in $\mathcal{D}$.

Next, we turn to the problem of minimizing $\mathcal{J}$ in the larger set $\hat{\mathcal{D}} := \{\hat{x} \in \hat{\mathcal{C}}^1[-1,1] : \hat{x}(-1) = 0, \hat{x}(1) = 1\}$. Suppose that $\hat{x}^\star$ is a local minimizer for $\mathcal{J}$ on $\hat{\mathcal{D}}$. Then, by the Weierstrass-Erdmann condition (2.23), we must have

$$-2\hat{x}^\star(c^-)\left[1-\dot{\hat{x}}^\star(c^-)\right] \;=\; -2\hat{x}^\star(c^+)\left[1-\dot{\hat{x}}^\star(c^+)\right],$$

and since $\hat{x}^\star$ is continuous,

$$\hat{x}^\star(c)\left[\dot{\hat{x}}^\star(c^+) - \dot{\hat{x}}^\star(c^-)\right] \;=\; 0.$$

But because $\dot{\hat{x}}^\star(c^+) \neq \dot{\hat{x}}^\star(c^-)$ at a corner point, such corners are only allowed at those $c \in (-1,1)$ such that $\hat{x}^\star(c) = 0$.

Observe that the cost functional is bounded below by 0, which is attained if either $\hat{x}^\star(t) = 0$ or $\dot{\hat{x}}^\star(t) = 1$ at each t in $[-1,1]$. Since $\hat{x}^\star(1) = 1$, we must have that $\dot{\hat{x}}^\star(t) = 1$ in the largest subinterval $(c,1]$ in which $\hat{x}^\star(t) > 0$; here, this interval must continue until $c = 0$, where $\hat{x}^\star(0) = 0$, since corner points are not allowed unless $\hat{x}^\star(c) = 0$. Finally, the only function in $\hat{\mathcal{C}}^1[-1,0]$, which vanishes at -1 and 0, and increases at each nonzero value, is $\hat{x} \equiv 0$. Overall, we have shown that the piecewise $\mathcal{C}^1$ function

$$\hat{x}^\star(t) \;=\; \begin{cases} 0, & -1 \leq t \leq 0 \\ t, & 0 < t \leq 1 \end{cases}$$

is the unique global minimum point for $\mathcal{J}$ on $\hat{\mathcal{D}}$.

Corollary 2.34 (Absence of Corner Points). *Consider the problem to minimize the functional*

$$\mathcal{J}(\hat{\mathbf{x}}) := \int_{t_1}^{t_2} \ell(t, \hat{\mathbf{x}}(t), \dot{\hat{\mathbf{x}}}(t))\mathrm{d}t$$

on $\hat{\mathcal{D}} := \{\hat{\mathbf{x}} \in \hat{\mathcal{C}}^1[t_1,t_2]^{n_x} : \hat{\mathbf{x}}(t_1) = \mathbf{x}_1, \hat{\mathbf{x}}(t_2) = \mathbf{x}_2\}$. *If* $\ell_{\dot{\mathbf{x}}}(t,\mathbf{y},\mathbf{z})$ *is a strictly monotone function of* $\mathbf{z} \in \mathbb{R}^{n_x}$ *(or, equivalently,* $\ell(t,\mathbf{y},\mathbf{z})$ *is a convex function in* $\mathbf{z}$ *on* $\mathbb{R}^{n_x}$*), for each* $(t,\mathbf{y}) \in [t_1,t_2] \times \mathbb{R}^{n_x}$*, then an extremal solution* $\hat{\mathbf{x}}^\star(t)$ *cannot have corner points.*

Proof. Let $\hat{\mathbf{x}}^\star$ by an extremal solution of $\mathcal{J}$ on $\hat{\mathcal{D}}$. By contradiction, assume that $\hat{\mathbf{x}}^\star$ has a corner point at $c \in (t_1,t_2)$. Then, we have $\dot{\hat{\mathbf{x}}}^\star(c^-) \neq \dot{\hat{\mathbf{x}}}^\star(c^+)$ and, $\ell_{\dot{\mathbf{x}}}$ being a strictly monotone function of $\dot{\hat{\mathbf{x}}}$,

$$\ell_{\dot{\mathbf{x}}}\left(c, \hat{\mathbf{x}}^\star(c), \dot{\hat{\mathbf{x}}}^\star(c^-)\right) \neq \ell_{\dot{\mathbf{x}}}\left(c, \hat{\mathbf{x}}^\star(c), \dot{\hat{\mathbf{x}}}^\star(c^+)\right)$$

(since $\ell_{\dot{\mathbf{x}}}$ cannot take twice the same value). This contradicts the condition (2.23) given by Theorem 2.31 above, i.e., $\hat{\mathbf{x}}^\star$ cannot have a corner point in (t_1,t_2). □

Example 2.35 (Minimum Path Problem (cont'd)). Consider the problem to minimize the distance between two fixed points, namely $A = (x_A, y_A)$ and $B = (x_B, y_B)$, in the (x,y)-plane. We have shown, in Example 2.18 (p. 75), that extremal trajectories for this problem correspond to straight lines. But could we have extremal trajectories with corner points? The answer is no, for $\ell_{\dot{x}} : z \mapsto \frac{1}{\sqrt{1+z^2}}$ is a convex function in z on $\mathbb{R}$.

2.6.3 Weierstrass' Necessary Conditions: Strong Minima

The Gâteaux derivatives of a functional are obtained by comparing its value at a point $\mathbf{x}$ with those at points $\mathbf{x} + \eta\boldsymbol{\xi}$ in a *weak* norm neighborhood. In contrast to these (weak) variations, we now consider a new type of (*strong*) variations whose smallness does *not* imply that of their derivatives.

Specifically, the variation $\hat{W} \in \hat{\mathcal{C}}^1[t_1, t_2]$ is defined as:

$$\hat{W}(t) := \begin{cases} v(t - \tau + \delta) & \text{if } \tau - \delta \le t < \tau \\ v(-\sqrt{\delta}(t - \tau) + \delta) & \text{if } \tau \le t < \tau + \sqrt{\delta} \\ 0 & \text{otherwise.} \end{cases},$$

where $\tau \in (t_1, t_2)$, $v > 0$, and $\delta > 0$ such that

$$\tau - \delta > t_1 \quad \text{and} \quad \tau + \sqrt{\delta} < t_2.$$

Both $\hat{W}(t)$ and its time derivative are illustrated in Fig. 2.4. below. Similarly, variations $\hat{\mathbf{W}}$ can be defined in $\hat{\mathcal{C}}^1[t_1, t_2]^{n_x}$.

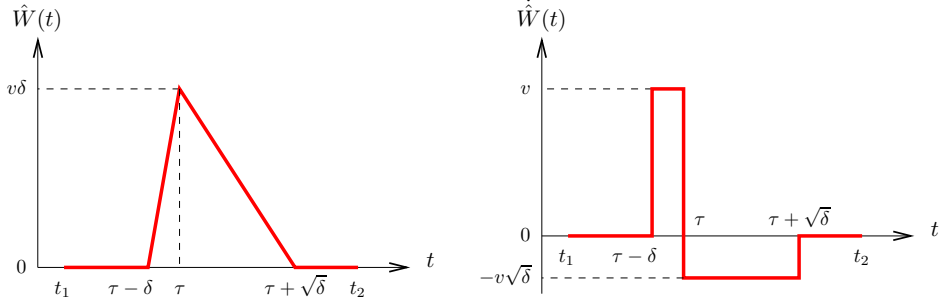


Figure 2.4. Strong variation $\hat{W}(t)$ (left plot) and its time derivative $\dot{\hat{W}}(t)$ (right plot).

The following theorem gives a necessary condition for a strong local minimum, whose proof is based on the foregoing class of variations.

Theorem 2.36 (Weierstrass' Necessary Condition). *Consider the problem to minimize the functional*

$$\mathcal{J}(\hat{\mathbf{x}}) := \int_{t_1}^{t_2} \ell(t, \hat{\mathbf{x}}(t), \dot{\hat{\mathbf{x}}}(t)) \mathrm{d}t,$$

on $\hat{\mathcal{D}} := \{\hat{\mathbf{x}} \in \hat{\mathcal{C}}^1[t_1, t_2]^{n_x} : \hat{\mathbf{x}}(t_1) = \mathbf{x}_1, \hat{\mathbf{x}}(t_2) = \mathbf{x}_2\}$. *Suppose* $\hat{\mathbf{x}}^\star(t)$, $t_1 \le t \le t_2$, *gives a* strong *(local) minimum for* $\mathcal{J}$ *on* $\hat{\mathcal{D}}$. *Then,*

$$\mathscr{E}(t, \hat{\mathbf{x}}^\star, \dot{\hat{\mathbf{x}}}^\star, \mathbf{v}) \;:=\; \ell(t, \hat{\mathbf{x}}^\star, \dot{\hat{\mathbf{x}}}^\star + \mathbf{v}) - \ell(t, \hat{\mathbf{x}}^\star, \dot{\hat{\mathbf{x}}}^\star) - \mathbf{v}^\mathsf{T} \ell_{\dot{\hat{\mathbf{x}}}}(t, \hat{\mathbf{x}}^\star, \dot{\hat{\mathbf{x}}}^\star) \;\ge\; 0, \qquad (2.25)$$

at every $t \in [t_1, t_2]$ *and for each* $\mathbf{v} \in \mathbb{R}^{n_x}$. *(*$\mathscr{E}$ *is referred to as the* excess function of Weierstrass.*)*

Proof. For the sake of clarity, we shall present and prove this condition for scalar functions $\hat{x} \in \hat{\mathcal{C}}^1[t_1, t_2]$ only. Its extension to vector functions $\hat{\mathbf{x}} \in \hat{\mathcal{C}}^1[t_1, t_2]^{n_x}$ poses no particular conceptual difficulty, and is left to the reader as an exercise ☺.

Let $\hat{x}_\delta(t) := \hat{x}^\star(t) + \hat{W}(t)$. Note that both $\hat{W}$ and $\hat{x}^\star$ being piecewise $\mathcal{C}^1$ functions, so is $\hat{x}_\delta$. These smoothness conditions are sufficient to calculate $\mathcal{J}(\hat{x}_\delta)$, as well as its derivative

with respect to δ at $\delta = 0$. By the definition of $\hat{W}$, we have

$$\begin{aligned}\mathcal{J}(\hat{x}_\delta) &= \mathcal{J}(\hat{x}^\star) - \int_{\tau-\delta}^{\tau+\sqrt{\delta}} \ell\left(t, \hat{x}^\star(t), \dot{\hat{x}}^\star(t)\right) \mathrm{d}t \\ &\quad + \int_{\tau-\delta}^{\tau} \ell\left(t, \hat{x}^\star(t) + v(t-\tau+\delta), \dot{\hat{x}}^\star(t) + v\right) \mathrm{d}t \\ &\quad + \int_{\tau}^{\tau+\sqrt{\delta}} \ell\left(t, \hat{x}^\star(t) + v(-\sqrt{\delta}(t-\tau)+\delta), \dot{\hat{x}}^\star(t) - v\sqrt{\delta}\right) \mathrm{d}t.\end{aligned}$$

That is,

$$\begin{aligned}\frac{\mathcal{J}(\hat{x}_\delta) - \mathcal{J}(\hat{x}^\star)}{\delta} &= \frac{1}{\delta}\int_{\tau-\delta}^{\tau} \ell\left(t, \hat{x}^\star(t) + v(t-\tau+\delta), \dot{\hat{x}}^\star(t) + v\right) - \ell\left(t, \hat{x}^\star(t), \dot{\hat{x}}^\star(t)\right) \mathrm{d}t \\ &+ \frac{1}{\delta}\int_{\tau}^{\tau+\sqrt{\delta}} \ell\left(t, \hat{x}^\star(t) + v(-\sqrt{\delta}(t-\tau)+\delta), \dot{\hat{x}}^\star(t) - v\sqrt{\delta}\right) - \ell\left(t, \hat{x}^\star(t), \dot{\hat{x}}^\star(t)\right) \mathrm{d}t.\end{aligned}$$

Letting $\delta \to 0$, the first term I_1^δ tends to

$$I_1^0 := \lim_{\delta\to 0} I_1^\delta = \ell\left(\tau, \hat{x}^\star(\tau), \dot{\hat{x}}^\star(\tau) + v\right) - \ell\left(\tau, \hat{x}^\star(\tau), \dot{\hat{x}}^\star(\tau)\right).$$

Letting g be the function such that $g(t) := v(-(t-\tau) + \sqrt{\delta})$, the second term I_2^δ becomes

$$\begin{aligned}I_2^\delta &= \frac{1}{\delta}\int_{\tau}^{\tau+\sqrt{\delta}} \ell\left(t, \hat{x}^\star(t) + \sqrt{\delta}g(t), \dot{\hat{x}}^\star(t) + \sqrt{\delta}\dot{g}(t)\right) - \ell\left(t, \hat{x}^\star(t), \dot{\hat{x}}^\star(t)\right) \mathrm{d}t \\ &= \frac{1}{\delta}\int_{\tau}^{\tau+\sqrt{\delta}} \ell_x^\star\sqrt{\delta}g + \ell_{\dot{x}}^\star\sqrt{\delta}\dot{g} + o(\sqrt{\delta})\,\mathrm{d}t \\ &= \frac{1}{\sqrt{\delta}}\int_{\tau}^{\tau+\sqrt{\delta}} \ell_x^\star g + \ell_{\dot{x}}^\star\dot{g}\,\mathrm{d}t + o(\sqrt{\delta}) \\ &= \frac{1}{\sqrt{\delta}}\int_{\tau}^{\tau+\sqrt{\delta}} \left(\ell_x^\star - \frac{\mathrm{d}}{\mathrm{d}t}\ell_{\dot{x}}^\star\right) g\,\mathrm{d}t + \left[\ell_{\dot{x}}^\star g\right]_\tau^{\tau+\sqrt{\delta}} + o(\sqrt{\delta}).\end{aligned}$$

Noting that $g(\tau + \sqrt{\delta}) = 0$ and $g(\tau) = -v\sqrt{\delta}$, and because $\hat{x}^\star$ verifies the Euler equation since it is a (local) minimizer for $\mathcal{J}$ on $\hat{\mathcal{D}}$, we get

$$I_2^0 := \lim_{\delta\to 0} I_2^\delta = -v\ell_{\dot{x}}^\star\left(\tau, \hat{x}^\star(\tau), \dot{\hat{x}}^\star(\tau)\right).$$

Finally, $\hat{x}^\star$ being a strong (local) minimizer for $\mathcal{J}$ on $\hat{\mathcal{D}}$, we have $\mathcal{J}(\hat{x}_\delta) \geq \mathcal{J}(\hat{x}^\star)$ for sufficiently small δ, hence

$$0 \leq \lim_{\delta\to 0} \frac{\mathcal{J}(\hat{x}_\delta) - \mathcal{J}(\hat{x}^\star)}{\delta} = I_1^0 + I_2^0,$$

and the result follows. □

Example 2.37 (Minimum Path Problem (cont'd)). Consider again the problem to minimize the distance between two fixed points $A = (x_A, y_A)$ and $B = (x_B, y_B)$, in the (x, y)-plane:

$$\min\ \mathcal{J}(\hat{y})\ :=\ \int_{x_A}^{x_B} \sqrt{1+\dot{\hat{y}}(x)^2}\,\mathrm{d}t,$$

on $\hat{\mathcal{D}}\ :=\ \{\hat{y}\ \in\ \hat{\mathcal{C}}^1[x_A, x_B]\ :\ \hat{y}(x_A)\ =\ y_A, \hat{y}(x_B)\ =\ y_B\}$. It has been shown in Example 2.18 that extremal trajectories for this problem correspond to straight lines, $\hat{y}^\star(x) = C_1 x + C_2$.

We now ask the question whether $\hat{y}^\star$ is a strong minimum for that problem? To answer this question, consider the excess function of Weierstrass at $\hat{y}^\star$,

$$\mathscr{E}(t, \hat{y}^\star, \dot{\hat{y}}^\star, v)\ :=\ \sqrt{1+(C_1+v)^2} - \sqrt{1+(C_1)^2} - \frac{v}{\sqrt{1+(C_1)^2}}.$$

a plot of this function is given in Fig. 2.5. for different values of C_1 and v in the range $[-10, 10]$. To confirm that the Weierstrass condition (2.25) is indeed satisfied at $\hat{y}^\star$, observe that the function $g : z \mapsto \sqrt{1+z^2}$ is convex in z on $\mathbb{R}$. Hence, for a fixed $z^\star \in \mathbb{R}$, we have (see Theorem A.17):

$$g(z^\star + v) - g(x^\star) - v\nabla g(x^\star)\ \geq\ 0\quad \forall v \in \mathbb{R},$$

which is precisely the Weierstrass' condition. This result based on convexity is generalized in the following remark.

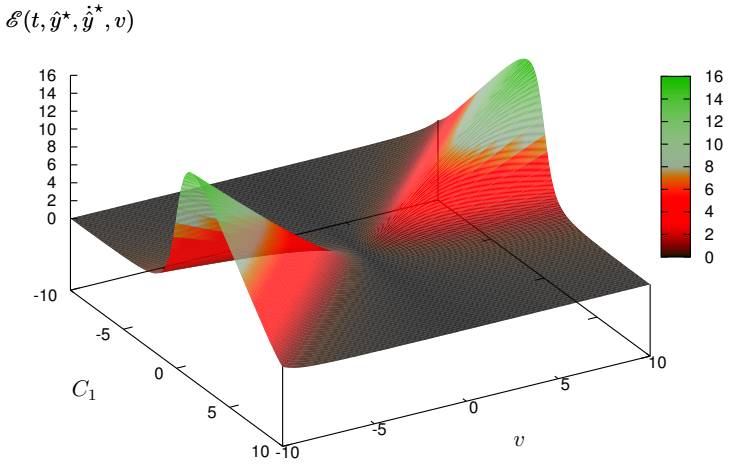


Figure 2.5. Excess function of Weierstrass for the minimum path problem.

The following corollary indicates that the Weierstrass condition is also useful to detect strong (local) minima in the class of $\mathcal{C}^1$ functions.

Corollary 2.38 (Weierstrass' Necessary Condition). *Consider the problem to minimize the functional*

$$\mathcal{J}(\mathbf{x}) := \int_{t_1}^{t_2} \ell(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\mathrm{d}t,$$

on $\mathcal{D} := \{\mathbf{x} \in \mathcal{C}^1[t_1, t_2]^{n_x} : \mathbf{x}(t_1) = \mathbf{x}_1, \mathbf{x}(t_2) = \mathbf{x}_2\}$. *Suppose* $\mathbf{x}^\star(t)$, $t_1 \leq t \leq t_2$, *gives a* strong *(local) minimum for* $\mathcal{J}$ *on* $\mathcal{D}$. *Then,*

$$\mathscr{E}(t, \mathbf{x}^\star, \dot{\mathbf{x}}^\star, \mathbf{v}) \;:=\; \ell(t, \mathbf{x}^\star, \dot{\mathbf{x}}^\star + \mathbf{v}) - \ell(t, \mathbf{x}^\star, \dot{\mathbf{x}}^\star) - \mathbf{v}^\mathsf{T} \ell_{\dot{\mathbf{x}}}(t, \mathbf{x}^\star, \dot{\mathbf{x}}^\star) \;\geq\; 0,$$

at every $t \in [t_1, t_2]$ *and for each* $\mathbf{v} \in \mathbb{R}^{n_x}$.

Proof. By Theorem 2.30, we know that $\mathbf{x}^\star$ also minimizes $\mathcal{J}$ on $\hat{\mathcal{D}}$ locally with respect to the strong norm $\|\cdot\|$, so the above Theorem 2.36 is applicable. □

Remark 2.39 (Weierstrass' Condition and Convexity). It is readily seen that the Weierstrass condition (2.25) is satisfied *automatically* when the Lagrange function $\ell(t, \mathbf{y}, \mathbf{z})$ is partially convex (and continuously differentiable) in $\mathbf{z} \in \mathbb{R}^{n_x}$, for each $(t, \mathbf{y}) \in [t_1, t_2] \times \mathbb{R}^{n_x}$.

Remark 2.40 (Weierstrass' Condition and Pontryagin's Maximum Principle). Interestingly enough, the Weierstrass' condition (2.25) can be rewritten as

$$\ell(t, x^\star, \dot{x}^\star + v) - (\dot{x}^\star + v)\, \ell_{\dot{x}}(t, x^\star, \dot{x}^\star) \;\geq\; \ell(t, x^\star, \dot{x}^\star) - \dot{x}^\star\, \ell_{\dot{x}}(t, x^\star, \dot{x}^\star).$$

That is, given the definition of the Hamiltonian function $\mathcal{H}$ (see Remark 2.17), we get

$$\mathcal{H}(t, x^\star, \dot{x}^\star) \;\geq\; \mathcal{H}(t, x^\star, \dot{x}^\star + v),$$

at every $t \in [t_1, t_2]$ and for each $v \in \mathbb{R}$. This necessary condition prefigures Pontryagin's Maximum Principle in optimal control theory.

2.7 PROBLEMS WITH EQUALITY AND INEQUALITY CONSTRAINTS

We now turn our attention to problems of the calculus of variations in the presence of constraints. Similar to finite-dimensional optimization, it is more convenient to use *Lagrange multipliers* in order to derive the necessary conditions of optimality associated with such problems; these considerations are discussed in §2.7.1 and §2.7.2, for equality and inequality constraints, respectively. The method of Lagrange multipliers is then used in §2.7.3 and §2.7.4 to obtain necessary conditions of optimality for problems subject to (equality) terminal constraints and isoperimetric constraints, respectively.

