Chapter One

Introduction

1.1 OPTIMAL CONTROL PROBLEM

We begin by describing, very informally and in general terms, the class of optimal control problems that we want to eventually be able to solve. The goal of this brief motivational discussion is to fix the basic concepts and terminology without worrying about technical details.

The first basic ingredient of an optimal control problem is a *control system*. It generates possible behaviors. In this book, control systems will be described by ordinary differential equations (ODEs) of the form

$$\dot{x} = f(t, x, u), \qquad x(t_0) = x_0 \tag{1.1}$$

where x is the *state* taking values in $\mathbb{R}^n$, u is the *control input* taking values in some *control set* $U \subset \mathbb{R}^m$, t is *time*, t_0 is the *initial time*, and x_0 is the *initial state*. Both x and u are functions of t, but we will often suppress their time arguments.

The second basic ingredient is the *cost functional*. It associates a cost