2.7.1 Method of Lagrange Multipliers: Equality Constraints

We saw in §2.5.1 that a usable characterization for a (local) extremal point $\mathbf{x}^\star$ to a functional $\mathcal{J}$ on a subset $\mathcal{D}$ of a normed linear space $(\mathcal{X}, \|\cdot\|)$ can be obtained by considering the Gâteaux variations $\delta\mathcal{J}(\mathbf{x}^\star; \boldsymbol{\xi})$ along the $\mathcal{D}$-admissible directions $\boldsymbol{\xi}$ at $\mathbf{x}^\star$. However, there are subsets $\mathcal{D}$ for which the set of $\mathcal{D}$-admissible directions is empty, possibly at every point in $\mathcal{D}$. In this case, the abovementioned characterization does not provide valuable information, and alternative conditions must be sought to characterize possible extremal points.

Example 2.41 (Empty Set of Admissible Directions). Consider the set $\mathcal{D}$ defined as

$$\mathcal{D} := \{\mathbf{x} \in \mathcal{C}^1[t_1, t_2]^2 : \sqrt{\frac{1}{2}x_1(t)^2 + \frac{1}{2}x_2(t)^2} = 1, \forall t \in [t_1, t_2]\},$$

i.e., $\mathcal{D}$ is the set of smooth curves lying on a cylinder whose axis is the time axis and whose cross sections are circles of radius $\sqrt{2}$ centered at $x_1(t) = x_2(t) = 0$. Let $\bar{\mathbf{x}} \in \mathcal{C}^1[t_1, t_2]^2$ such that $\bar{x}_1(t) = \bar{x}_2(t) = 1$ for each $t \in [t_1, t_2]$. Clearly, $\bar{\mathbf{x}} \in \mathcal{D}$. But, for every nonzero direction $\boldsymbol{\xi} \in \mathcal{C}^1[t_1, t_2]^2$ and every scalar $\eta \neq 0$, $\bar{\mathbf{x}} + \eta\boldsymbol{\xi} \notin \mathcal{D}$. That is, the set of $\mathcal{D}$-admissible directions is empty, for any functional $\mathcal{J} : \mathcal{C}^1[t_1, t_2]^2 \to \mathbb{R}$ (see Definition 2.12).

In this subsection, we shall discuss the method of Lagrange multipliers. The idea behind it is to characterize the (local) extremals of a functional $\mathcal{J}$ defined in a normed linear space $\mathcal{X}$, when restricted to one or more level sets of other such functionals. We have already encountered many examples of constraining relations in the previous section, and most of them can be described in terms of level sets of appropriately defined functionals:

Example 2.42. The set $\mathcal{D}$ defined as

$$\mathcal{D} := \{x \in \mathcal{C}^1[t_1, t_2] : x(t_1) = x_1, x(t_2) = x_2\},$$

can be seen as the intersection of the x_1-level set of the functional $\mathcal{G}_1(x) := x(t_1)$ with that of the x_2-level set of the functional $\mathcal{G}_2(x) := x(t_2)$. That is,

$$\mathcal{D} \;=\; \Gamma_1(x_1) \cap \Gamma_2(x_2),$$

where $\Gamma_i(K) := \{x \in \mathcal{C}^1[t_1, t_2] : \mathcal{G}_i(x) = K\}$, $i = 1, 2$.

Example 2.43. Consider the same set $\mathcal{D}$ as in Example 2.41 above. Then, $\mathcal{D}$ can be seen as the intersection of 1-level set of the (family of) functionals $\mathcal{G}_t$ defined as:

$$\mathcal{G}_t(\mathbf{x}) := \sqrt{\frac{1}{2}x_1(t)^2 + \frac{1}{2}x_2(t)^2},$$

for $t_1 \leq t \leq t_2$. That is,

$$\mathcal{D} \;=\; \bigcap_{t \in [t_1, t_2]} \Gamma_t(1),$$

where $\Gamma_\theta(K) := \{x \in \mathcal{C}^1[t_1, t_2] : \mathcal{G}_\theta(x) = K\}$. Note that we get an uncountable number of functionals in this case! This also illustrate why problems having path (or Lagrangian) constraints are admittedly hard to solve...

The following Lemma gives conditions under which a point $\bar{\mathbf{x}}$ in a normed linear space $(\mathcal{X}, \|\cdot\|)$ cannot be a (local) extremal of a functional $\mathcal{J}$, when *constrained* to the level set of another functional $\mathcal{G}$.

Lemma 2.44. *Let $\mathcal{J}$ and $\mathcal{G}$ be functionals defined in a neighborhood of $\bar{\mathbf{x}}$ in a normed linear space $(\mathcal{X}, \|\cdot\|)$, and let $K := \mathcal{G}(\bar{\mathbf{x}})$. Suppose that there exist (fixed) directions $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \in \mathcal{X}$ such that the Gâteaux derivatives of $\mathcal{J}$ and $\mathcal{G}$ satisfy the Jacobian condition*

$$\begin{vmatrix} \delta\mathcal{J}(\bar{\mathbf{x}}; \boldsymbol{\xi}_1) & \delta\mathcal{J}(\bar{\mathbf{x}}; \boldsymbol{\xi}_2) \\ \delta\mathcal{G}(\bar{\mathbf{x}}; \boldsymbol{\xi}_1) & \delta\mathcal{G}(\bar{\mathbf{x}}; \boldsymbol{\xi}_2) \end{vmatrix} \neq 0,$$

and are continuous in a neighborhood of $\bar{\mathbf{x}}$ *(in each direction* $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2$*). Then,* $\bar{\mathbf{x}}$ *cannot be a local extremal for* $\mathcal{J}$ *when constrained to* $\Gamma(K) := \{\mathbf{x} \in \mathcal{X} : \mathcal{G}(\mathbf{x}) = K\}$*, the level set of* $\mathcal{G}$ *through* $\bar{\mathbf{x}}$*.*

Proof. Consider the auxiliary functions

$$j(\eta_1, \eta_2) \;:=\; \mathcal{J}(\bar{\mathbf{x}} + \eta_1 \boldsymbol{\xi_1} + \eta_2 \boldsymbol{\xi}_2) \qquad \text{and} \qquad g(\eta_1, \eta_2) \;:=\; \mathcal{G}(\bar{\mathbf{x}} + \eta_1 \boldsymbol{\xi_1} + \eta_2 \boldsymbol{\xi}_2).$$

Both functions j and g are defined in a neighborhood of $(0,0)$ in $\mathbb{R}^2$, since $\mathcal{J}$ and $\mathcal{G}$ are themselves defined in a neighborhood of $\bar{\mathbf{x}}$. Moreover, the Gâteaux derivatives of $\mathcal{J}$ and $\mathcal{G}$ being continuous in the directions $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ around $\bar{\mathbf{x}}$, the partials of j and g with respect to η_1 and η_2,

$$\begin{aligned} j_{\eta_1}(\eta_1, \eta_2) \;&=\; \delta\mathcal{J}(\bar{\mathbf{x}} + \eta_1 \boldsymbol{\xi}_1 + \eta_2 \boldsymbol{\xi}_2; \boldsymbol{\xi}_1), \quad j_{\eta_2}(\eta_1, \eta_2) \;=\; \delta\mathcal{J}(\bar{\mathbf{x}} + \eta_1 \boldsymbol{\xi}_1 + \eta_2 \boldsymbol{\xi}_2; \boldsymbol{\xi}_2), \\ g_{\eta_1}(\eta_1, \eta_2) \;&=\; \delta\mathcal{J}(\bar{\mathbf{x}} + \eta_1 \boldsymbol{\xi}_1 + \eta_2 \boldsymbol{\xi}_2; \boldsymbol{\xi}_1), \quad g_{\eta_2}(\eta_1, \eta_2) \;=\; \delta\mathcal{J}(\bar{\mathbf{x}} + \eta_1 \boldsymbol{\xi}_1 + \eta_2 \boldsymbol{\xi}_2; \boldsymbol{\xi}_2), \end{aligned}$$

exist and are continuous in a neighborhood of $(0,0)$. Observe also that the non-vanishing determinant of the hypothesis is equivalent to the condition

$$\left| \begin{array}{cc} \frac{\partial j}{\partial \eta_1} & \frac{\partial j}{\partial \eta_2} \\ \frac{\partial g}{\partial \eta_1} & \frac{\partial g}{\partial \eta_2} \end{array} \right|_{\eta_1 = \eta_2 = 0} \neq\; 0.$$

Therefore, the inverse function theorem 2.A.61 applies, i.e., the application $\Phi := (j, g)$ maps a neighborhood of $(0,0)$ in $\mathbb{R}^2$ onto a region containing a full neighborhood of $(\mathcal{J}(\bar{\mathbf{x}}), \mathcal{G}(\bar{\mathbf{x}}))$. That is, one can find pre-image points (η_1^-, η_2^-) and (η_1^+, η_2^+) such that

$$\begin{aligned} & \mathcal{J}(\bar{\mathbf{x}} + \eta_1^- \boldsymbol{\xi}_1 + \eta_2^- \boldsymbol{\xi}_2) \;<\; \mathcal{J}(\bar{\mathbf{x}}) \;<\; \mathcal{J}(\bar{\mathbf{x}} + \eta_1^+ \boldsymbol{\xi}_1 + \eta_2^+ \boldsymbol{\xi}_2) \\ \text{while} \quad & \mathcal{G}(\bar{\mathbf{x}} + \eta_1^- \boldsymbol{\xi}_1 + \eta_2^- \boldsymbol{\xi}_2) \;=\; \mathcal{G}(\bar{\mathbf{x}}) \;=\; \mathcal{G}(\bar{\mathbf{x}} + \eta_1^+ \boldsymbol{\xi}_1 + \eta_2^+ \boldsymbol{\xi}_2), \end{aligned}$$

as illustrated in Fig. 2.6. This shows that $\bar{\mathbf{x}}$ cannot be a local extremal for $\mathcal{J}$. □

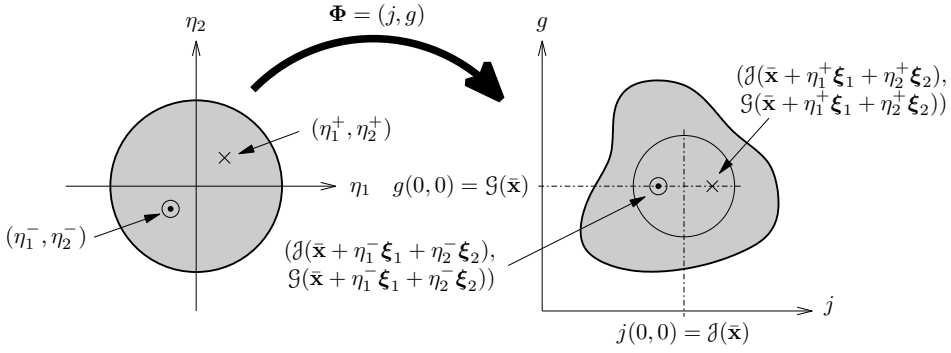


Figure 2.6. Mapping of a neighborhood of $(0,0)$ in $\mathbb{R}^2$ onto a (j, g) set containing a full neighborhood of $(j(0,0), g(0,0)) = (\mathcal{J}(\bar{\mathbf{x}}), \mathcal{G}(\bar{\mathbf{x}}))$.

With this preparation, it is easy to derive necessary conditions for a local extremal in the presence of equality or inequality constraints, and in particular the existence of the Lagrange multipliers.

Theorem 2.45 (Existence of the Lagrange Multipliers). *Let $\mathcal{J}$ and $\mathcal{G}$ be functionals defined in a neighborhood of $\mathbf{x}^\star$ in a normed linear space $(\mathcal{X}, \|\cdot\|)$, and having continuous Gâteaux derivatives in this neighborhood.* [3] *Let also $K := \mathcal{G}(\mathbf{x}^\star)$, and suppose that $\mathbf{x}^\star$ is a (local) extremal for $\mathcal{J}$ constrained to $\Gamma(K) := \{\mathbf{x} \in \mathcal{X} : \mathcal{G}(\mathbf{x}) = K\}$. Suppose further that $\delta\mathcal{G}(\mathbf{x}^\star; \bar{\boldsymbol{\xi}}) \neq 0$ for some direction $\bar{\boldsymbol{\xi}} \in \mathcal{X}$. Then, there exists a scalar $\lambda \in \mathbb{R}$ such that*

$$\delta\mathcal{J}(\mathbf{x}^\star; \boldsymbol{\xi}) + \lambda\, \delta\mathcal{G}(\mathbf{x}^\star; \boldsymbol{\xi}) \;=\; 0, \quad \forall \boldsymbol{\xi} \in \mathcal{X}.$$

Proof. Since $\mathbf{x}^\star$ is a (local) extremal for $\mathcal{J}$ constrained to $\Gamma(K)$, by Lemma 2.44, we must have that the determinant

$$\begin{vmatrix} \delta\mathcal{J}(\mathbf{x}^\star; \bar{\boldsymbol{\xi}}) & \delta\mathcal{J}(\mathbf{x}^\star; \boldsymbol{\xi}) \\ \delta\mathcal{G}(\mathbf{x}^\star; \bar{\boldsymbol{\xi}}) & \delta\mathcal{G}(\mathbf{x}^\star; \boldsymbol{\xi}) \end{vmatrix} \;=\; 0,$$

for any $\boldsymbol{\xi} \in \mathcal{X}$. Hence, with

$$\lambda \;:=\; -\frac{\delta\mathcal{J}(\mathbf{x}^\star; \bar{\boldsymbol{\xi}})}{\delta\mathcal{G}(\mathbf{x}^\star; \bar{\boldsymbol{\xi}})},$$

it follows that $\delta\mathcal{J}(\mathbf{x}^\star; \boldsymbol{\xi}) + \lambda\, \delta\mathcal{G}(\mathbf{x}^\star; \boldsymbol{\xi}) = 0$, for each $\boldsymbol{\xi} \in \mathcal{X}$. □

As in the finite-dimensional case, the parameter λ in Theorem 2.45 is called a *Lagrange multiplier*. Using the terminology of directional derivatives appropriate to $\mathbb{R}^{n_x}$, the Lagrange condition $\delta\mathcal{J}(\mathbf{x}^\star; \boldsymbol{\xi}) = -\lambda\delta\mathcal{G}(\mathbf{x}^\star; \boldsymbol{\xi})$ says simply that the directional derivatives of $\mathcal{J}$ are proportional to those of $\mathcal{G}$ at $\mathbf{x}^\star$. Thus, in general, Lagrange's condition means that the level sets of both $\mathcal{J}$ and $\mathcal{G}$ at $\mathbf{x}^\star$ share the same tangent hyperplane at $\mathbf{x}^\star$, i.e., they meet tangentially. Note also that the Lagrange's condition can also be written in the form

$$\delta(\mathcal{J} + \lambda\mathcal{G})(\mathbf{x}^\star; \cdot) \;=\; 0,$$

which suggests consideration of the Lagrangian function $\mathcal{L} := \mathcal{J} + \lambda\mathcal{G}$ as in Remark 1.53 (p. 25).

It is straightforward, albeit technical, to extend the method of Lagrange multipliers to problems involving any finite number of constraint functionals:

Theorem 2.47 (Existence of the Lagrange Multipliers (Multiple Constraints)). *Let $\mathcal{J}$ and $\mathcal{G}_1, \ldots, \mathcal{G}_{n_g}$ be functionals defined in a neighborhood of $\mathbf{x}^\star$ in a normed linear space $(\mathcal{X}, \|\cdot\|)$, and having continuous Gâteaux derivatives in this neighborhood. Let also $K_i := \mathcal{G}_i(\mathbf{x}^\star)$, $i = 1, \ldots, n_g$, and suppose that $\mathbf{x}^\star$ is a (local) extremal for $\mathcal{J}$ constrained to $\mathcal{G}_{\mathbf{K}} := \{\mathbf{x} \in \mathcal{X} : \mathcal{G}_i(\mathbf{x}) = K_i, i = 1, \ldots, n_g\}$. Suppose further that*

$$\begin{vmatrix} \delta\mathcal{G}_1(\mathbf{x}^\star; \bar{\boldsymbol{\xi}}_1) & \cdots & \delta\mathcal{G}_1(\mathbf{x}^\star; \bar{\boldsymbol{\xi}}_{n_g}) \\ \vdots & \ddots & \vdots \\ \delta\mathcal{G}_{n_g}(\mathbf{x}^\star; \bar{\boldsymbol{\xi}}_1) & \cdots & \delta\mathcal{G}_{n_g}(\mathbf{x}^\star; \bar{\boldsymbol{\xi}}_{n_g}) \end{vmatrix} \;\neq\; 0,$$

for n_g (independent) directions $\bar{\boldsymbol{\xi}}_1, \ldots, \bar{\boldsymbol{\xi}}_{n_g} \in \mathcal{X}$. Then, there exists a vector $\boldsymbol{\lambda} \in \mathbb{R}^{n_g}$ such that

$$\delta\mathcal{J}(\mathbf{x}^\star; \boldsymbol{\xi}) + [\delta\mathcal{G}_1(\mathbf{x}^\star; \boldsymbol{\xi}) \;\cdots\; \delta\mathcal{G}_{n_g}(\mathbf{x}^\star; \boldsymbol{\xi})]\, \boldsymbol{\lambda} \;=\; 0, \quad \forall \boldsymbol{\xi} \in \mathcal{X}.$$

[3] In fact, it suffices to require that $\mathcal{J}$ and $\mathcal{G}$ have *weakly continuous* Gâteaux derivatives in a neighborhood of $\mathbf{x}^\star$ for the result to hold:

Definition 2.46 (Weakly Continuous Gâteaux Variations). *The Gâteaux variations $\delta\mathcal{J}(\mathbf{x}; \boldsymbol{\xi})$ of a functional $\mathcal{J}$ defined on a normed linear space $(\mathcal{X}, \|\cdot\|)$ are said to be* weakly continuous *at $\bar{\mathbf{x}}$ provided that $\delta\mathcal{J}(\mathbf{x}; \boldsymbol{\xi}) \to \delta\mathcal{J}(\bar{\mathbf{x}}; \boldsymbol{\xi})$ as $\mathbf{x} \to \bar{\mathbf{x}}$, for each $\boldsymbol{\xi} \in \mathcal{X}$.*

Proof. See, e.g., [55, Theorem 5.16] for a proof. □

Remark 2.48 (Link to Nonlinear Optimization). The previous theorem is the generalization of Theorem 1.50 (p. 24) to optimization problems in normed linear spaces. Note, in particular, that the requirement that $\mathbf{x}^\star$ be a regular point (see Definition 1.47) for the Lagrange multipliers to exist is generalized by a non-singularity condition in terms of the Gâteaux derivatives of the constraint functionals $\mathcal{G}_1, \ldots, \mathcal{G}_{n_g}$. Yet, this condition is in general *not* sufficient to guarantee uniqueness of the Lagrange multipliers.

Remark 2.49 (Hybrid Method of Admissible Directions and Lagrange Multipliers). If $\mathbf{x}^\star \in \mathcal{D}$, with $\mathcal{D}$ a subset of a normed linear space $(\mathcal{X}, \|\cdot\|)$, and the $\mathcal{D}$-admissible directions form a *linear* subspace of $\mathcal{X}$ (i.e., $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \in \mathcal{D} \Rightarrow \eta_1\boldsymbol{\xi}_1 + \eta_2\boldsymbol{\xi}_2 \in \mathcal{D}$ for every scalars $\eta_1, \eta_2 \in \mathbb{R}$), then the conclusions of Theorem 2.47 remain valid when further restricting the continuity requirement for $\mathcal{J}$ to $\mathcal{D}$ and considering $\mathcal{D}$-admissible directions only. Said differently, Theorem 2.47 can be applied to determine (local) extremals to the functional $\mathcal{J}|_{\mathcal{D}}$ constrained to $\Gamma(\mathbf{K})$. This extension leads to a more efficient but admittedly hybrid approach to certain problems involving multiple constraints: *Those constraints on $\mathcal{J}$ which determine a domain $\mathcal{D}$ having a linear subspace of $\mathcal{D}$-admissible directions, may be taken into account by simply restricting the set of admissible directions when applying the method of Lagrangian multipliers to the remaining constraints.*

Example 2.50. Consider the problem to minimize $\mathcal{J}(x) := \int_{-1}^{0} [\dot{x}(t)]^3 \, \mathrm{d}t$, on $\mathcal{D} := \{x \in \mathcal{C}^1[-1, 0] : x(-1) = 0, x(0) = 1\}$, under the constraining relation $\mathcal{G}(x) := \int_{-1}^{0} t\,\dot{x}(t) \, \mathrm{d}t = -1$.

A possible way of dealing with this problem is by characterizing $\mathcal{D}$ as the intersection of the 0- and 1-level sets of the functionals $\mathcal{G}_1(x) := x(-1)$, $\mathcal{G}_2(x) := x(0)$, respectively, and then apply Theorem 2.47 with $n_g = 3$ constraints. But, since the $\mathcal{D}$-admissible directions for $\mathcal{J}$ at any point $x \in \mathcal{D}$ form a linear subspace of $\mathcal{C}^1[-1, 0]$, we may as well utilize a hybrid approach, as discussed in Remark 2.49 above, to solve the problem.

Necessary conditions of optimality for problems with end-point constraints and isoperimetric constraints shall be obtained with this hybrid approach in §2.7.3 and §2.7.4, respectively.

2.7.2 Extremals with Inequality Constraints

The method of Lagrange multipliers can also be used to address problems of the calculus of variations having inequality constraints (or mixed equality and inequality constraints), as shown by the following:

Theorem 2.51 (Existence of Uniqueness of the Lagrange Multipliers (Inequality Constraints)). *Let $\mathcal{J}$ and $\mathcal{G}_1, \ldots, \mathcal{G}_{n_g}$ be functionals defined in a neighborhood of $\mathbf{x}^\star$ in a normed linear space $(\mathcal{X}, \|\cdot\|)$, and having* continuous *and* linear *Gâteaux derivatives in this neighborhood. Suppose that $\mathbf{x}^\star$ is a (local)* minimum *point for $\mathcal{J}$ constrained to $\Gamma(\mathbf{K}0 := \{\mathbf{x} \in \mathcal{X} : \mathcal{G}_i(\mathbf{x}) \leq K_i, i = 1, \ldots, n_g\}$, for some constant vector $\mathbf{K}$. Suppose further that $n_a \leq n_g$ constraints, say $\mathcal{G}_1, \ldots, \mathcal{G}_{n_a}$ for simplicity, are active at $\mathbf{x}^\star$, and*

satisfy the regularity condition

$$\det \mathsf{G} \;=\; \left| \begin{array}{ccc} \delta \mathcal{G}_1(\mathbf{x}^\star; \bar{\boldsymbol{\xi}}_1) & \cdots & \delta \mathcal{G}_1(\mathbf{x}^\star; \bar{\boldsymbol{\xi}}_{n_a}) \\ \vdots & \ddots & \vdots \\ \delta \mathcal{G}_{n_a}(\mathbf{x}^\star; \bar{\boldsymbol{\xi}}_1) & \cdots & \delta \mathcal{G}_{n_a}(\mathbf{x}^\star; \bar{\boldsymbol{\xi}}_{n_a}) \end{array} \right| \neq 0,$$

for n_a (independent) directions $\bar{\boldsymbol{\xi}}_1, \ldots, \bar{\boldsymbol{\xi}}_{n_a} \in \mathcal{X}$ (the remaining constraints being inactive). Then, there exists a vector $\boldsymbol{\nu} \in \mathbb{R}^{n_g}$ such that

$$\delta \mathcal{J}(\mathbf{x}^\star; \boldsymbol{\xi}) + \left[\begin{array}{ccc} \delta \mathcal{G}_1(\mathbf{x}^\star; \boldsymbol{\xi}) & \ldots & \delta \mathcal{G}_{n_g}(\mathbf{x}^\star; \boldsymbol{\xi}) \end{array} \right] \boldsymbol{\nu} \;=\; 0, \quad \forall \boldsymbol{\xi} \in \mathcal{X}, \tag{2.26}$$

and,

$$\nu_i \;\geq 0 \tag{2.27}$$

$$(\mathcal{G}_i(\mathbf{x}^\star) - K_i)\, \nu_i \;= 0, \tag{2.28}$$

for $i = 1, \ldots, n_g$.

Proof. Since the inequality constraints $\mathcal{G}_{n_a+1}, \ldots, \mathcal{G}_{n_g}$ are inactive, the conditions (2.27) and (2.28) are trivially satisfied by taking $\nu_{n_a+1} = \cdots = \nu_{n_g} = 0$. On the other hand, since the inequality constraints $\mathcal{G}_1, \ldots, \mathcal{G}_{n_a}$ are active and satisfy a regularity condition at $\mathbf{x}^\star$, the conclusion that there exists a unique vector $\boldsymbol{\lambda} \in \mathbb{R}^{n_g}$ such that (2.26) holds follows from Theorem 2.47; moreover, (2.28) is trivially satisfied for $i = 1 \ldots, n_a$. Hence, it suffices to prove that the Lagrange multipliers $\nu_1 = \cdots = \nu_{n_a}$ cannot assume negative values when $\mathbf{x}^\star$ is a (local) minimum.

We show the result by contradiction. Without loss of generality, suppose that $\nu_1 < 0$, and consider the $(n_a + 1) \times n_a$ matrix A defined by

$$\mathsf{A} \;=\; \left(\begin{array}{ccc} \delta \mathcal{J}(\mathbf{x}^\star; \bar{\boldsymbol{\xi}}_1) & \cdots & \delta \mathcal{J}(\mathbf{x}^\star; \bar{\boldsymbol{\xi}}_{n_a}) \\ \delta \mathcal{G}_1(\mathbf{x}^\star; \bar{\boldsymbol{\xi}}_1) & \cdots & \delta \mathcal{G}_1(\mathbf{x}^\star; \bar{\boldsymbol{\xi}}_{n_a}) \\ \vdots & \ddots & \vdots \\ \delta \mathcal{G}_{n_a}(\mathbf{x}^\star; \bar{\boldsymbol{\xi}}_1) & \cdots & \delta \mathcal{G}_{n_a}(\mathbf{x}^\star; \bar{\boldsymbol{\xi}}_{n_a}) \end{array} \right).$$

By hypothesis, $\operatorname{rank} \mathsf{A}^\mathsf{T} \geq n_a - 1$, hence the null space of A has dimension lower than or equal to 1. But from (2.26), the nonzero vector $(1, \nu_1, \ldots, \nu_{n_a})^\mathsf{T} \in \ker \mathsf{A}^\mathsf{T}$. That is, $\ker \mathsf{A}^\mathsf{T}$ has dimension equal to 1, and

$$\mathsf{A}^\mathsf{T}\mathbf{y} < 0 \quad \text{only if } \exists \eta \in \mathbb{R} \text{ such that } \mathbf{y} = \eta(1, \nu_1, \ldots, \nu_{n_a})^\mathsf{T}.$$

Because $\nu_1 < 0$, there does not exist a $\mathbf{y} \neq 0$ in $\ker \mathsf{A}^\mathsf{T}$ such that $y_i \geq 0$ for every $i = 1, \ldots, n_a + 1$. Thus, by Gordan's Theorem 1.A.78, there exists a nonzero vector $\mathbf{p} \geq \mathbf{0}$ in $\mathbb{R}^{n_a}$ such that $\mathsf{A}\mathbf{p} < \mathbf{0}$, or equivalently,

$$\sum_{k=1}^{n_a} p_k \delta \mathcal{J}\left(\mathbf{x}^\star; \bar{\boldsymbol{\xi}}_k\right) < 0$$

$$\sum_{k=1}^{n_a} p_k \delta \mathcal{G}_i\left(\mathbf{x}^\star; \bar{\boldsymbol{\xi}}_k\right) < 0, \quad i = 1, \ldots, n_a.$$

The Gâteaux derivatives of $\mathcal{J}, \mathcal{G}_1, \ldots, \mathcal{G}_{n_a}$ being linear (by assumption), we get

$$\begin{aligned} \delta\mathcal{J}\left(\mathbf{x}^\star; \textstyle\sum_{k=1}^{n_a} p_k \bar{\boldsymbol{\xi}}_k\right) &< 0 \\ \delta\mathcal{G}_i\left(\mathbf{x}^\star; \textstyle\sum_{k=1}^{n_a} p_k \bar{\boldsymbol{\xi}}_k\right) &< 0, \quad i = 1, \ldots, n_a. \end{aligned} \tag{2.29}$$

In particular,

$$\exists \delta > 0 \text{ such that } \mathbf{x}^\star + \eta \textstyle\sum_{k=1}^{n_a} p_k \bar{\boldsymbol{\xi}}_k \in \Gamma(\mathbf{K}), \quad \forall 0 \leq \eta \leq \delta.$$

That is, $\mathbf{x}^\star$ being a local minimum of $\mathcal{J}$ on $\mathcal{G}_{\mathbf{K}}$,

$$\exists \delta' \in (0, \delta] \text{ such that } \mathcal{J}\left(\mathbf{x}^\star + \eta \textstyle\sum_{k=1}^{n_a} p_k \bar{\boldsymbol{\xi}}_k\right) \geq \mathcal{J}(\mathbf{x}^\star), \quad \forall 0 \leq \eta \leq \delta',$$

and we get

$$\delta\mathcal{J}\left(\mathbf{x}^\star; \textstyle\sum_{k=1}^{n_a} p_k \bar{\boldsymbol{\xi}}_k\right) = \lim_{\eta \to^+} \frac{\mathcal{J}(\mathbf{x}^\star + \eta \sum_{k=1}^{n_a} p_k \bar{\boldsymbol{\xi}}_k) - \mathcal{J}(\mathbf{x}^\star)}{\eta} \geq 0,$$

which contradicts the inequality (2.29) obtained earlier. □

2.7.3 Problems with End-Point Constraints: Transversal Conditions

So far, we have only considered those problems with either free or fixed end-times t_1, t_2 and end-points $\mathbf{x}(t_1)$, $\mathbf{x}(t_2)$. Yet, many problems of the calculus of variations do not fall into this formulation. As just an example, in the Brachistochrone problem 2.1, the extremity point B could very well be constrained to lie on a curve, rather than a fixed point.

In this subsection, we shall consider problems having end-point constraints of the form $\boldsymbol{\phi}(t_2, \mathbf{x}(t_2)) = \mathbf{0}$, with t_2 being specified or not. Note that this formulation allows to deal with fixed end-point problems as in §2.5.2, e.g., by specifying the end-point constraint as $\boldsymbol{\phi} := \mathbf{x}(t_2) - \mathbf{x}_2$. In the case t_2 is free, t_2 shall be considered as an additional variable in the optimization problem. Like in §2.5.5, we shall then define the functions $\mathbf{x}(t)$ by extension on a "sufficiently" large interval $[t_1, T]$, and consider the linear space $\mathcal{C}^1[t_1, T]^{n_x} \times \mathbb{R}$, supplied with the weak norm $\|(\mathbf{x}, t)\|_{1,\infty} := \|\mathbf{x}\|_{1,\infty} + |t|$. In particular, Theorem 2.47 applies readily by specializing the normed linear space $(\mathcal{X}, \|\cdot\|)$ to $(\mathcal{C}^1[t_1, T]^{n_x} \times \mathbb{R}, \|(\cdot, \cdot)\|_{1,\infty})$, and considering the Gâteaux derivative $\delta\mathcal{J}(\mathbf{x}, t_2; \boldsymbol{\xi}, \tau)$ at $(\mathbf{x}, t_2)$ in the direction $(\boldsymbol{\xi}, \tau)$. These considerations yield necessary conditions of optimality for problems with end-point constraints as given in the following:

Theorem 2.52 (Transversal Conditions). *Consider the problem to minimize the functional*

$$\mathcal{J}(\mathbf{x}, t_2) := \int_{t_1}^{t_2} \ell(t, \mathbf{x}(t), \dot{\mathbf{x}}(t)) \mathrm{d}t,$$

on $\mathcal{D} := \{(\mathbf{x}, t_2) \in \mathcal{C}^1[t_1, T]^{n_x} \times [t_1, T] : \mathbf{x}(t_1) = \mathbf{x}_1, \boldsymbol{\phi}(t_2, \mathbf{x}(t_2)) = 0\}$, *with* $\ell \in \mathcal{C}^1([t_1, T] \times \mathbb{R}^{2 \times n_x})$ *and* $\boldsymbol{\phi} \in \mathcal{C}^1([t_1, T] \times \mathbb{R}^{n_x})^{n_x}$. *Suppose that* $(\mathbf{x}^\star, t_2^\star)$ *gives a (local) minimum for* $\mathcal{J}$ *on* $\mathcal{D}$, *and* $\operatorname{rank}(\boldsymbol{\phi}_t \quad \boldsymbol{\phi}_{\mathbf{x}}) = n_x$ *at* $(\mathbf{x}^\star(t_2^\star), t_2^\star)$. *Then,* $\mathbf{x}^\star$ *is a solution to the Euler equation* (2.12) *satisfying both the end-point constraints* $\mathbf{x}^\star(t_1) = \mathbf{x}_1$ *and the transversal condition*

$$\left[\left(\ell - \dot{\mathbf{x}}^\mathsf{T} \ell_{\dot{\mathbf{x}}}\right) d_t + \ell_{\dot{\mathbf{x}}}{}^\mathsf{T} \mathbf{d}_{\mathbf{x}}\right]_{\mathbf{x}=\mathbf{x}^\star, t=t_2^\star} = 0 \quad \forall (d_t \quad \mathbf{d}_{\mathbf{x}}) \in \ker(\boldsymbol{\phi}_t \quad \boldsymbol{\phi}_{\mathbf{x}}). \tag{2.30}$$

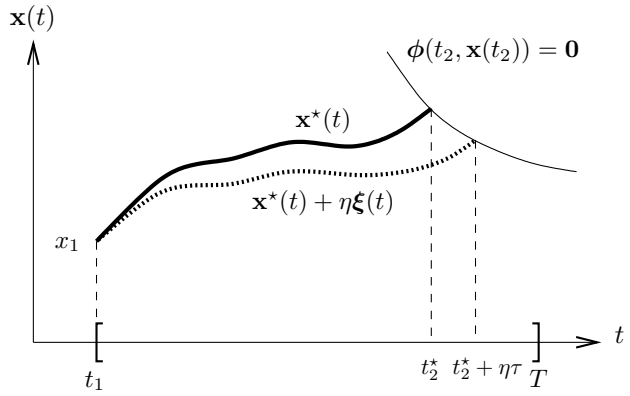


Figure 2.7. Extremal trajectory $x^\star(t)$ in $[t_1, t_2^\star]$, and a neighboring admissible trajectory $x^\star(t) + \eta\boldsymbol{\xi}(t)$ in $[t_1, t_2^\star + \eta\tau]$.

In particular, the transversal condition (2.30) *reduces to*

$$\left[\phi_x \left(\ell - \dot{x}\ell_{\dot{x}}\right) - \phi_t \ell_{\dot{x}}\right]_{x=x^\star, t=t_2^\star} = 0, \tag{2.31}$$

in the scalar case ($n_x = 1$).

Proof. Observe first that by fixing $t_2 := t_2^\star$ and varying $\mathbf{x}^\star$ in the $\mathcal{D}$-admissible direction $\boldsymbol{\xi} \in \mathcal{C}^1[t_1, t_2^\star]^{n_x}$ such that $\boldsymbol{\xi}(t_1) = \boldsymbol{\xi}(t_2^\star) = \mathbf{0}$, we show as in the proof of Theorem 2.14 that $\mathbf{x}^\star$ must be a solution to the Euler equation (2.12) on $[t_1, t_2^\star]$. Observe also that the right end-point constraints may be expressed as the zero-level set of the functionals

$$\mathcal{P}_k := \phi_k(t_2, \mathbf{x}(t_2)) \quad k = 1, \ldots, n_x.$$

Then, using the Euler equation, it is readily obtained that

$$\begin{aligned}
\delta\mathcal{J}(\mathbf{x}^\star, t_2^\star; \boldsymbol{\xi}, \tau) &= \ell_{\dot{\mathbf{x}}} \left[\mathbf{x}^\star(t_2^\star)\right]^\mathsf{T} \boldsymbol{\xi}(t_2^\star) + \ell \left[\mathbf{x}^\star(t_2^\star)\right] \tau \\
\delta\mathcal{P}_k(\mathbf{x}^\star, t_2^\star; \boldsymbol{\xi}, \tau) &= (\phi_k)_t \left[\mathbf{x}^\star(t_2^\star)\right] \tau + (\phi_k)_{\mathbf{x}} \left[\mathbf{x}^\star(t_2^\star)\right]^\mathsf{T} \left(\boldsymbol{\xi}(t_2^\star) + \dot{\mathbf{x}}(t_2^\star)\tau\right),
\end{aligned}$$

where the usual compressed notation is used. Based on the differentiability assumptions on ℓ and $\boldsymbol{\phi}$, it is clear that these Gâteaux derivatives exist and are continuous. Further, since the rank condition $\operatorname{rank}\begin{pmatrix}\boldsymbol{\phi}_t & \boldsymbol{\phi}_{\mathbf{x}}\end{pmatrix} = n_x$ holds at $(\mathbf{x}^\star(t_2^\star), t_2^\star)$, one can always find n_x (independent) directions $(\bar{\boldsymbol{\xi}}_k, \bar{\tau}_k) \in \mathcal{C}^1[t_1, t_2]^{n_x} \times \mathbb{R}$ such that the regularity condition,

$$\begin{vmatrix}
\delta\mathcal{P}_1(\mathbf{x}^\star, t_2^\star; \bar{\boldsymbol{\xi}}_1, \bar{\tau}_1) & \cdots & \delta\mathcal{P}_1(\mathbf{x}^\star, t_2^\star; \bar{\boldsymbol{\xi}}_{n_x}, \bar{\tau}_{n_x}) \\
\vdots & \ddots & \vdots \\
\delta\mathcal{P}_{n_x}(\mathbf{x}^\star, t_2^\star; \bar{\boldsymbol{\xi}}_1, \bar{\tau}_1) & \cdots & \delta\mathcal{P}_{n_x}(\mathbf{x}^\star, t_2^\star; \bar{\boldsymbol{\xi}}_{n_x}, \bar{\tau}_{n_x})
\end{vmatrix} \neq 0,$$

is satisfied.

Now, consider the linear subspace $\Xi := \{(\boldsymbol{\xi}, \tau) \in \mathcal{C}^1[t_1, T]^{n_x} \times \mathbb{R} : \boldsymbol{\xi}(t_1) = \mathbf{0}\}$. Since $(\mathbf{x}^\star, t_2^\star)$ gives a (local) minimum for $\mathcal{J}$ on $\mathcal{D}$, by Theorem 2.47 (and Remark 2.49), there is

a vector $\boldsymbol{\lambda} \in \mathbb{R}^{n_x}$ such that

$$\begin{aligned} 0 &= \delta \left(\mathcal{J} + \textstyle\sum_{k=1}^{n_x} \lambda_k \mathcal{P}_k \right) (\mathbf{x}^\star, t_2^\star; \boldsymbol{\xi}, \tau) \\ &= \Big(\ell \left[\mathbf{x}^\star(t_2^\star) \right] - \dot{\mathbf{x}}^{\mathsf{T}}(t_2^\star) \ell_{\dot{\mathbf{x}}} \left[\mathbf{x}^\star(t_2^\star) \right] + \boldsymbol{\lambda}^{\mathsf{T}} \boldsymbol{\phi}_t \left[\mathbf{x}^\star(t_2^\star) \right] \Big) \tau \\ &\quad + \Big[\ell_{\dot{\mathbf{x}}} \left[\mathbf{x}^\star(t_2^\star) \right] + \boldsymbol{\lambda}^{\mathsf{T}} \boldsymbol{\phi}_{\mathbf{x}} \left[\mathbf{x}^\star(t_2^\star) \right] \Big]^{\mathsf{T}} \left(\boldsymbol{\xi}(t_2^\star) + \dot{\mathbf{x}}(t_2^\star) \tau \right) \end{aligned}$$

for each $(\boldsymbol{\xi}, \tau) \in \Xi$. In particular, restricting attention to those $(\boldsymbol{\xi}, \tau) \in \Xi$ such that $\boldsymbol{\xi}(t_1) = \mathbf{0}$ and $\boldsymbol{\xi}(t_2^\star) = -\dot{\mathbf{x}}(t_2^\star)\tau$, we get

$$0 \;=\; \Big(\ell \left[\mathbf{x}^\star(t_2^\star) \right] - \dot{\mathbf{x}}^{\mathsf{T}}(t_2^\star) \ell_{\dot{\mathbf{x}}} \left[\mathbf{x}^\star(t_2^\star) \right] + \boldsymbol{\lambda}^{\mathsf{T}} \boldsymbol{\phi}_t \left[\mathbf{x}^\star(t_2^\star) \right] \Big) \tau,$$

for every τ sufficiently small. Similarly, considering those variations $(\boldsymbol{\xi}, \tau) \in \Xi$ such that $\boldsymbol{\xi} = (0, \ldots, 0, \xi_i, 0, \ldots, 0)^{\mathsf{T}}$ with $\xi_i(t_1) = 0$ and $\tau = 0$, we obtain

$$0 \;=\; \Big[\ell_{\dot{x}_i} \left[\mathbf{x}^\star(t_2^\star) \right] + \boldsymbol{\lambda}^{\mathsf{T}} \boldsymbol{\phi}_{x_i} \left[\mathbf{x}^\star(t_2^\star) \right] \Big]^{\mathsf{T}} \xi_i(t_2^\star),$$

for every $\xi_i(t_2^\star)$ sufficiently small, and for each $i = 1, \ldots, n_x$. Dividing the last two equations by τ and $\xi_i(t_2^\star)$, respectively, yields the system of equations

$$\boldsymbol{\lambda}^{\mathsf{T}} \begin{pmatrix} \boldsymbol{\phi}_t & \boldsymbol{\phi}_{\mathbf{x}} \end{pmatrix} \;=\; \Big(\ell \left[\mathbf{x}^\star(t_2^\star) \right] - \dot{\mathbf{x}}^{\mathsf{T}}(t_2^\star) \ell_{\dot{\mathbf{x}}} \left[\mathbf{x}^\star(t_2^\star) \right] \quad \ell_{\dot{\mathbf{x}}} \left[\mathbf{x}^\star(t_2^\star) \right] \Big)^{\mathsf{T}}.$$

Finally, since $\operatorname{rank} \begin{pmatrix} \boldsymbol{\phi}_t & \boldsymbol{\phi}_{\mathbf{x}} \end{pmatrix} = n_x$, we have $\dim \ker \begin{pmatrix} \boldsymbol{\phi}_t & \boldsymbol{\phi}_{\mathbf{x}} \end{pmatrix} = 1$, and

$$\Big(\ell \left[\mathbf{x}^\star(t_2^\star) \right] - \dot{\mathbf{x}}^{\mathsf{T}}(t_2^\star) \ell_{\dot{\mathbf{x}}} \left[\mathbf{x}^\star(t_2^\star) \right] \quad \ell_{\dot{\mathbf{x}}} \left[\mathbf{x}^\star(t_2^\star) \right] \Big)^{\mathsf{T}} \mathbf{d} \;= 0,$$

for each $\mathbf{d} \in \ker \begin{pmatrix} \boldsymbol{\phi}_t & \boldsymbol{\phi}_{\mathbf{x}} \end{pmatrix}$. □

Example 2.53 (Minimum Path Problem with Variable End-Point). Consider the problem to minimize the distance between a fixed point $A = (x_A, y_A)$ and the $B = (x_B, y_B) \in \{(x, y) \in \mathbb{R}^2 : y = ax + b\}$, in the (x, y)-plane. We want to find the curve $y(x)$, $x_A \leq x \leq x_B$ such that the functional $\mathcal{J}(y, x_B) := \int_{x_A}^{x_B} \sqrt{1 + \dot{y}(x)^2} \mathrm{d}x$ is minimized, subject to the bound constraints $y(x_A) = y_A$ and $y(x_B) = a x_B + b$. We saw in Example 2.18 that $\dot{\bar{y}}(x)$ must be constant along an extremal trajectory $\bar{y} \in \mathcal{C}^1[x_A, x_B]$, i.e.

$$y(x) \;=\; C_1 x + C_2.$$

Here, the constants of integration C_1 and C_2 must verify the end-point conditions

$$\begin{array}{rcl} y(x_A) & = & y_A \;=\; C_1 x_A + C_2 \\ y(x_B) & = & a\, x_B + b \;=\; C_1 x_B + C_2, \end{array}$$

which yields a system of two equations in the variables C_1, C_2 and x_B. The additional relation is provided by the transversal condition (2.31) as

$$-a\, \dot{y}(x) \;=\; -a\, C_1 \;=\; 1.$$

Note that this latter condition expresses the fact that an extremal $\bar{y}(x)$ must be orthogonal to the boundary end-point constraint $y = ax + b$ at $x = x_B$, in the (x, y)-plane. Finally, provided that $a \neq 0$, we get:

$$\begin{aligned} \bar{y}(x) &= y_A - \frac{1}{a}(x - x_A) \\ \bar{x}_B &= \frac{a(y_A - b) - x_A}{1 + a^2}. \end{aligned}$$

2.7.4 Problems with Isoperimetric Constraints

An *isoperimetric problem* of the calculus of variations is a problem wherein one or more constraints involves the integral of a given functional over part or all of the integration horizon $[t_1, t_2]$. Typically,

$$\min_{\mathbf{x}(t)} \int_{t_1}^{t_2} \ell(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\mathrm{d}t \quad \text{subject to:} \quad \int_{t_1}^{t_2} \psi(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\mathrm{d}t = K.$$

Such isoperimetric constraints arise frequently in geometry problems such as the determination of the curve [resp. surface] enclosing the largest surface [resp. volume] subject to a fixed perimeter [resp. area].

The following theorem provides a characterization of the extremals of an isoperimetric problem, based on the method of Lagrange multipliers introduced earlier in §2.7.1.

Theorem 2.54 (First-Order Necessary Conditions for Isoperimetric Problems). *Consider the problem to minimize the functional*

$$\mathcal{J}(\mathbf{x}) := \int_{t_1}^{t_2} \ell(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\mathrm{d}t,$$

on $\mathcal{D} := \{\mathbf{x} \in \mathcal{C}^1[t_1, t_2]^{n_x} : \mathbf{x}(t_1) = \mathbf{x}_1, \mathbf{x}(t_2) = \mathbf{x}_2\}$*, subject to the isoperimetric constraints*

$$\mathcal{G}_i(\mathbf{x}) := \int_{t_1}^{t_2} \psi_i(t, \mathbf{x}(t), \dot{\mathbf{x}}(t))\mathrm{d}t = K_i, \quad i = 1, \ldots, n_g,$$

with $\ell \in \mathcal{C}^1([t_1, t_2] \times \mathbb{R}^{2\times n_x})$ *and* $\psi_i \in \mathcal{C}^1([t_1, t_2] \times \mathbb{R}^{2\times n_x})$*,* $i = 1, \ldots, n_g$*. Suppose that* $\mathbf{x}^\star \in \mathcal{D}$ *gives a (local) minimum for this problem, and*

$$\begin{vmatrix} \delta\mathcal{G}_1(\mathbf{x}^\star; \bar{\boldsymbol{\xi}}_1) & \cdots & \delta\mathcal{G}_{n_g}(\mathbf{x}^\star; \bar{\boldsymbol{\xi}}_{n_g}) \\ \vdots & \ddots & \vdots \\ \delta\mathcal{G}_1(\mathbf{x}^\star; \bar{\boldsymbol{\xi}}_1) & \cdots & \delta\mathcal{G}_{n_g}(\mathbf{x}^\star; \bar{\boldsymbol{\xi}}_{n_g}) \end{vmatrix} \neq 0,$$

for n_g *(independent) directions* $\bar{\boldsymbol{\xi}}_1, \ldots, \bar{\boldsymbol{\xi}}_{n_g} \in \mathcal{C}^1[t_1, t_2]^{n_x}$*. Then, there exists a vector* $\boldsymbol{\lambda}$ *such that* $\mathbf{x}^\star$ *is a solution to the so called Euler-Lagrange's equation*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}_{\dot{x}_i}(t, \mathbf{x}, \dot{\mathbf{x}}, \boldsymbol{\lambda}) = \mathcal{L}_{x_i}(t, \mathbf{x}, \dot{\mathbf{x}}, \boldsymbol{\lambda}), \quad i = 1, \ldots, n_x, \tag{2.32}$$

where

$$\mathcal{L}(t, \mathbf{x}, \dot{\mathbf{x}}, \boldsymbol{\lambda}) = \ell(t, \mathbf{x}, \dot{\mathbf{x}}) + \boldsymbol{\lambda}^\mathsf{T}\boldsymbol{\psi}(t, \mathbf{x}, \dot{\mathbf{x}}).$$

Proof. Remark first that, from the differentiability assumptions on ℓ and ψ_i, $i = 1, \ldots, n_g$, the Gâteaux derivatives $\delta\mathcal{J}(\mathbf{x}^\star; \boldsymbol{\xi})$ and $\delta\mathcal{G}_i(\mathbf{x}^\star; \boldsymbol{\xi})$, $i = 1, \ldots, n_g$, exist and are continuous for every $\boldsymbol{\xi} \in \mathcal{C}^1[t_1, t_2]^{n_x}$. Since $\mathbf{x}^\star \in \mathcal{D}$ gives a (local) minimum for $\mathcal{J}$ on $\mathcal{D}$ constrained to $\Gamma(\mathbf{K}) := \{\mathbf{x} \in \mathcal{C}^1[t_1, t_2]^{n_x} : \mathcal{G}_i(\mathbf{x}) = K_i, i = 1, \ldots, n_g\}$, and $\mathbf{x}^\star$ is a regular point for the constraints, by Theorem 2.47 (and Remark 2.49), there exists a constant vector $\boldsymbol{\lambda} \in \mathbb{R}^{n_g}$ such that

$$\delta\mathcal{J}(\mathbf{x}^\star; \boldsymbol{\xi}) + [\delta\mathcal{G}_1(\mathbf{x}^\star; \boldsymbol{\xi}) \ldots \delta\mathcal{G}_{n_g}(\mathbf{x}^\star; \boldsymbol{\xi})]\, \boldsymbol{\lambda} \;=\; 0,$$

for each $\mathcal{D}$-admissible direction $\boldsymbol{\xi}$. Observe that this latter condition is equivalent to that of finding a minimizer to the functional

$$\hat{\mathcal{J}}(\mathbf{x}) \;:=\; \int_{t_1}^{t_2} \ell(t, \mathbf{x}, \dot{\mathbf{x}}) + \boldsymbol{\lambda}^\mathsf{T} \boldsymbol{\psi}(t, \mathbf{x}, \dot{\mathbf{x}})\, \mathrm{d}t \;:=\; \int_{t_1}^{t_2} \mathcal{L}(t, \mathbf{x}, \dot{\mathbf{x}}, \boldsymbol{\lambda})\, \mathrm{d}t,$$

on $\mathcal{D}$. The conclusion then directly follows upon applying Theorem 2.14. □

Remark 2.55 (First Integrals). Similar to free problems of the calculus of variations (see Remark 2.17), it can be shown that the Hamiltonian function $\mathcal{H}$ defined as

$$\mathcal{H} \;:=\; \mathcal{L} - \dot{\mathbf{x}}^\mathsf{T} \mathcal{L}_{\dot{\mathbf{x}}},$$

is constant along an extremal trajectory provided that $\mathcal{L}$ does not depend on the independent variable t.

Example 2.56 (Problem of the Surface of Revolution of Minimum Area). Consider the problem to find the smooth curve $x(t) \geq 0$, having a fixed length $\mu > 0$, joining two given points $A = (t_A, x_A)$ and $B = (t_B, x_B)$, and generating a surface of revolution around the t-axis of minimum area (see Fig 2.8.). In mathematical terms, the problem consists of finding a minimizer of the functional

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

on $\mathcal{D} := \{x \in \mathcal{C}^1[t_A, t_B] : x(t_A) = x_A, x(t_B) = x_B\}$, subject to the isoperimetric constraint

$$\Theta(x) := \int_{t_A}^{t_B} \sqrt{1 + \dot{x}(t)^2}\mathrm{d}t \;=\; \mu.$$

Let us drop the coefficient 2π in $\mathcal{J}$, and introduce the Lagrangian $\mathcal{L}$ as

$$\mathcal{L}(x, \dot{x}, \lambda) \;:=\; x(t)\sqrt{1 + \dot{x}(t)^2} + \lambda\sqrt{1 + \dot{x}(t)^2}.$$

Since $\mathcal{L}$ does not depend on the independent variable t, we use the fact that $\mathcal{H}$ must be constant along an extremal trajectory $\bar{x}(t)$,

$$\mathcal{H} \;:=\; \mathcal{L} - \dot{x}^\mathsf{T} \mathcal{L}_{\dot{x}} \;=\; C_1,$$

for some constant $C_1 \in \mathbb{R}$. That is,

$$\bar{x}(t) + \lambda \;=\; C\sqrt{1 + \dot{\bar{x}}(t)^2}.$$

Let $\bar{y}$ be defined as $\dot{\bar{x}} = \sinh \bar{y}$. Then, $\bar{x} = C_1 \cosh \bar{y} - \lambda$ and $\mathrm{d}\bar{x} = C_1 \sinh \bar{y}\, \mathrm{d}\bar{y}$. That is, $\mathrm{d}t = (\sinh \bar{y})^{-1} \mathrm{d}\bar{x} = C_1 \mathrm{d}y$, from which we get $t = C_1 \bar{y} + C_2$, with $C_2 \in \mathbb{R}$ another constant of integration. Finally,

$$\bar{x}(t) = C_1 \cosh\left(\frac{t - C_2}{C_1}\right) - \lambda.$$

We obtain the equation of a family of *catenaries*, where the constants C_1, C_2 and the Lagrange multiplier λ are to be determined from the boundary conditions $\bar{x}(t_A) = x_A$ and $\bar{x}(t_B) = x_B$, as well as the isoperimetric constraint $\Theta(\bar{x}) = \mu$.

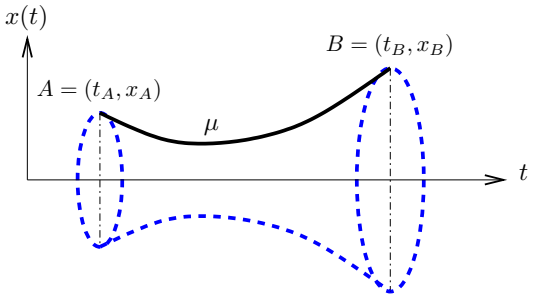


Figure 2.8. Problem of the surface of revolution of minimum area.

2.8 NOTES AND REFERENCES

The material presented in this chapter is mostly a summary of the material in Chapters 2 through 7 of Troutman's book [55]. The books by Kamien and Schwartz [28] and Culioli [17] have also been very useful in writing this chapter.

Note that sufficient conditions which do not rely on the joint-convexity of the Lagrangian function have not been presented herein. This is the case in particular for Jacobi's sufficient condition which uses the concept of *conjugacy*. However, this advanced material falls out of the scope of this textbook. We refer the interested reader to the books by Troutman [55, Chapter 9] and Cesari [16, Chapter 2], for a thorough description, proof and discussion of these additional optimality conditions.

It should also be noted that problems having Lagrangian constraints have not been addressed in this chapter. This discussion is indeed deferred until the following chapter on optimal control.

Finally, we note that a number of results exist regarding the existence of a solution to problems of the calculus of variations, such as Tonelli's existence theorem. Again, the interested (and somewhat courageous) reader is referred to [16] for details.

Appendix: Technical Lemmas

Lemma 2.A.57 (duBois-Reymond's Lemma). *Let D be a subset of $\mathbb{R}$ containing $[t_1, t_2]$, $t_1 < t_2$. Let also h be a continuous function on D, such that*

$$\int_{t_1}^{t_2} h(t)\dot{\xi}(t)\mathrm{d}t = 0,$$

for every continuously differentiable function ξ such that $\xi(t_1) = \xi(t_2) = 0$. Then, h is constant over $[t_1, t_2]$.

Proof. For every real constant C and every function ξ as in the lemma, we have $\int_{t_1}^{t_2} C\dot{\xi}(t)\mathrm{d}t = C\left[\xi(t_2) - \xi(t_1)\right] = 0$. That is,

$$\int_{t_1}^{t_2} \left[h(t) - C\right] \dot{\xi}(t)\mathrm{d}t = 0.$$

Let

$$\tilde{\xi}(t) := \int_{t_1}^{t} \left[h(\tau) - \tilde{C}\right] \mathrm{d}\tau \quad \text{and} \quad \tilde{C} := \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} h(t)\mathrm{d}t.$$

By construction, $\tilde{\xi}$ is continuously differentiable on $[t_1, t_2]$ and satisfies $\tilde{\xi}(t_1) = \tilde{\xi}(t_2) = 0$. Therefore,

$$\int_{t_1}^{t_2} \left[h(t) - \tilde{C}\right] \dot{\tilde{\xi}}(t)\mathrm{d}t = \int_{t_1}^{t_2} \left[h(t) - \tilde{C}\right]^2 \mathrm{d}t = 0,$$

hence, $h(t) = \tilde{C}$ in $[t_1, t_2]$. □

Lemma 2.A.58. *Let $\mathsf{P}(t)$ and $\mathsf{Q}(t)$ be given continuous $(n_x \times n_x)$ symmetric matrix functions on $[t_1, t_2]$, and let the quadratic functional*

$$\int_{t_1}^{t_2} \left[\dot{\boldsymbol{\xi}}(t)^{\mathsf{T}} \mathsf{P}(t) \dot{\boldsymbol{\xi}}(t) + \boldsymbol{\xi}(t)^{\mathsf{T}} \mathsf{Q}(t) \boldsymbol{\xi}(t)\right] \mathrm{d}t, \tag{2.A.1}$$

be defined for all $\boldsymbol{\xi} \in \mathcal{C}^1[t_1, t_2]^{n_x}$ such that $\boldsymbol{\xi}(t_1) = \boldsymbol{\xi}(t_2) = \mathbf{0}$. Then, a necessary condition for (2.A.1) *to be nonnegative for all such $\boldsymbol{\xi}$ is that $\mathsf{P}(t)$ be semi-definite positive for each $t_1 \le t \le t_2$.*

Proof. See. e.g., [22, §29.1, Theorem 1]. □

Theorem 2.A.59 (Differentiation Under the Integral Sign). *Let $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, such that $\ell := \ell(t, \eta)$, be a continuous function with continuous partial derivative ℓ_η on $[t_1, t_2] \times [\eta_1, \eta_2]$. Then,*

$$f(\eta) := \int_{t_1}^{t_2} \ell(t, \eta)\, \mathrm{d}t$$

is in $\mathcal{C}^1[\eta_1, \eta_2]$, with the derivatives

$$\frac{\mathrm{d}}{\mathrm{d}\eta} f(\eta) = \frac{\mathrm{d}}{\mathrm{d}\eta} \int_{t_1}^{t_2} \ell(t, \eta)\, \mathrm{d}t = \int_{t_1}^{t_2} \ell_\eta(t, \eta)\, \mathrm{d}t.$$

Proof. See. e.g., [55, Theorem A.13]. □

Theorem 2.A.60 (Leibniz). *Let $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, such that $\ell := \ell(t, \eta)$, be a continuous function with continuous partial derivative ℓ_η on $[t_1, t_2] \times [\eta_1, \eta_2]$. Let also $h : \mathbb{R} \to \mathbb{R}$ be continuously differentiable on $[\eta_1, \eta_2]$ with range in $[t_1, t_2]$,*

$$h(\eta) \in [t_1, t_2], \quad \forall \eta \in [\eta_1, \eta_2].$$

Then,

$$\frac{\mathrm{d}}{\mathrm{d}\eta} \int_{t_1}^{h(\eta)} \ell(t, \eta)\, \mathrm{d}t = \int_{t_1}^{h(\eta)} \ell_\eta(t, \eta)\, \mathrm{d}t + \ell(h(\eta), \eta) \frac{\mathrm{d}}{\mathrm{d}\eta} h(\eta).$$

Proof. See. e.g., [55, Theorem A.14]. □

Lemma 2.A.61 (Inverse Function Theorem). *Let* $\mathbf{x}_0 \in \mathbb{R}^{n_x}$ *and* $\eta > 0$. *If a function* $\mathbf{\Phi} : \mathcal{B}_\eta(\mathbf{x}_0) \to \mathbb{R}^{n_x}$ *has continuous first partial derivatives in each component with non-vanishing Jacobian determinant at* $\mathbf{x}_0$, *then* $\mathbf{\Phi}$ *provides a continuously invertible mapping between* $\mathcal{B}_\eta(\mathbf{x}_0)$ *and a region containing a full neighborhood of* $\mathbf{\Phi}(\mathbf{x}_0)$.

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.

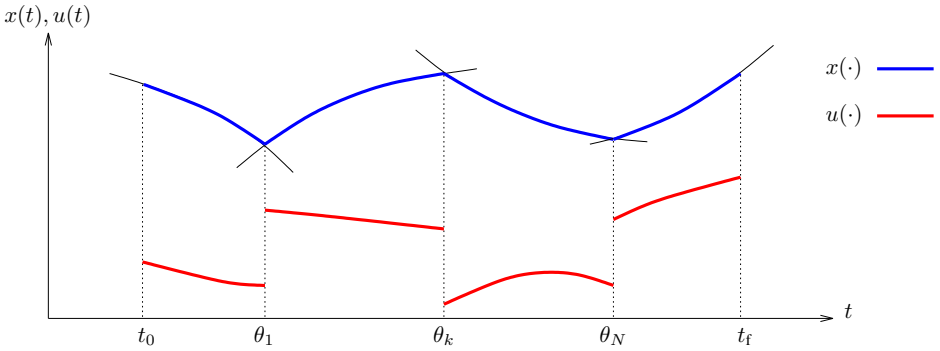


Figure 3.1. Illustration of a piecewise continuous control $u \in \hat{\mathcal{C}}[t_0, t_f]$ (red line), and the corresponding piecewise continuously differentiable response $x \in \hat{\mathcal{C}}^1[t_0, t_f]$ (blue line).

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:

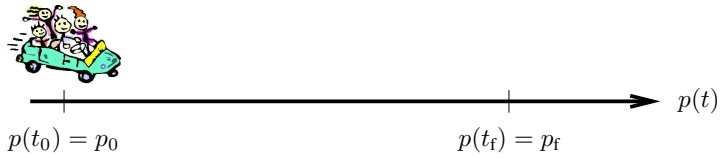


Figure 3.2. A simple control problem.

"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,

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

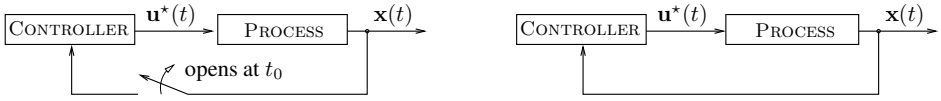
Obtaining a so-called open-loop optimal control law is much easier from a practical viewpoint:

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 $\eta = 0$.

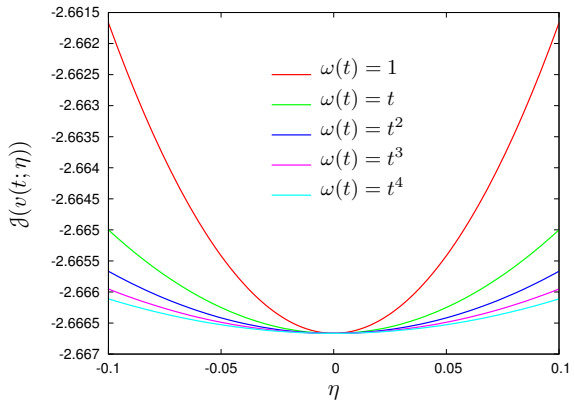


Figure 3.4. Function $\mathcal{J}(v(t;\eta))$ for various perturbations $\omega(t) = t^k$, $k = 0, \ldots, 4$, in Example 3.10.

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.

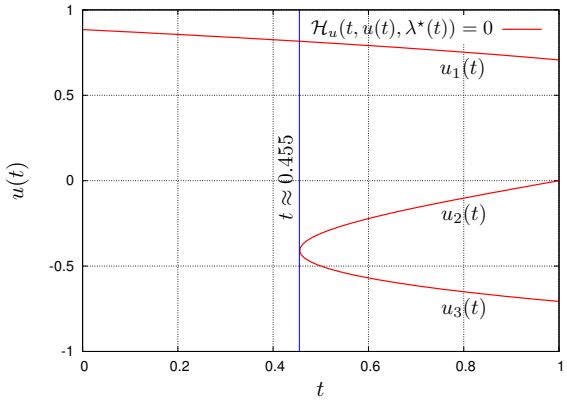


Figure 3.5. Values of the control variable $u(t)$, $0 \le t \le 1$, satisfying the Euler-Lagrange equations in Example 3.16.

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

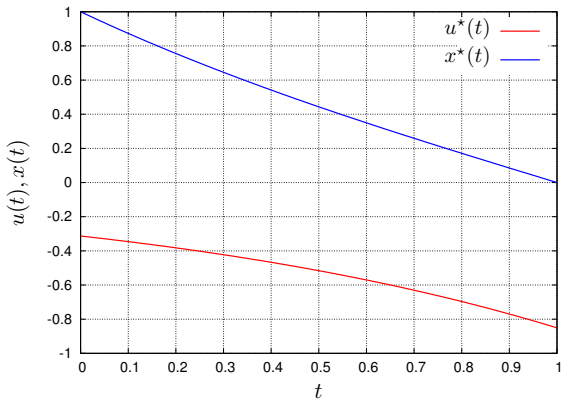


Figure 3.6. Optimal trajectories $u^\star(t)$ and $x^\star(t)$, $0 \le t \le 1$, in Example 3.22.

$\psi_k(t_\mathrm{f}, \mathbf{x}(t_\mathrm{f})) \le 0$ (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:

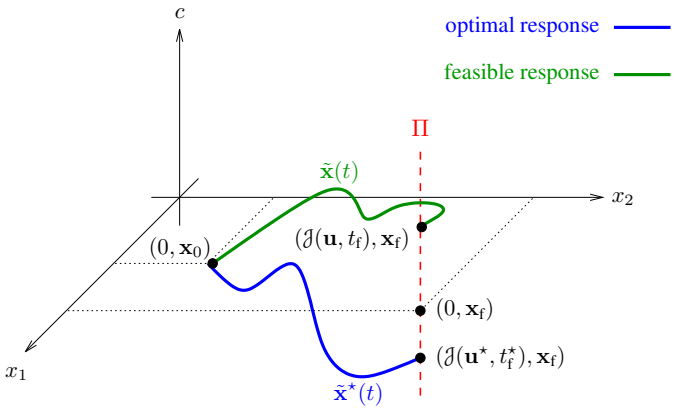


Figure 3.7. Geometric representation of Problem 3.25. (The c-axis is vertical for clarity.)

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.

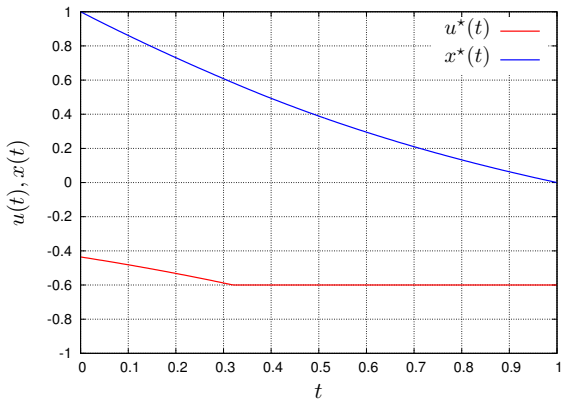


Figure 3.8. Optimal trajectories $u^\star(t)$ and $x^\star(t)$, $0 \le t \le 1$, in Example 3.28.

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.

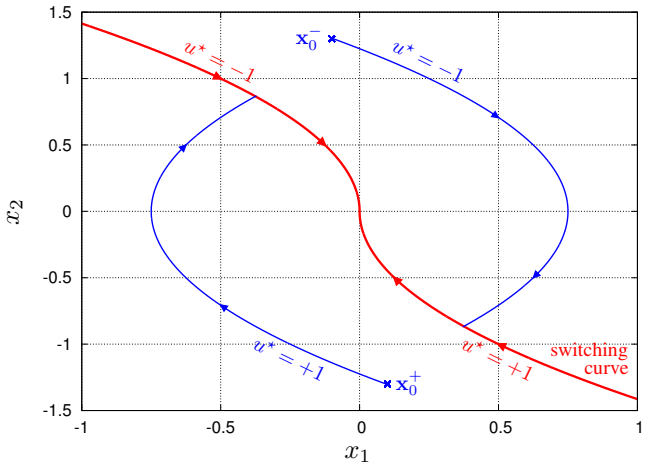


Figure 3.9. Switching curve and typical optimal responses for Example 3.31 — Red line: switching curve; blue line: typical path.

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.

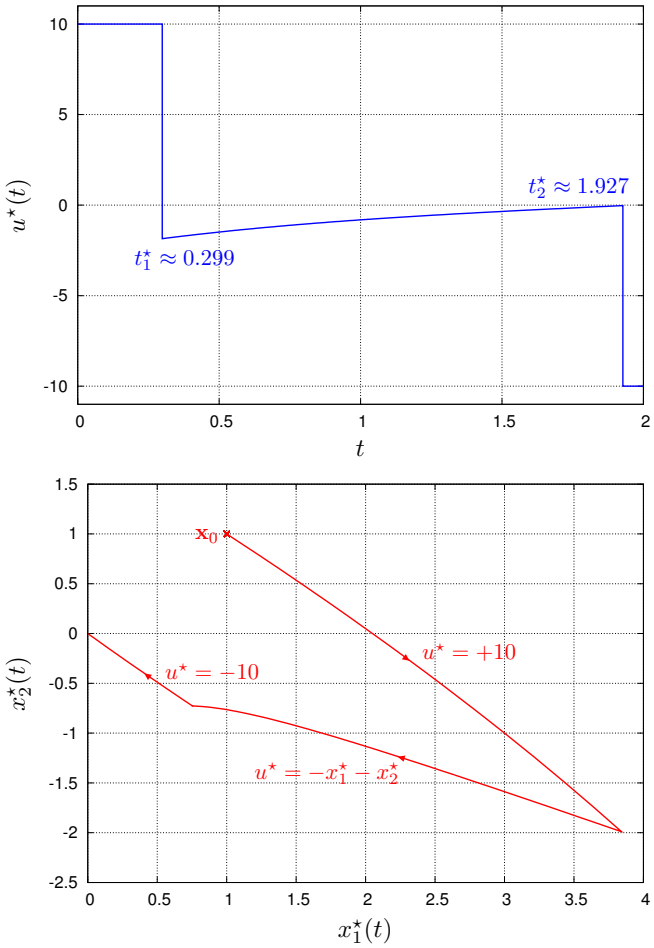


Figure 3.10. Optimal control and response for Example 3.31 – left plot: optimal control vs. time; right plot: optimal response in the phase space.

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.

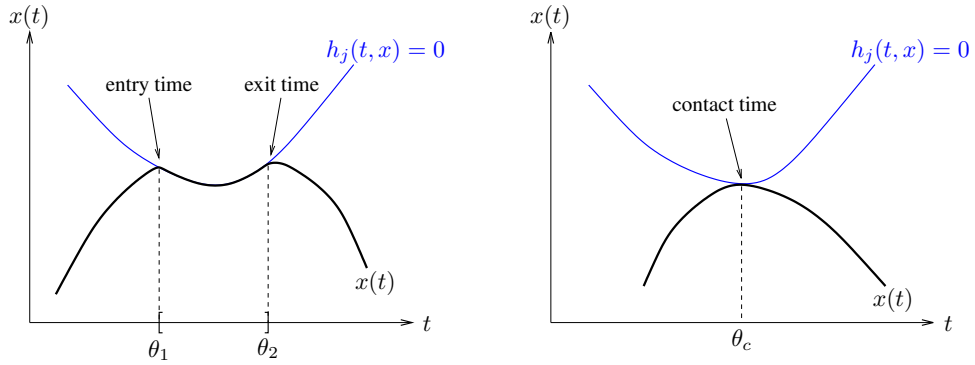


Figure 3.11. Junction types for optimal control problems with pure state inequality constraints.

[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 $\varrho < 0$.

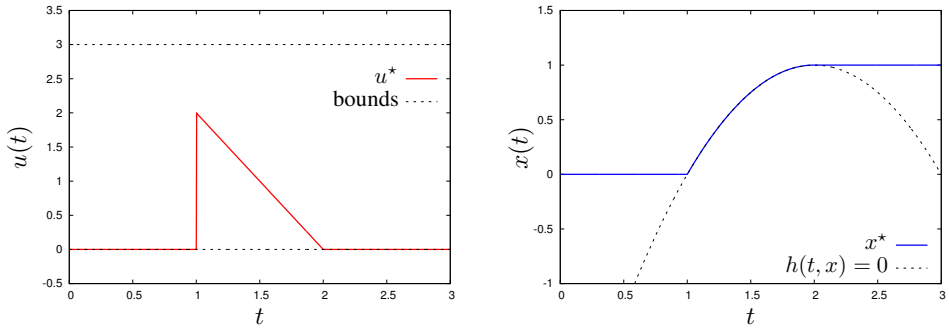


Figure 3.12. Optimal control and response for Example 3.38.

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.

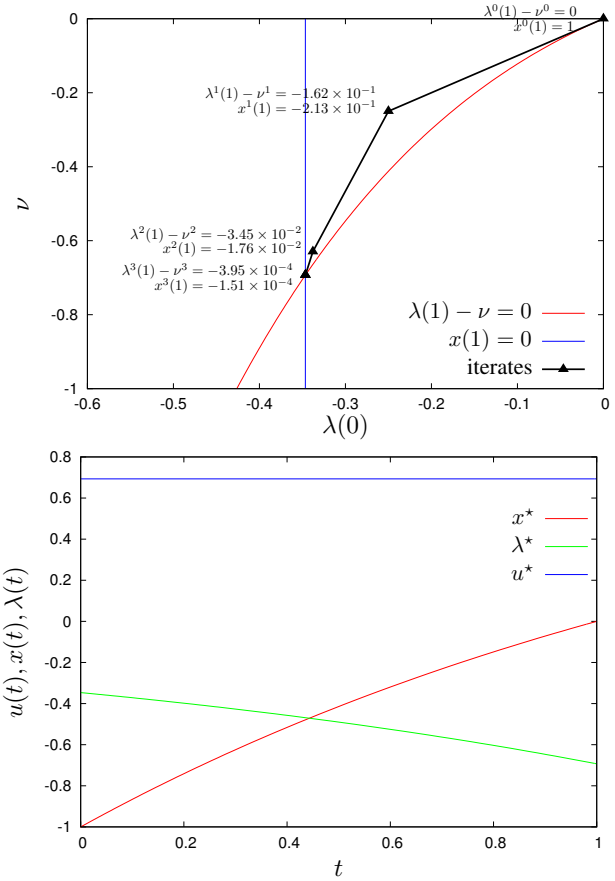


Figure 3.13. Illustration of the indirect shooting method in Example 3.42. Top plot: Newton iterates; bottom plot: Optimal trajectories

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.

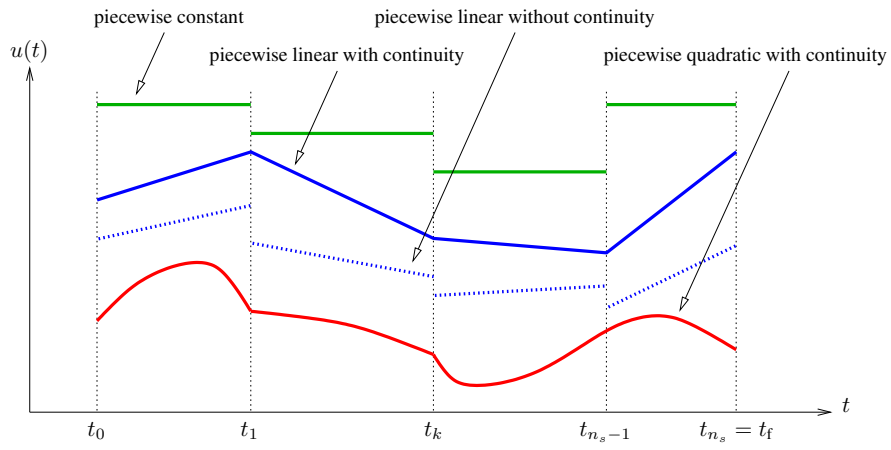


Figure 3.14. Examples of control variable profiles.

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.

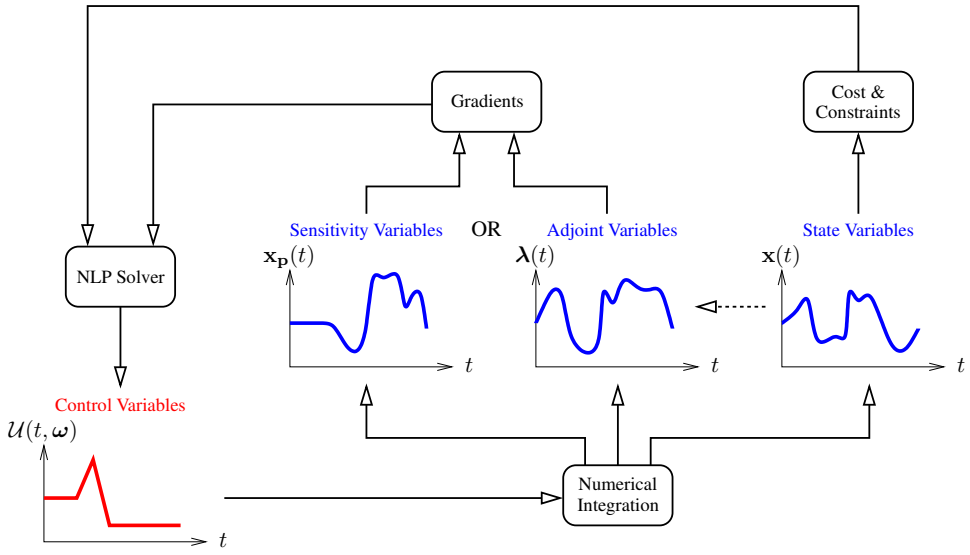


Figure 3.15. Direct sequential methods.

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

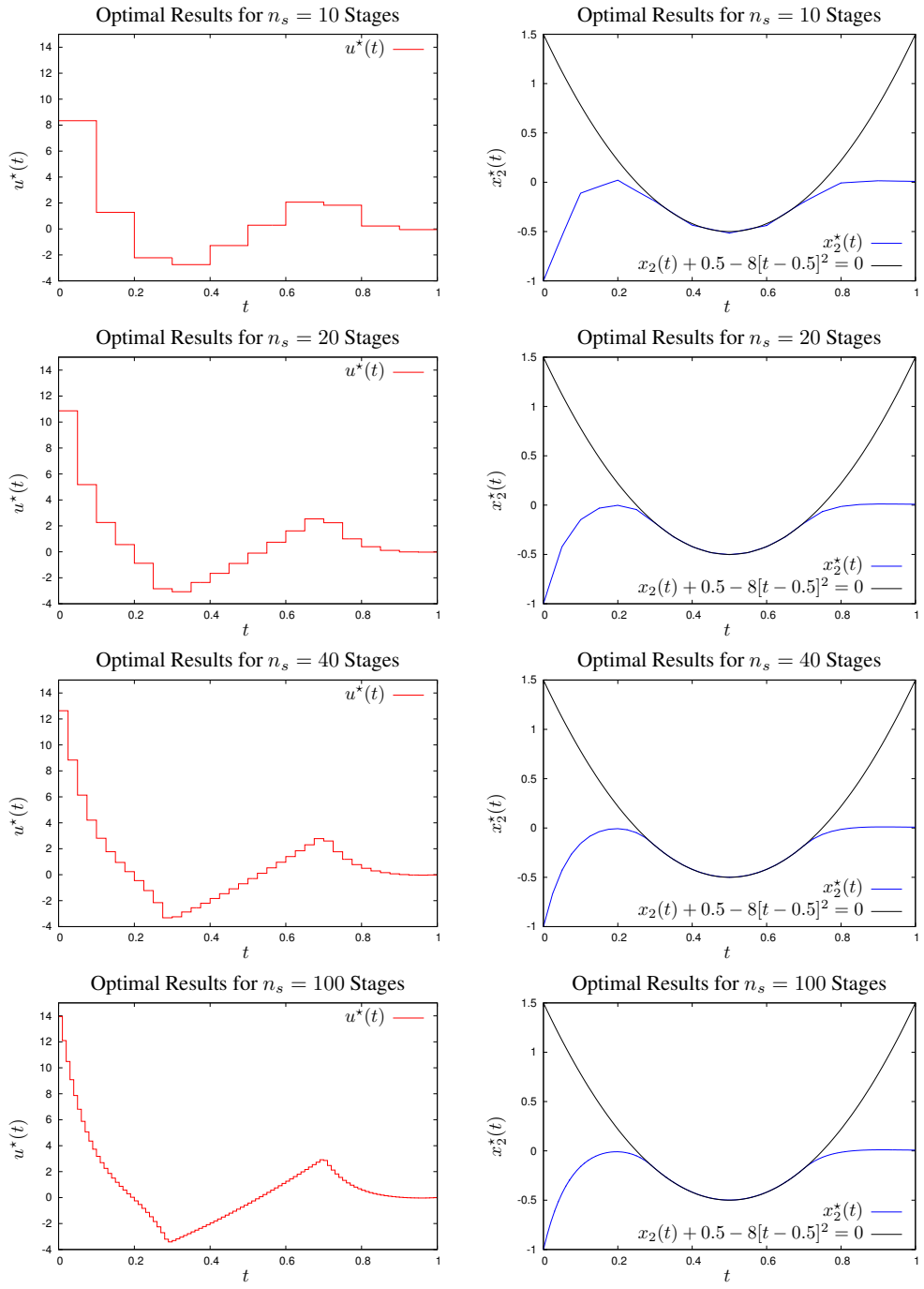


Figure 3.16. Results of the direct sequential approach as applied to Example 3.44 with $\varrho = 5 \times 10^{-3}$ (case A), for $n_s = 10$, 20, 40, and 100. Left plots: optimal piecewise control $u^\star$; right plots: optimal response ${x_2}^\star$

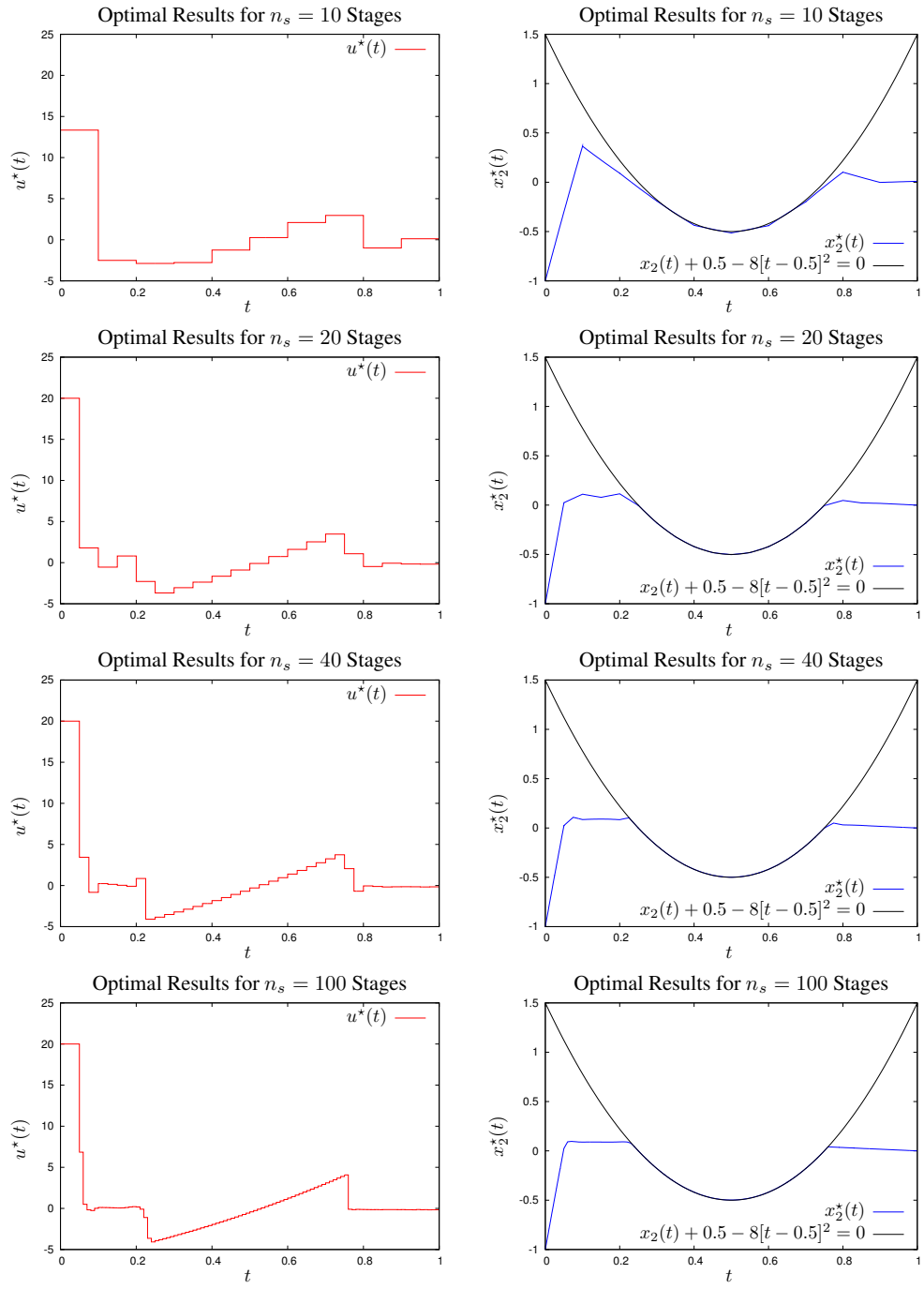


Figure 3.17. Results of the direct sequential approach as applied to Example 3.44 with $\varrho = 0$ (case B), for $n_s = 10$, 20, 40, and 100. Left plots: optimal piecewise control $u^\star$; right plots: optimal response $x_2{}^\star$

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.

APPENDIX A

A.1 NOTATIONS

The following notations are used throughout the textbook. Scalars are denoted by lowercase Roman or Greek letters, e.g., k, α, and μ. $\mathbb{R}^n$ denotes the n-dimensional *real Euclidean space*, composed of all real vectors of dimension n; such vectors are denoted by boldface lowercase Roman or Greek letters, e.g., $\mathbf{x}$, $\mathbf{y}$, and $\boldsymbol{\nu}$. All vectors are column vectors unless stated otherwise. Row vectors are the transpose of column vectors, e.g., $\mathbf{x}^\mathsf{T}$ denotes the row vector $(x_1, \ldots, x_n)$. Matrices are denoted by san serif capital Roman or boldface capital Greek letters, e.g., A, B, and $\boldsymbol{\Psi}$.

$\mathcal{C}([a,b])$ [resp. $\mathcal{C}^k([a,b])$] stands for the set of real-valued, continuous [resp. k times continuously differentiable] functions on the interval $[a,b]$; such functions are denoted by lowercase Roman or Greek letters, e.g., f, g, and φ. $\mathcal{C}([a,b])^n$ [resp. $\mathcal{C}^k([a,b])^n$] stands for the set of vector-valued, continuous [resp. k times continuously differentiable] functions on the interval $[a,b]$; vector-valued functions are denoted by boldface lowercase Roman letters, e.g., $\mathbf{h}$ and $\boldsymbol{\psi}$.

Finally, the following abbreviations are used in this textbook:

$:=$	*"defined as..."*
$\in$	*"is an element of..."* or *"is in..."*
$\notin$	*"is not an element of..."*
$\exists$	*"there exists..."*
$\forall$	*"for each..."* or *"for every..."*
$\square$	*"end of proof."*

A.2 ELEMENTARY CONCEPTS FROM REAL ANALYSIS

We recall some elementary concepts from real analysis [50].

Definition A.1 (Open Ball). *The* open ball *of radius ε centered at $\bar{\mathbf{x}}$ is defined to be the set*

$$\mathcal{B}_\varepsilon(\bar{\mathbf{x}}) := \{\mathbf{x} \in \mathbb{R}^n : \|\bar{\mathbf{x}} - \mathbf{x}\| < \varepsilon\},$$

in any norm $\|\cdot\|$. The corresponding deleted open ball *is defined by*

$$\dot{\mathcal{B}}_\varepsilon(\bar{\mathbf{x}}) := \mathcal{B}_\varepsilon(\bar{\mathbf{x}}) \setminus \{\bar{\mathbf{x}}\}.$$

Definition A.2 (Interior Point, Openness, Limit Point, Closedness, Boundedness, Compactness, Boundary Point, Closure). *Let D be a set in $\mathbb{R}^n$, $n \geq 1$.*

Interior Point *A point $\mathbf{x} \in \mathbb{R}^n$ is said to be an* interior point *of D if there is an open ball $\mathcal{B}_\varepsilon(\mathbf{x})$ such that $\mathcal{B}_\varepsilon(\mathbf{x}) \subset D$. The* interior *of a set D, denoted* $\operatorname{int}(D)$*, is the set of interior points of D. A point $\mathbf{x} \in \mathbb{R}^n$ is said to be an* exterior point *of D if it is an interior point of $\mathbb{R}^n \setminus D$.*

Openness *D is said to be* open *if every point of D is an interior point of D. Obviously, if D is open then* $\operatorname{int}(D) = D$.

Limit Point *A point $\bar{\mathbf{x}} \in \mathbb{R}^n$ is said to be a* limit point *of the set D if every open ball $\mathcal{B}_\varepsilon(\bar{\mathbf{x}})$ contains a point $\mathbf{x} \neq \bar{\mathbf{x}}$ such that $\mathbf{x} \in D$. Note in particular that $\bar{\mathbf{x}}$ does not necessarily have to be an element of D to be a limit point of D.*

Closedness *D is said to be* closed *if every limit point of D is an element of D. Note that there do exist sets that are both open and closed, as well as sets that are neither closed nor open.*

Boundedness *D is said to be* bounded *if there is a real number M such that*

$$\|\mathbf{x}\| \leq M \quad \forall \mathbf{x} \in D$$

in any norm.

Compactness *In $\mathbb{R}^n$, D is said to be* compact *if it is both closed and bounded.*

Boundary Point *A point $\bar{\mathbf{x}} \in \mathbb{R}^n$ is said to be a* boundary point *of the set D if every neighborhood of $\bar{\mathbf{x}}$ contains points both inside and outside of D. The set of boundary point of D is denoted by ∂D.*

Closure *The* closure *of D is the set* $\operatorname{cl}(D) := D \cup L$ *where L denotes the set of all limit points of D.*

A.3 CONVEX ANALYSIS

This subsection summarizes a number of important definitions and results related to convex sets (§ A.3.1) and convex functions (§ A.3.2 and A.3.3). Indeed, the notions of convex sets and convex functions play a crucial role in the theory of optimization. In particular, strong theoretical results are available for convex optimization problems (see, e.g., § 1.3).

A.3.1 Convex Sets

Definition A.3 (Convex Set). *A set $C \subset \mathbb{R}^n$ is said to be* convex *if for every points $\mathbf{x}, \mathbf{y} \in C$, the points*

$$\mathbf{z} := \lambda\mathbf{x} + (1-\lambda)\mathbf{y} \quad \forall\lambda \in [0,1],$$

are also in the set C.

It is important to note that for a set C to be convex, (A.3) must hold for **all** pairs of points in the set C. Geometrically, for C to be convex every point on the line segment connecting any two points in C must also be in C. Fig. A.1. show an example of a convex set. Note that the line segment joining the points $\mathbf{x}$ and $\mathbf{y}$ lies completely inside C, and this is true for all pairs of points in C. On the other hand, Fig. A.2. shows an example of a *nonconvex* set (i.e., a set that is not convex). Observe that not all points on the line segment connecting $\mathbf{x}$ and $\mathbf{y}$ lie in the set D, immediately indicating that D in nonconvex.

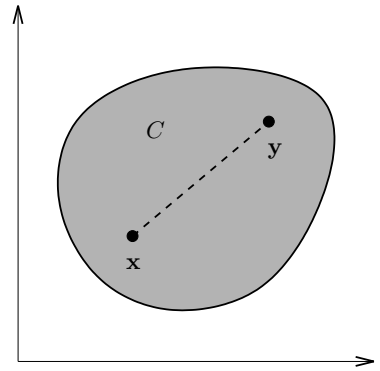


Figure A.1. A convex set.

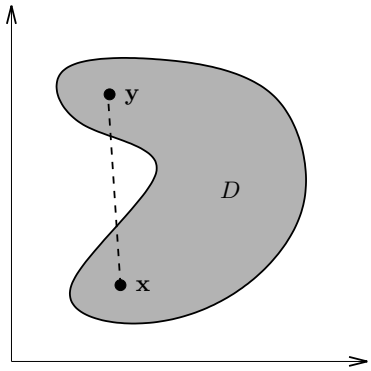


Figure A.2. A nonconvex set.

Example A.4. The set $C := \{\mathbf{x} \in \mathbb{R}^2 : x_1 \geq 0, x_2 \geq 0\}$ is convex.

Lemma A.5 (Intersection of Convex Sets). *Let C_1 and C_2 be convex sets in $\mathbb{R}^n$. Then, $C_1 \cap C_2$ is convex, i.e., the intersection of two convex sets in a convex set.*

Proof. The proof is left to the reader as an exercise. □

Remark A.6. Note that by Definition A.3, an empty set is convex (this is because no counterexample can be found to show that it is nonconvex). That is, Lemma A.5 holds even if C_1 and C_2 do not share any common elements. Note also that Lemma A.5 can be readily extended, by induction, to the intersection of any family of convex sets.

Definition A.7 (Hyperplane, Halfspace). *Let $\mathbf{a} \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Then, $H := \{\mathbf{x} : \mathbf{a}^\mathsf{T}\mathbf{x} = c\}$ is said to be a* hyperplane *and $H^+ := \{\mathbf{x} : \mathbf{a}^\mathsf{T}\mathbf{x} \geq c\}$ is said to be a* halfspace.

Theorem A.8 (Separation of a Convex Set and a Point). *Let C be a nonempty, convex set in $\mathbb{R}^n$ and let $\mathbf{y} \notin C$. Then, there exists a nonzero vector $\mathbf{a} \in \mathbb{R}^n$ and a scalar $c \in \mathbb{R}$ such that:*

$$\mathbf{a}^\mathsf{T}\mathbf{y} > c \quad \textit{and} \quad \mathbf{a}^\mathsf{T}\mathbf{x} \leq c \quad \forall\mathbf{x} \in C.$$

Proof. See, e.g., [6, Theorem 2.4.4] for a proof. □

In fact, $\mathbf{a}^\mathsf{T}\mathbf{y} = c$ defines a separating hyperplane, as illustrated in Fig. A.3. below.

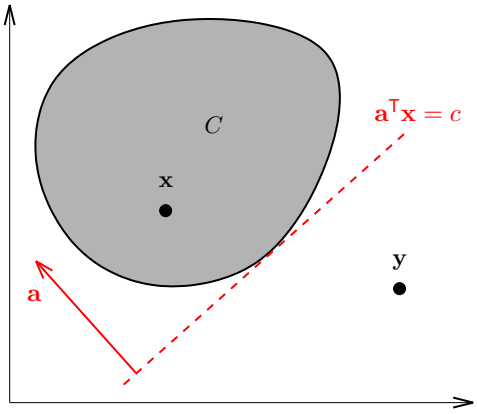


Figure A.3. Illustration of the Separation Theorem.

Theorem A.9 (Separation of Two Convex Sets). *Let C_1 and C_2 be two nonempty, convex set in $\mathbb{R}^n$ and suppose that $C_1 \cap C_2 = \emptyset$. Then, there exists a hyperplane that separates C_1 and C_2; that is, there exists a nonzero vector $\mathbf{p} \in \mathbb{R}^n$ such that*

$$\mathbf{p}^\mathsf{T}\mathbf{x_1} \geq \mathbf{p}^\mathsf{T}\mathbf{x_2} \quad \forall \mathbf{x_1} \in \mathrm{cl}\,(C_1)\,, \ \forall \mathbf{x_2} \in \mathrm{cl}\,(C_2)\,.$$

Proof. See, e.g., [6, Theorem 2.4.8] for a proof. □

Definition A.10 (Cone, Convex Cone). *A nonempty set $C \subset \mathbb{R}^n$ is said to be a* cone *if for every point $\mathbf{x} \in C$,*

$$\alpha\mathbf{x} \in C \quad \forall \alpha \geq 0.$$

If, in addition, C is convex then it is said to be a convex cone.

A.3.2 Convex and Concave Functions

Definition A.11 (Convex Function, Strictly Convex Function). *A function $f : C \to \mathbb{R}$ defined on a convex set $C \in \mathbb{R}^n$ is said to be* convex *if*

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}), \tag{A.1}$$

for each $\mathbf{x}, \mathbf{y} \in C$ and each $\lambda \in (0,1)$; that is, the value of the function on the line segment connecting any two points in the convex set C lies below the line segment in $C \times \mathbb{R}$ connecting the value of the function at the same two points in C. Moreover, f is said to be strictly convex *if*

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}), \tag{A.2}$$

for each $\mathbf{x}, \mathbf{y} \in C$ and each $\lambda \in (0,1)$.

The left plot in Fig. A.4. illustrates the definition: the line segment connecting the values of the function at any two points $\mathbf{x}$ and $\mathbf{y}$ in C lies above the function between $\mathbf{x}$ and $\mathbf{y}$. It

should be noted that this alone does not establish that the function is convex on C; the set C itself should be a convex set. A function that is not convex is said to be *nonconvex*. the right plot in Fig. A.4. shows an example of a nonconvex function on the set C. Note that the dotted portion of the line segment connecting the values of the function at $\mathbf{x}$ and $\mathbf{y}$ lies below the function. Yet, this function is convex on the set C'.

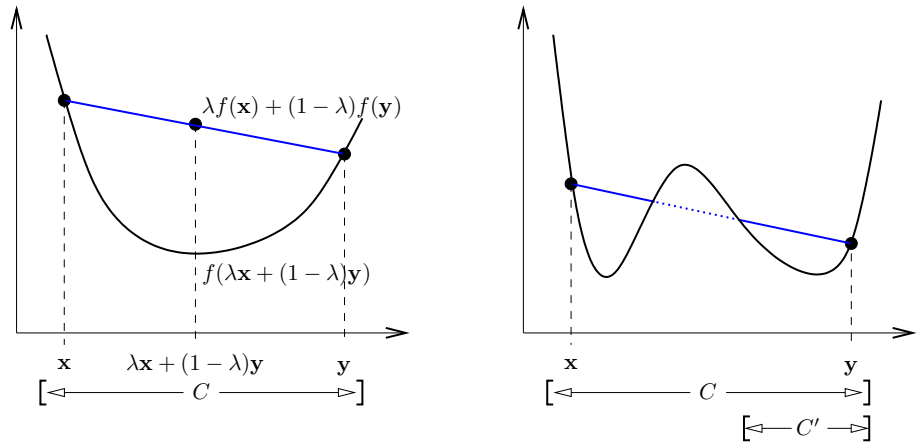


Figure A.4. Illustration of a convex function on C (left plot) and a nonconvex function on C (right plot).

Example A.12. The function $f(x) = |x|$ is convex on $\mathbb{R}$.

Definition A.13 (Concave Function, Strictly Concave Function). *A function $g : C \to \mathbb{R}$ defined on a convex set $C \in \mathbb{R}^n$ is said to be* concave *if the function $f := -g$ is convex on C. The function g is said to be* strictly concave *on C if $-g$ is strictly convex on C.*

Often, it is required that only those $\mathbf{x} \in \mathbb{R}^n$ with $g_i(\mathbf{x}) \leq 0$, $i = 1, \ldots, m$, are feasible points of an optimization problem (i.e., a finite number of inequality constraints are imposed – see, e.g., Chapter 1).

Theorem A.14. *Let C be a convex set in $\mathbb{R}^n$ and let $f : C \to \mathbb{R}$ be a convex function. Then, the level set $C_\alpha := \{\mathbf{x} \in C : f(\mathbf{x}) \leq \alpha\}$, where α is a real number, is a convex set.*

Proof. Let $\mathbf{x}_1, \mathbf{x}_2 \in C_\alpha$. Clearly, $\mathbf{x}_1, \mathbf{x}_2 \in C$, and $f(\mathbf{x}_1) \leq \alpha$ and $f(\mathbf{x}_2) \leq \alpha$. Let $\lambda \in (0, 1)$ and $\mathbf{x} = \lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$. By convexity of C, $\mathbf{x} \in C$. Moreover, by convexity of f on C,

$$f(\mathbf{x}) \leq \lambda f(\mathbf{x}_1) + (1-\lambda) f(\mathbf{x}_2) \leq \lambda\alpha + (1-\lambda)\alpha = \alpha,$$

i.e., $\mathbf{x} \in C_\alpha$. □

Corollary A.15. *Let C be a convex set in $\mathbb{R}^n$ and let $g_i : C \to \mathbb{R}$, $i = 1, \ldots, m$, be convex functions on C. Then, the set defined by*

$$F := \{\mathbf{x} \in C : g_i(\mathbf{x}) \leq 0, \forall i = 1, \ldots, n\}$$

is convex.

Proof. The result is immediately evident from Theorem A.14 and Lemma A.5. □

It is not uncommon that the feasible set in an optimization problems be also defined in terms of equality constraints. Imposing an equality constraint such as $h(\mathbf{x}) = 0$ is obviously equivalent to imposing the pair of inequality constraints $h(\mathbf{x}) \leq 0$ and $-h(\mathbf{x}) \leq 0$. In particular, an affine equality constraint, $\mathbf{a}^\mathsf{T}\mathbf{x} = b$, defines a convex feasible set, for it is both convex and concave for the pair of inequality constraints. With a few trivial exceptions, **most nonlinear equality constraints define a nonconvex feasible set**.

A.3.3 How to Detect Convexity?

In an optimization problem, convexity of the objective function and constraints is crucial, because convex programs posses nicer theoretical properties and can be more efficiently solved numerically than general nonconvex programs. Henceforth, it is important to know whether a given function is convex or not.

Proposition A.16 (Operations that Preserve Convexity of Functions).

- Stability under Nonnegative Weighted Sums. *Let C be a convex set in $\mathbb{R}^n$. If $\mathbf{f} : C \to \mathbb{R}^m$ and $\mathbf{g} : C \to \mathbb{R}^m$ are convex on C, then their linear combination $\lambda\mathbf{f} + \mu\mathbf{g}$, with* nonnegative *coefficients λ and μ, is also convex on C.*
- Stability under Composition with an Affine Mapping. *Let C_1 and C_2 be convex sets in $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively. If $g : C_1 \to \mathbb{R}$ is a convex function on C_1, and $\mathbf{h} : C_2 \to \mathbb{R}^m$ is an affine mapping (i.e., $\mathbf{h}(\mathbf{x}) := \mathbf{A}(\mathbf{x}) + \mathbf{b}$) with* range$(\mathbf{h}) \subset C_1$, *then the composite function $f : C_2 \to \mathbb{R}$ defined as $f(\mathbf{x}) := g[\mathbf{h}(\mathbf{x})]$ is convex on C_2.*
- Stability under (Scalar) Composition with a Nondecreasing Convex Function. *Let C_1 and C_2 be convex sets in $\mathbb{R}$ and $\mathbb{R}^n$, respectively. If $g : C_1 \to \mathbb{R}$ is a nondecreasing, convex function on C_1, and $h : C_2 \to \mathbb{R}$ is a convex function with range$(h) \subset C_1$, then the composite function $f : C_2 \to \mathbb{R}$ defined as $f(\mathbf{x}) := g[h(\mathbf{x})]$ is convex on C_2.*
- Stability under Pointwise Supremum. *Let C be a convex set in $\mathbb{R}^n$. If $g_\alpha : C \to \mathbb{R}^m$, $\alpha = 1, 2, \ldots$, are convex functions on C, then the function $\mathbf{x} \mapsto \sup_\alpha g_\alpha(\mathbf{x})$ is convex on C.*

Proof. The proofs are left to the reader as an exercise. □

We shall now have a look to which standard functions these operations can be applied to. The usual way of checking convexity of a "simple" function is based on *differential criteria* of convexity.

Theorem A.17 (First-Order Condition of Convexity). *Let C be a convex set in $\mathbb{R}^n$ with a nonempty interior, and let $f : C \to \mathbb{R}$ be a function. Suppose f is continuous on C and differentiable on* int(C). *Then f is convex on* int(C) *if and only if*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\mathsf{T}[\mathbf{y} - \mathbf{x}]$$

holds for any two points $\mathbf{x}, \mathbf{y} \in C$.

Theorem A.18 (Second-Order Condition of Convexity). *Let C be a convex set in $\mathbb{R}^n$ with a nonempty interior, and let $f : C \to \mathbb{R}$ be a function. Suppose f is continuous on C and twice differentiable on* $\operatorname{int}(C)$. *Then f is convex on* $\operatorname{int}(C)$ *if and only if its Hessian matrix $\mathbf{H}(\mathbf{x})$ is positive semidefinite at each $\mathbf{x} \in \operatorname{int}(C)$.*

With the foregoing result, it is straightforward to verify that a great variety of functions is convex. However, a difficulty arises, e.g., if the set C is closed, since convexity can be established on the interior of C only. The following result can be used to overcome this difficulty.

Lemma A.19. *Let C be a convex set in $\mathbb{R}^n$ with a nonempty interior, and let $f : C \to \mathbb{R}$ be a function. If f is continuous on C and convex on* $\operatorname{int}(C)$*, then it is also convex on C.*

With the foregoing rules, convexity can be established for a great variety of complicated functions. This is illustrated in the following example.

Example A.20. Consider the exponential posynomial function $f : \mathbb{R}^n \to \mathbb{R}$ defined as

$$f(\mathbf{x}) = \sum_{i=1}^{N} c_i \exp(\mathbf{a}_i^{\mathsf{T}} \mathbf{x}),$$

with positive coefficients c_i. The function $\mathbf{x} \mapsto \exp(\mathbf{x})$ is convex on $\mathbb{R}^n$, for its Hessian matrix is positive definite at each $\mathbf{x} \in \mathbb{R}^n$. All functions $\mathbf{x} \mapsto \exp(\mathbf{a}_i^{\mathsf{T}} \mathbf{x})$ are therefore convex on $\mathbb{R}^n$ (stability of convexity under composition with an affine mapping). Finally, f is convex on $\mathbb{R}^n$ (stability of convexity under taking linear combinations with nonnegative coefficients).

A.4 LINEAR SPACES

The problems considered in the Chapters 2 and 3 of this textbook consist of optimizing a real valued function $\mathcal{J}$ defined on a subset $\mathcal{D}$ of a linear space $\mathcal{X}$. This section gives a summary of standard results for linear spaces, presupposing some familiarity with vector space operations in $\mathbb{R}^d$.

The principal requirement for a *(real) linear space*, also called *(real) vector space*, is that it contain the sums and (real) scalar multiples of its elements. In other words, a linear space must be closed under the operations of addition and scalar multiplication.

Definition A.21 (Linear Space). *A* real linear space *is a nonempty set $\mathcal{X}$ for which two operations called addition (denoted $+$) and (real) scalar multiplication (denoted $\cdot$) are defined. Addition is commutative and associative, making $\mathcal{X}$ an Abelian group under addition. Multiplication by scalars from the real number field is associative, and distributive with respect to $+$ as well as addition of scalars.*

We remark without proof that the set of real-valued functions f, g, on a (nonempty) set S forms a real linear space (or vector space) with respect to the operations of pointwise addition:

$$(f + g)(x) \;=\; f(x) + g(x) \quad \forall x \in S,$$

and scalar multiplication:

$$(\alpha f)(x) \;=\; \alpha f(x) \quad \forall x \in S,\ \alpha \in \mathbb{R}.$$

Likewise, for each $d = 1, 2, \ldots$ the set of all d-dimensional real vector valued functions on this set S forms a linear space with respect to the operations of component-wise addition and scalar multiplication.

If continuity is definable on S, then $\mathcal{C}(S)$ $(:= \mathcal{C}^0(S))$, the set of continuous real-valued functions on S, will be a real linear space since the sum of continuous functions, or the multiple of a continuous function by a real constant, is again a continuous function. Similarly, for each *open* subset D of a Euclidean space and each $k = 1, 2, \ldots, \mathcal{C}^k(D)$, the set of functions on D having continuous partial derivatives of order lower than or equal to k, is a real linear space, since the laws of differentiation guarantee that the sum or scalar multiple of such functions will be another. In addition, if D is bounded with boundary ∂D, and $\bar{D} := D \cup \partial D$, then $\mathcal{C}^k(\bar{D})$, the subset of $\mathcal{C}^k(\bar{D}) \cup \mathcal{C}(\bar{D})$ consisting of those functions whose partial derivatives of order lower than or equal to k each admit continuous extension to $\bar{D}$, is a real linear space. For example, a function x, which is continuous on $[a, b]$, is in $\mathcal{C}^1([a, b])$ if it is continuously differentiable in (a, b) and its derivative $\dot{x}$ has finite limiting values from the right at a (denoted $\dot{x}(a^+)$) and from the left at b (denoted $\dot{x}(b^-)$).

Example A.22. The function $x \mapsto x^{\frac{3}{2}}$ defines a function in $\mathcal{C}^1([0, 1])$, but $x \mapsto x^{\frac{1}{2}}$ does not.

For $d = 1, 2, \ldots$, $[\mathcal{C}(S)]^d$, $[\mathcal{C}^k(D)]^d$, and $[\mathcal{C}^k(\bar{D})]^d$, the sets of d-dimensional vector valued functions whose components are in $\mathcal{C}(S)$, $\mathcal{C}^k(D)$, and $\mathcal{C}^k(\bar{D})$, respectively, also form real linear spaces.

Definition A.23 (Linear Subspace). *A* linear subspace*, or simply a* subspace*, of the linear space $\mathcal{X}$ is a subset which is itself a linear space under the same operations.*

We note that subsets $\mathcal{D}$ of these spaces provide natural domain for optimization of real-valued functions in Chapters 2 and 3. However, these subsets do *not* in general constitute linear spaces themselves.

Example A.24. The subset

$$\mathcal{D} \;:=\; \{x \in \mathcal{C}([a, b]) : x(a) = 0, x(b) = 1\},$$

is *not* a linear space since if $x \in \mathcal{D}$ then $2x \notin \mathcal{D}$. $(2x(b) = 2 \neq 1.)$ On the other hand,

$$\mathcal{D} \;:=\; \{x \in \mathcal{C}([a, b]) : x(a) = 0, x(b) = 0\},$$

is a linear space.

Definition A.25 (Functional). *A function defined on a linear space $\mathcal{X}$ with range in $\mathbb{R}$ is called a* functional*.*

Definition A.26 (Continuous Functional). *Let $(\mathcal{X}, \|\cdot\|)$ be a normed linear space. A functional $\mathcal{F} : \mathcal{X} \to \mathbb{R}$ is said to be* continuous at $\mathbf{x} \in \mathcal{X}$*, if $\{\mathcal{F}(\mathbf{x}_k)\} \to \mathcal{F}(\mathbf{x})$, for any*

convergent sequence $\{\mathbf{x}_k\} \to \mathbf{x}$ *in* $\mathcal{X}$. *A functional is said to be* continuous *on* $\mathcal{X}$, *if it is continuous at any* $\mathbf{x} \in \mathcal{X}$.

Analysis in $\mathbb{R}^d$ is described most easily through inequalities between the lengths of its vectors. Similarly, in the real linear space $\mathcal{X}$, we shall assume that we can assign to each $\mathbf{x} \in \mathcal{X}$ a nonnegative number, denoted $\|\mathbf{x}\|$:

Definition A.27 (Norm). *A* norm $\|\cdot\|$ *on a linear space* $\mathcal{X}$ *is a nonnegative functional such that*

$$\begin{aligned} \|\mathbf{x}\| =& 0 \quad \textit{if and only if} \quad \mathbf{x} = \mathbf{0} \quad \forall \mathbf{x} \in \mathcal{X} \textit{ (positive definite)} \\ \|\alpha\mathbf{x}\| =& |\alpha| \, \|\mathbf{x}\| \quad \forall \mathbf{x} \in \mathcal{X}, \ \alpha \in \mathbb{R} \textit{ (positive homogeneous)} \\ \|\mathbf{x} + \mathbf{y}\| \leq& \|\mathbf{x}\| + \|\mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{X} \textit{ (triangle inequality).} \end{aligned}$$

There may be more than one norm for a linear space, although in a specific example, one may be more natural or more useful than another. Every norm also satisfies the so-called *reverse triangle inequality*:

$$|\|\mathbf{x}\| - \|\mathbf{y}\|| \leq \|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{X} \text{ (reverse triangle inequality).}$$

Definition A.28 (Normed Linear Space). *A* normed linear space *is a linear space with the topology induced by the norm defined on it: neighborhoods of any point* $\bar{\mathbf{x}}$ *are the balls*

$$\mathcal{B}_\eta(\bar{\mathbf{x}}) := \{\mathbf{x} \in \mathcal{X} : \|\mathbf{x} - \bar{\mathbf{x}}\| < \eta\},$$

with $\eta > 0$.

Two possible norms on the linear space of continuous real-values functions on $[a, b]$ are:

$$\|x\|_\infty := \max_{a \leq t \leq b} |x(t)| \tag{A.3}$$

$$\|x\|_p := \left(\int_a^b |x(t)|^p \, \mathrm{d}t \right)^{\frac{1}{p}}. \tag{A.4}$$

Further, since $\mathcal{C}^k[a, b] \subset \mathcal{C}[a, b]$, for each $k = 1, 2, \ldots$, it follows that (A.3) and (A.4) also define norms on $\mathcal{C}^k[a, b]$. However, these norms do not take cognizance of the differential properties of the functions and supply control only over their continuity. Alternative norms on $\mathcal{C}^k[a, b]$ supplying control over the k first derivatives are:

$$\|x\|_{k,\infty} := \|x\|_\infty + \|x^{(k)}\|_\infty = \max_{a \leq t \leq b} |x(t)| + \max_{a \leq t \leq b} |x^{(k)}(t)| \tag{A.5}$$

$$\|x\|_{k,p} := \|x\|_p + \|x^{(k)}\|_p = \left(\int_a^b |x(t)|^p \, \mathrm{d}t \right)^{\frac{1}{p}} + \left(\int_a^b |x^{(k)}(t)|^p \, \mathrm{d}t \right)^{\frac{1}{p}}. \tag{A.6}$$

Norms can be defined in a likewise fashion for the linear spaces $\mathcal{C}[a, b]^d$, and $\mathcal{C}^k[a, b]^d$, with $d = 1, 2, \ldots$. For example,

$$\|\mathbf{x}\|_\infty := \max_{a \leq t \leq b} \|\mathbf{x}(t)\|, \tag{A.7}$$

defines a norm on $\mathcal{C}[a, b]^d$, and

$$\|\mathbf{x}\|_{k,\infty} := \|\mathbf{x}\|_\infty + \|\mathbf{x}^{(k)}\|_\infty = \max_{a \leq t \leq b} \|\mathbf{x}(t)\| + \max_{a \leq t \leq b} \|\mathbf{x}^{(k)}(t)\|, \tag{A.8}$$

defines a norm on $\mathcal{C}^k[a,b]^d$, where $\|\mathbf{x}(t)\|$ stands for any norm in $\mathbb{R}^d$.

Definition A.29 (Equivalent Norms). *Let $\|\cdot\|$ and $\|\cdot\|'$ be two norm on a linear space $\mathcal{X}$. These norm are said to be* equivalent norms *if there exist positive real numbers α, β such that*

$$\alpha\|\mathbf{x}\| \leq \|\mathbf{x}\|' \leq \beta\|\mathbf{x}\|.$$

While all norms can be shown to be equivalent on a finite dimensional linear space, this result does *not* hold on infinite dimensional spaces. This is illustrated in the following:

Example A.30. Consider the linear space of continuously differentiable real-valued functions $\mathcal{C}^1[0,1]$, supplied with the maximum norms $\|\cdot\|_\infty$ and $\|\cdot\|_{1,\infty}$, as defined in (A.3) and (A.5), respectively. Let sequence of functions $\{x_k\} \in \mathcal{C}^1[0,1]$ be defined as

$$x_k(t) \;:=\; 2^{2k}t^k(1-t)^k.$$

It is easily shown that

$$\|x_k\|_\infty = \left|x_k\left(\tfrac{1}{2}\right)\right| = 1 \quad \text{for each } k \geq 1.$$

On the other hand, the maximum value of the first derivative $|\dot{x}_k(t)|$ on $[0,1]$ is attained at $t^\pm = \frac{1}{2}\frac{\sqrt{2k-1}\pm 1}{\sqrt{2k-1}}$, yielding

$$\|\dot{x}_k\|_\infty = |\dot{x}_k(t^\pm)| = 2\sqrt{2k-1}\frac{k}{k-1}\left(\frac{\sqrt{2k-1}+1}{\sqrt{2k-1}}\right)^k\left(\frac{\sqrt{2k-1}-1}{\sqrt{2k-1}}\right)^k,$$

As k grows large, we thus have $\|\dot{x}_k\|_{1,\infty} = \|x_k\|_\infty + \|\dot{x}_k\|_\infty \sim \sqrt{k}$. Hence, for any $\beta > 0$, there is always a k such that

$$\|\dot{x}_k\|_{1,\infty} \;>\; \beta\|\dot{x}_k\|_\infty.$$

This proves that the norms $\|\cdot\|_\infty$ and $\|\cdot\|_{1,\infty}$ are not equivalent on $\mathcal{C}^1[0,1]$.

Definition A.31 (Convergent Sequence). *A sequence $\{\mathbf{x}_k\}$ in a normed linear space $(\mathcal{X}, \|\cdot\|)$ is said to be* convergent *if there is an element $\bar{x} \in \mathcal{X}$ such that:*

$$\forall\varepsilon > 0, \exists N(\varepsilon) > 0 \textit{ such that } \|\mathbf{x}_k - \bar{\mathbf{x}}\| < \varepsilon, \forall k \geq N.$$

We say that $\{\mathbf{x}_k\}$ converges to $\bar{\mathbf{x}}$.

Hence, the convergence of a sequence in a linear space is reduced to the convergence of a sequence of real numbers via the use of a norm. A point $\bar{\mathbf{x}}$ in a normed linear space $\mathcal{X}$, is said to be a *limit point* of a set $\mathcal{D} \in \mathcal{X}$ if there exists a sequence $\{\mathbf{x}_k\}$ in $\mathcal{D}$ such that $\{\mathbf{x}_k\} \to \bar{\mathbf{x}}$.

Definition A.32 (Cauchy Sequence). *A sequence $\{\mathbf{x}_k\}$ in a normed linear space $(\mathcal{X}, \|\cdot\|)$ is said to be a* Cauchy sequence *if*

$$\forall\varepsilon > 0, \exists N > 0 \textit{ such that } \|\mathbf{x}_n - \mathbf{x}_m\| < \varepsilon, \forall n, m \geq N.$$

While every convergent sequence is a Cauchy sequence, **the reverse does not necessarily hold true in a normed linear space**. This motivates the following:

Definition A.33 (Completeness, Banach Space). *A normed linear space* $(\mathcal{X}, \|\cdot\|)$ *in which every Cauchy sequence is a convergent sequence in* $\mathcal{X}$ *is said to be* complete. *A complete normed linear space is called a* Banach space.

Example A.34 (Complete Function Space). The linear space of continuous functions, $\mathcal{C}([a,b])$, equipped with the maximum norm $\|\cdot\|_\infty$, is a Banach space. The linear space of continuously differentiable functions, $\mathcal{C}^1([a,b])$, equipped with the maximum norm $\|\cdot\|_{1,\infty}$, is a Banach space too.

Example A.35 (Incomplete Function Space). Consider the function space $\mathcal{C}^1[0,1]$, supplied with the norm $\|\cdot\|_p$ as defined in (A.4), and let $\{x_k\} \in \mathcal{C}^1[0,1]$ be defined as earlier in Example A.30,

$$x_k(t) \;:=\; 2^{2k} t^k (1-t)^k.$$

It can be established that the limit $\bar{x}$ of $\{x_k\}$ as $k \to +\infty$ is a real-valued function given by

$$\bar{x}(t) = \begin{cases} 1 & \text{if } t = 1 \\ 0 & \text{otherwise,} \end{cases}$$

which is not in $\mathcal{C}^1[0,1]$. In fact, $\bar{x}(t) \in L^p[0,1]$[1], and $\{x_k\}$ is convergent in the function space $L^p[0,1]$. Therefore, $\{x_k\}$ is a Cauchy sequence in $\mathcal{C}^1[0,1]$ relative to the norm $\|\cdot\|_p$, which does not have a limit in $\mathcal{C}^1[0,1]$. We have thus established that $(\mathcal{C}^1[0,1], \|\cdot\|_p$ is not a complete normed linear space.

Definition A.36 (Ball). *Let* $(\mathcal{X}, \|\cdot\|)$ *be a normed linear space. Given a point* $\bar{\mathbf{x}} \in \mathcal{X}$ *and a real number* $r > 0$*, a* ball centered at $\bar{\mathbf{x}}$ and of radius r *is the set*

$$\mathcal{B}_r(\bar{\mathbf{x}}) := \{\mathbf{x} \in \mathcal{X} : \|\mathbf{x} - \bar{\mathbf{x}}\| < r\}$$

Definition A.37 (Open Set, Closed Set). *A subset* $\mathcal{D}$ *of a normed linear space* $(\mathcal{X}, \|\cdot\|)$ *is said to be* open *if it contains a ball arount each of its points. A subset* $\mathcal{K}$ *of* $\mathcal{X}$ *is said to be* closed *if its complement in* $\mathcal{X}$ *is open.*

Theorem A.38 (Closed Set). *Let* $\mathcal{K}$ *be a nonempty subset of a normed linear space* $(\mathcal{X}, \|\cdot\|)$*. Then,* $\mathcal{K}$ *is closed if and only if every convergent sequence* $\{\mathbf{x}_k\} \in \mathcal{K}$ *converges to an element* $\bar{\mathbf{x}} \in \mathcal{K}$*.*

Definition A.39 (Totally Bounded Set). *Let* $(\mathcal{X}, \|\cdot\|)$ *be a normed linear space. A set* $\mathcal{D} \subset \mathcal{X}$ *is said to be* totally bounded *if*

$$\forall \varepsilon > 0, \exists n \geq 1 \textit{ (finite) and} (d_1, \ldots, d_n) \in \mathcal{D} \textit{ such that } \mathcal{D} \subseteq \bigcup_{k=1}^{n} \mathcal{B}_\varepsilon(d_k).$$

[1] $L^p(\Omega)$, $p \geq 1$, stands for the linear space of p-integrable functions, i.e., all functions $f : \Omega \to \mathbb{R}$ with $\int_\Omega f(x)^p \mathrm{d}x < \infty$.

Definition A.40 ((Sequentially) Compact Normed Linear Space). *A normed linear space* $(\mathcal{X}, \|\cdot\|)$ *is said to be* sequentially compact *if every sequence in* $\mathcal{X}$ *has a convergent subsequence in* $\mathcal{X}$*.*

Theorem A.41 (Characterization of Compact Normed Linear Spaces). *A normed linear space* $(\mathcal{X}, \|\cdot\|)$ *is (sequentially) compact if and only if* $\mathcal{X}$ *is* totally bounded *and* complete*.*

Definition A.42 ((Sequentially) Compact Set). *Let* $(\mathcal{X}, \|\cdot\|)$ *be a normed linear space. A set* $\mathcal{K} \subset \mathcal{X}$ *is said to be* sequentially compact*, or simply* compact*, if every subsequence in* $\mathcal{K}$ *has a subsequence that converges to a point in* $\mathcal{K}$*.*

Theorem A.43 (Characterization of Compact Sets). *A subset* $\mathcal{K}$ *of a compact normed linear space* $(\mathcal{X}, \|\cdot\|)$ *is compact if and only if it is closed.*

In particular, it should be noted that a compact subset of a normed linear space is both closed and bounded. However, the converse is true only for finite dimensional spaces.

We close this subsection with the extension of Weierstrass' Theorem 1.14 (p. 7) to general normed linear spaces:

Theorem A.44 (Weierstrass' Theorem for Normed Linear Spaces). *A continuous functional* $\mathcal{J}$ *on a compact subset* $\mathcal{K}$ *of a compact normed linear space* $(\mathcal{X}, \|\cdot\|)$ *assumes both its maximum and minimum values at points in* $\mathcal{K}$*. In particular, these values are finite.*

A.5 FIRST-ORDER ORDINARY DIFFERENTIAL EQUATIONS

This section states some fundamental properties of the solutions of *ordinary differential equations (ODEs)*, such as existence, uniqueness, continuous dependence on initial conditions and parameters, and differentiabililty. These properties are *essential* for the state equation $\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x})$ to be a useful mathematical model of a physical system.

A.5.1 Existence and Uniqueness

For a mathematical model of a given system to predict the future state of that system from its current state at t_0, the *initial value problem (IVP)*

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}); \quad \mathbf{x}(t_0) = \mathbf{x}_0, \tag{A.9}$$

with $\mathbf{x} \in \mathbb{R}^{n_x}$ and $\mathbf{f} : \mathbb{R} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$, must have a unique solution. By a solution of (A.9) over an interval $[t_0, t_\mathrm{f}]$, we mean a continuous vector-valued function $\mathbf{x} : [t_0, t_\mathrm{f}] \to \mathbb{R}^{n_x}$, such that $\dot{\mathbf{x}}(t)$ is defined and $\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t))$, for all $t \in [t_0, t_\mathrm{f}]$. If $\mathbf{f}(t, \mathbf{x})$ is continuous both in t and $\mathbf{x}$, then the solution $\mathbf{x}(t)$ will be continuously differentiable. We shall assume herein that $\mathbf{f}(t, \mathbf{x})$ is continuous in $\mathbf{x}$, but only piecewise continuous in t, in which case, a solution $\mathbf{x}(t)$ could only be piecewise continuously differentiable, i.e., $\mathbf{x} \in \hat{\mathcal{C}}^1[t_0, t_\mathrm{f}]$. The assumption that $\mathbf{f}(t, \mathbf{x})$ be piecewise continuous in t allows us to include the case wherein $\mathbf{f}(t, \mathbf{x}(t)) := \mathbf{g}(t, \mathbf{u}(t), \mathbf{x}(t))$ depends on a time-varying input $\mathbf{u}(t) \in \mathbb{R}^{n_u}$ that may experience step changes with time, e.g., $\mathbf{u} \in \hat{\mathcal{C}}[t_0, t_\mathrm{f}]$.

Prior to giving local existence and uniqueness conditions for the solution to an IVP in ODEs, we need the following:

Definition A.45 (Local Lipschitzness). *The function* $\mathbf{f}(\mathbf{x})$ *is said to be* Lipschitz *at* $\mathbf{x}_0 \in \mathbb{R}^{n_x}$ *if there exist constants* $K \geq 0$ *and* $\eta > 0$ *such that*[2]

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| \leq K\|\mathbf{x} - \mathbf{y}\|, \tag{A.10}$$

for every $\mathbf{x}, \mathbf{y} \in \mathcal{B}_\eta(\mathbf{x}_0)$. *Moreover,* $\mathbf{f}(\mathbf{x})$ *is said to be* locally Lipschitz *on* X*, an open connected subset of* $\mathbb{R}^{n_x}$*, if it is Lipschitz at each* $\mathbf{x}_0 \in X$.

Likewise, the function $\mathbf{f}(t, \mathbf{x})$*,* $t \in [t_0, t_\mathrm{f}]$*, is said to be Lipschitz at* $\mathbf{x}_0 \in \mathbb{R}^{n_x}$ *provided the Lipschitz condition* (A.10) *holds* uniformly *in* $t \in [t_0, t_\mathrm{f}]$*;* $\mathbf{f}(t, \mathbf{x})$ *is said to be locally Lipschitz on* X *for* $t \in [t_0, t_\mathrm{f}]$ *provided that it is Lipschitz at any point* $\mathbf{x}_0 \in X$.

Note, in particular, that the Lipschitz property is stronger than continuity, but weaker than continuous differentiability. We are now ready to state the following:

Theorem A.46 (Local Existence and Uniqueness). *Let* $\mathbf{f}(t, \mathbf{x})$ *be piecewise continuous in* t*, and Lipschitz at* $\mathbf{x}_0 \in \mathbb{R}^{n_x}$ *for* $t \in [t_0, t_\mathrm{f}]$*. Then, there exists some* $\delta > 0$ *such that the state equation* $\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x})$ *with* $\mathbf{x}(t_0) = \mathbf{x}_0$ *has a unique solution over* $[t_0, t_0 + \delta]$.

The key assumption in Theorem A.46 is the Lipschitz condition at $\mathbf{x}_0$. Strictly, only continuity of $\mathbf{f}(t, \mathbf{x})$ with respect to $\mathbf{x}$ is needed to ensure existence of a solution. However, continuity is not sufficient to ensure uniqueness of that solution as illustrated subsequently:

Example A.47. The scalar differential equation

$$\dot{x}(t) = [x(t)]^{\frac{1}{3}},$$

with initial condition $x(0) = 0$ has a solution $x(t) = [\frac{2}{3}t]^{\frac{3}{2}}$ for each $t > 0$. This solution is not unique, however, since $x(t) = 0, \forall t > 0$ is another solution.

The foregoing Theorem A.46 gives conditions under which a solution to an IVP in ODEs of the form (A.9) exists and is unique over an interval $[t_0, t_0+\delta]$, where δ may be very small. In other words, we have no control on δ, and cannot guarantee existence and uniqueness over a given time interval $[t_0, t_\mathrm{f}]$. Starting at time t_0, with an initial state $\mathbf{x}(t_0) = \mathbf{x}_0$, Theorem A.46 shows that there is a positive constant δ (dependent on $\mathbf{x}_0$) such that the state equation (A.9) has a unique solution over the time interval $[t_0, t_0 + \delta]$. Then, taking $t_0 + \delta$ as a new initial time and $\mathbf{x}(t_0 + \delta)$ as a new initial state, one may try to apply Theorem A.46 to establish existence and uniqueness of the solution beyond $t_0 + \delta$. If the conditions of the theorem are satisfied at $(t_0 + \delta, \mathbf{x}(t_0 + \delta))$, then there exists $\delta_2 > 0$ such that the equation has a unique solution over $[t_0+\delta, t_0+\delta+\delta_2]$, that passes through the point $(t_0+\delta, \mathbf{x}(t_0+\delta))$. The solutions over $[t_0, t_0+\delta]$ and $[t_0+\delta, t_0+\delta+\delta_2]$ can now be pieced together to establish the existence of a unique solution over $[t_0, t_0 + \delta + \delta_2]$. This idea can be repeated to keep extending the solution. However, in general, the interval of existence of the solution cannot be extended indefinitely because the conditions of Theorem A.46 may cease to hold. In other words, there is a *maximum interval* $[t_0, T)$ where the unique solution starting at (t_0, x_0) exists. Clearly, T may be less than t_f, in which case the solution leaves any compact set over which $\mathbf{f}$ is locally Lipschitz in $\mathbf{x}$ as $t \to T$. The term *finite escape time* is used to describe this phenomenon.

[2]Here, $\|\cdot\|$ stands for any norm in $\mathbb{R}^{n_x}$.

Example A.48. Consider the scalar differential equation

$$\dot{x}(t) = [x(t)]^2,$$

with $x(0) = 1$. The function $f(x) = x^2$ is locally Lipschitz on $\mathbb{R}$. Hence, it is Lipschitz on any compact subset of $\mathbb{R}$. However, the unique solution

$$x(t) = \frac{1}{1-t},$$

passing through the point $(0, 1)$, exists over $[0, 1)$. As $t \to 1$, $x(t)$ leaves any compact set.

In view of the preceding discussion, one may ask the question whether additional conditions could be imposed, if any, so that a solution can be extended indefinitely. Prior to giving one such condition, we need the following:

Definition A.49 (Global Lipschitzness). *The function* $\mathbf{f}(\mathbf{x})$ *is said to be* globally Lipschitz *if there exists a constant* $K \geq 0$ *such that*

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| \leq K\|\mathbf{x} - \mathbf{y}\|, \tag{A.11}$$

for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{n_x}$. *(Here,* K *must be the same for every pair of points* x*,* y *in* $\mathbb{R}^{n_x}$*.) Likewise, the function* $\mathbf{f}(t, \mathbf{x})$ *is said to be globally Lipschitz for* $t \in [t_0, t_\mathrm{f}]$*, provided that* (A.11) *holds for every* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{n_x}$*,* uniformly *in* $t \in [t_0, t_\mathrm{f}]$.

The global Lipschitzness property is sufficient for a solution to be extended indefinitely:

Theorem A.50 (Global Existence and Uniqueness I). *Let* $\mathbf{f}(t, \mathbf{x})$ *be piecewise continuous in* t*, and globally Lipschitz in* $\mathbf{x}$*, for* $t \in [t_0, t_\mathrm{f}]$*. Then, the state equation* $\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x})$ *with* $\mathbf{x}(t_0) = \mathbf{x}_0$ *has a unique solution over* $[t_0, t_\mathrm{f}]$.

Example A.51. Consider the linear system

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{b}(t), \tag{A.12}$$

where the elements of $\mathbf{A} \in \mathbb{R}^{n_x \times n_x}$ and $\mathbf{b} \in \mathbb{R}^{n_x}$ are piecewise continuous functions of t. Over any finite time interval $[t_0, t_\mathrm{f}]$, the elements of $\mathbf{A}(t)$ are bounded. Hence, $\|\mathbf{A}(t)\| \leq a$, where $\|\cdot\|$ stands for any any matrix norm, and we have

$$\|\mathbf{f}(t, \mathbf{x}) - \mathbf{f}(t, \mathbf{y})\| = \|\mathbf{A}(t)\,(\mathbf{x} - \mathbf{y})\,\| = \|\mathbf{A}(t)\|\,\|\mathbf{x} - \mathbf{y}\| \leq a\|\mathbf{x} - \mathbf{y}\|,$$

for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{n_x}$. Therefore, Theorem A.50 applies and the linear system (A.12) has a unique solution over $[t_0, t_\mathrm{f}]$. Since t_1 can be arbitrarily large, we can also conclude that a linear system has a unique solution, provided that $\mathbf{A}(t)$ and $\mathbf{b}(t)$ are piecewise continuous for all $t \geq t_0$. Hence, a linear system *cannot* have a finite escape time.

In view of the conservative nature of the global Lipschitz condition, it would be useful to have a global existence and uniqueness theorem requiring the function $\mathbf{f}$ to be only locally Lipschitz. The next theorem achieves that at the expense of having to know more about the solution of the system:

Theorem A.52 (Global Existence and Uniqueness II). *Let* $\mathbf{f}(t, \mathbf{x})$ *be piecewise continuous in* t*, and locally Lipschitz in* $\mathbf{x}$*, for all* $t \geq t_0$ *and all* $\mathbf{x} \in D \subset \mathbb{R}^{n_x}$*. Let also* X *be a compact subset of* D*,* $\mathbf{x}_0 \in X$*, and suppose it is known that every solution of*

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

lies entirely in X*. Then, there is a unique solution that is defined for all* $t \geq t_0$*.*

Example A.53. Consider the scalar differential equation

$$\dot{x}(t) = f(x) := -[x(t)]^3,$$

with $x(0) = a$. Observe that the function f is locally Lipschitz on $\mathbb{R}$, but does *not* satisfy a global Lipschitz condition since the Jacobian $f_x(x) = -3x^2$ is not globally bounded. That is, Theorem A.50 does not apply. Now, remarking that, at any time instant $t \geq 0$, $\dot{x}(t) \leq 0$ when $x(t) \geq 0$ and $\dot{x}(t) \geq 0$ when $x(t) \leq 0$, a solution cannot leave the compact set $X := \{x \in \mathbb{R} : |x| \leq |a|\}$. Thus, without calculating the solution, we conclude by Theorem A.52 that the differential equation has a unique solution for all $t \geq 0$.

A.5.2 Continuous Dependence on Initial Conditions and Parameters

We now turn to the problem of continuous dependence on initial conditions and parameters for the IVP in ODEs

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}, \mathbf{p}); \quad \mathbf{x}(t_0; \mathbf{p}) = \mathbf{x}_0, \tag{A.13}$$

with $\mathbf{p} \in \mathbb{R}^{n_p}$ being constant parameters, e.g., representing physical parameters of the system. Continuous dependence is an important property that any model of interest should possess. It is defined as follows:

Definition A.54 (Continuous Dependence on Initial Conditions and Parameters). *Let* $\mathbf{x}(t; \mathbf{p}^0)$ *be a solution of* (A.13)*, with* $\mathbf{x}(t_0; \mathbf{p}^0) = \mathbf{x}_0^0$*, defined on* $[t_0, t_\mathrm{f}]$*. Then,* $\mathbf{x}(t; \mathbf{p}^0)$ *is said to depend continuously on* $\mathbf{x}_0$ *and* $\mathbf{p}$ *if, for any* $\varepsilon > 0$*, there is* $\delta > 0$ *such that for all* $\mathbf{x}_0^1 \in \mathcal{B}_\delta\left(\mathbf{x}_0^0\right)$ *and* $\mathbf{p}^1 \in \mathcal{B}_\delta\left(\mathbf{p}^0\right)$*,* (A.13) *has a unique solution* $\mathbf{x}(t; \mathbf{p}^1)$*, defined on* $[t_0, t_\mathrm{f}]$*, with* $\mathbf{x}(t_0; \mathbf{p}^1) = \mathbf{x}_0^1$*, and* $\mathbf{x}(t; \mathbf{p}^1)$ *satisfies*

$$\|\mathbf{x}(t; \mathbf{p}^1) - \mathbf{x}(t; \mathbf{p}^0)\| < \varepsilon, \quad \forall t \in [t_0, t_\mathrm{f}].$$

We can now state the main theorem on the continuity of solutions in terms of initial states and parameters:

Theorem A.55 (Continuous Dependence on Initial Conditions and Parameters). *Let* $X \subset \mathbb{R}^{n_x}$ *be an open connected set, and* $\mathbf{p}_0 \in \mathbb{R}^{n_p}$*. Suppose that* $\mathbf{f}(t, \mathbf{x}, \mathbf{p})$ *is piecewise continuous in* $(t, \mathbf{x}, \mathbf{p})$ *and locally Lipschitz (uniformly in* t *and* $\mathbf{p}$*) on* $[t_0, t_\mathrm{f}] \times X \times \mathcal{B}_\eta\left(\mathbf{p}^0\right)$*, for some* $\eta > 0$*. Let* $\mathbf{x}(t; \mathbf{p}^0)$ *be a solution of* $\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}, \mathbf{p})$ *with* $\mathbf{x}(t_0; \mathbf{p}^0) := \mathbf{x}_0^0 \in X$*, and suppose that* $\mathbf{x}(t; \mathbf{p}^0)$ *exists and belongs to* X *for all* $t \in [t_0, t_\mathrm{f}]$*. Then,* $\mathbf{x}(t; \mathbf{p}^0)$ *depends continuously on* $\mathbf{x}_0$ *and* $\mathbf{p}$*.*

A.5.3 Differentiability of Solutions

Suppose that $\mathbf{f}(t, \mathbf{x}, \mathbf{p})$ is continuous in $(t, \mathbf{x}, \mathbf{p})$ and has continuous first partial derivatives with respect to $\mathbf{x}$ and $\mathbf{p}$ for all $(t, \mathbf{x}, \mathbf{p}) \in [t_0, t_\mathrm{f}] \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_p}$. Suppose also that $\mathbf{h}(\mathbf{p})$ is continuous and has continuous first partial derivatives with respect to $\mathbf{p}$ in $\mathbb{R}^{n_p}$. Let $\mathbf{p}^0$ be a nominal value of $\mathbf{p}$, and suppose that the nominal state equation

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}, \mathbf{p}^0); \quad \mathbf{x}(t_0; \mathbf{p}^0) = \mathbf{h}(\mathbf{p}^0), \tag{A.14}$$

has a unique solution $\mathbf{x}(t; \mathbf{p}^0)$ over $[t_0, t_\mathrm{f}]$. From Theorem A.55, we know that for all $\mathbf{p}$ sufficiently close to $\mathbf{p}^0$, the state equation

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

has a unique solution $\mathbf{x}(t; \mathbf{p})$ over $[t_0, t_\mathrm{f}]$ that is close to the nominal solution $\mathbf{x}(t; \mathbf{p}^0)$. The continuous differentiability of $\mathbf{f}$ with respect to $\mathbf{x}$ and $\mathbf{p}$, together with the continuous differentiability of $\mathbf{h}$ with respect to $\mathbf{p}$, implies the additional property that $\mathbf{x}(t; \mathbf{p})$ is differentiable with respect to $\mathbf{p}$ near $\mathbf{p}^0$, at each $t \in [t_0, t_\mathrm{f}]$.[3] This is easily seen upon writing

$$\mathbf{x}(t; \mathbf{p}) = \mathbf{h}(\mathbf{p}) + \int_{t_0}^{t} \mathbf{f}(\tau, \mathbf{x}(\tau; \mathbf{p}), \mathbf{p}) \, \mathrm{d}\tau,$$

and then taking partial derivatives with respect to $\mathbf{p}$,

$$\mathbf{x}_\mathbf{p}(t; \mathbf{p}) = \mathbf{h}_\mathbf{p}(\mathbf{p}) + \int_{t_0}^{t} [\mathbf{f}_\mathbf{x}(\tau, \mathbf{x}(\tau; \mathbf{p}), \mathbf{p}) \mathbf{x}_\mathbf{p}(\tau; \mathbf{p}) + \mathbf{f}_\mathbf{p}(\tau, \mathbf{x}(\tau; \mathbf{p}), \mathbf{p})] \, \mathrm{d}\tau. \tag{A.15}$$

(See Theorem 2.A.59 on p. 102 for differentiation under the integral sign.)

A.6 NOTES AND REFERENCES

There are many excellent textbooks on real analysis. We just mention here the textbooks by Rudin [46] and Sohrab [50].

The material presented in Section A.3 is mostly a summary of the material in Chapters 2 and 3 of the book by Bazaraa, Sherali and Shetty [6], where most of the omitted proofs can be found. See also Chapters 2 and 3 of the book by Boyd and Vandenberghe [10].[4] The classical, comprehensive text on convex analysis is Rockafellar's book [44].

Similarly, there are many excellent textbooks on functional analysis. A particularly accessible introductory textbook is that by Kreysig [33], whose only real prerequisites are a solid understanding of calculus and some familiarity with Linear Algebra. The classical textbooks in functional analysis are Rudin's book [47] as well as Lax's book [34].

Finally, the summary on nonlinear differential equations presented in Section A.3 is mostly taken from the excellent textbook by Khalil [30, Chapter 3].

[3] This result can be readily extended to the solution $\mathbf{x}(t; \mathbf{p})$ being $\mathcal{C}^k$ with respect to $\mathbf{p}$ near $\mathbf{p}^0$, at each $t \in [t_0, t_\mathrm{f}]$, when $\mathbf{f}$ and $\mathbf{h}$ are themselves $\mathcal{C}^k$ with respect to $(\mathbf{x}, \mathbf{p})$ and $\mathbf{p}$, respectively.

[4] An electronic copy of this book can be obtained at `http://www.stanford.edu/ boyd/cvxbook/`.

Bibliography

1. F. Allgower and A. Zheng. *Non-Linear Model Predictive Control*, volume 26 of *Progress in Systems and Control Theory*. Birkhäuser, Basel, Switzerland, 2000.

2. P. J. Antsaklis and A. M. Michel. *Linear Systems*. Series in electrical and computer engineering. McGraw-Hill, New York, 1997.

3. U. M. Ascher and L. R. Petzold. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. SIAM, Philadelphia (PA), 1998.

4. G. Bader and U. M. Ascher. A new basis implementation for mixed order boundary value ODE solver. *SIAM Journal on Scientific Computing*, 8:483–500, 1987.

5. R. Banga, J. R. anf Irizarry-Rivera and W. D. Seider. Stochastic optimization for optimal and model-predictive control. *Computers & Chemical Engineering*, 22(4/5):603–612, 1995.

6. M. S. Bazarra, H. D. Sherali, and C. M. Shetty. *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons, New York, 2nd edition, 1993.

7. D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont (MA), 2nd edition, 1999.

8. J. T. Betts. *Practical Methods for Optimal Control Using Nonlinear Programming*. Advances in Design and Control. SIAM, Philadelphia (PA), 2001.

9. L. T. Biegler, A. Cervantes, and A. Wächter. Advances in simultaneous strategies for dynamic process optimization. *Chemical Engineering Science*, 57:575–593, 2002.

10. S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge (UK), 2nd edition, 2006.

11. K. E. Brenan, S. L. Campbell, and L. R. Petzold. *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, volume 14 of *Classics in Applied Mathematics*. SIAM, Philadelphia (PA), 1996.

12. Jr. Bryson, A. E and Y.-C. Ho. *Applied Optimal Control: Optimization, Estimation and Control*. Hemisphere Publishing Corporation, Washington, D.C., 1975.

13. R. Bulirsch and H. J. Montrone, F. Pesch. Abort landing in the presence of windshear as a minimax optimal control problem. Part 1: Necessary conditions. *Journal of Optimization Theory & Applications*, 70:1–23, 1991.

14. C. Büskens and H. Maurer. SQP-methods for solving optimal control problems with control and state constraints: adjoint variables, sensitivity analysis and real-time control. *Journal of Computational and Applied Mathematics*, 120:85–108, 2000.

15. Y. Cao, S. Li, and L. R. Petzold. Adjoint sensitivity analysis for differential-algebraic equations: The adjoint DAE system and its numerical solution. *SIAM Journal on Scientific Computing*, 24(3):1076–1089, 2003.

16. L. Cesari. *Optimization-Theory and Application – Problems with Ordinary Differential Equations*. Springer-Verlag, New York, 1983.

17. J.-C. Culioli. *Introduction à l'Optimisation*. Ellipses, Paris, France, 1994.

18. S. A. Dadebo and K. B. McAuley. Dynamic optimization of constrained chemical engineering problems using dynamic programming. *Computers & Chemical Engineering*, 19(5):513–525, 1995.

19. R. M. Errico. What is an adjont model? *Bulletin of the American Meteorological Society*, 78(11):2577–2591, 1997.

20. W. F. Feehery, J. E. Tolsma, and P. I. Barton. Efficient sensitivity analysis of large-scale differential-algebraic systems. *Applied Numerical Mathematics*, 25(1):41–54, 1997.

21. T. Fuller. Relay control systems optimized for various performance criteria. In *Proceedings of the First IFAC World Congress*, volume 1, pages 510–519, London (UK), 1961. Butterworths.

22. I. M. Gelfand and S. V. Fomin. *Calculus of Variations*. Prentice-Hall, London (UK), 1963.

23. B. Gollan. On optimal control problems with state constraints. *Journal of Optimization Theory & Applications*, 32(1):75–80, 1980.

24. A. Griewank. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. Frontiers in Applied Mathematics. SIAM, Philadelphia (PA), 2000.

25. R. F. Hartl, S. P. Sethi, and R. G. Vickson. A survey of the Maximum Principles for optimal control problems with state constraints. *SIAM Review*, 37(2):181–218, 1995.

26. J. Herskovits. A view on nonlinear optimization. In J. Herskovits, editor, *Advances in Structural Optimization*, pages 71–117, Dordrecht, the Netherlands, 1995. Kluwer Academic Publishers.

27. D. H. Jacobson, M. M. Lele, and J. L. Speyer. New necessary conditions of optimality for control problems with state-variable inequality constraints. *Journal of Mathematical Analysis & Applications*, 35:255–284, 1971.

28. M. I. Kamien and N. L. Schwartz. *Dynamic Optimization: The Calculus of Variations and Optimal Control in Economics and Management*, volume 31 of *Advanced Textbooks in Economics*. North-Holland, Amsterdam, The Netherlands, 2nd edition, 1991.

29. H. J. Kelley, R. E. Kopp, and H. G. Moyer. Singular extremals. In G. Leitmann, editor, *Topics in Optimization*, pages 63–101, New York, 1967. Academic Press.

30. H. K. Khalil. *Nonlinear Systems*. Prentice Hall, Upper Saddle River (NJ), 3rd edition, 2002. (ISBN: 0-13-067389-7).

31. M. A. Kramer and J. R. Leis. The simultaneous solution and sensitivity analysis of systems described by ordinary differential equations. *ACM Transactions on Mathematical Software*, 14:45–60, 1988.

32. E. Kreindler. Additional necessary conditions for optimal control with state-variable inequality constraints. *Journal of Optimization Theory & Applications*, 38(2):241–250, 1982.

33. E. Kreyszig. *Introductory Functional Analysis with Applications*. John Wiley & Sons, New York, 1978.

34. P. D. Lax. *Functional Analysis*. Wiley-Interscience, New York, 2002.

35. J. S. Logsdon and L. T. Biegler. Accurate solution of differential-algebraic optimization problems. *Industrial & Engineering Chemistry Research*, 28:1628–1639, 1989.

36. D. Luenberger. *Linear and Nonlinear Programming*. Addison-Wesley, Reading (MA), 2nd edition, 1984.

37. R. Luus. *Iterative Dynamic Programming*. Chapman & Hall, Boca Raton, FL, 2000.

38. J. Macki and A. Strauss. *Introduction to Optimal Control Theory*. Undergraduate texts in mathematics. springer-Verlag, New York, 2nd edition, 1982.

39. T. Maly and L. R. Petzold. Numerical methods and software for sensitivity analysis of differential-algebraic systems. *Applied Numerical Mathematics*, 20(1-2):57–79, 1996.

40. J. P. McDanell and W. F. Powers. Necessary conditions for joining optimal singular and nonsingular subarcs. *SIAM Journal of Control*, 9(2):161–173, 1971.

41. L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko. *The Mathematical Theory of Optimal Processes*. Pergamon Press, New York, 1964.

42. S. J. Qin and T. A. Badgwell. An overview of industrial model predictive technology. In *5th Chemical Process Control Conference*, pages 232–256, Tahoe City (CA), 1997.

43. H. Robbins. Junction phenomena for optimal control with state-variable inequality constraints of third order. *Journal of Optimization Theory & Applications*, 31:85–99, 1980.

44. R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton (NJ), 1970.

45. E. Rosenwasser and R. Yusupov. *Sensitivity of Automatic Control Systems*. CRC Press, Boca Raton (FL), 2000.

46. W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, New York, 3rd edition, 1976.

47. W. Rudin. *Functional Analysis*. McGraw-Hill, New York, 2nd edition, 1991.

48. I. B. Russak. On general problems with bounded state variables. *Journal of Optimization Theory & Applications*, 6(6):424–452, 1970.

49. A. Seierstad and K. Sydsæter. Sufficient conditions in optimal control theory. *International Economic Review*, 18(2):367–391, 1977.

50. H. H. Sohrab. *Basic Real Analysis*. Birkhäuser, Boston (MA), 2003.

51. B. Srinivasan and D. Bonvin. Real-time optimization of batch processes by tracking the necessary conditions of optimality. *Industrial & Engineering Chemistry Research*, 46(2):492–504, 2007.

52. B. Srinivasan, D. Bonvin, E. Visser, and S. Palanki. Dynamic optimization of batch processes: II. Role of measurements in handling uncertainty. *Computers & Chemical Engineering*, 44:27–44, 2003.

53. J. G. Taylor. comments on a multiplier condition for problems with state inequality constraints. *IEEE Transactions on Automatic Control*, AC-12:743–744, 1972.

54. K. L. Teo, C. J. Goh, and K. H. Wong. *A Unified Computational Approach to Optimal Control Problems*. Longman Scientific & Technical, New York, 1991.

55. J. L. Troutman. *Variational Calculus and Optimal Control: Optimization with Elementary Convexity*. Undergraduate Texts in Mathematics. Springer, New York, 2nd edition, 1995.

56. V. S. Vassiliadis, R. W. H. Sargent, and C. C. Pantelides. Solution of a class of multistage dynamic optimization problems. 1. Problems without path constraints. *Industrial & Engineering Chemistry Research*, 33(9):2111–2122, 1994.

57. V. S. Vassiliadis, R. W. H. Sargent, and C. C. Pantelides. Solution of a class of multistage dynamic optimization problems. 2. Problems with path constraints. *Industrial & Engineering Chemistry Research*, 33(9):2123–2133, 1994.

58. M. I. Zelikin and V. F. Borisov. *Theory of Chattering Control*. Birkhäuser, Boston (MA), 1994.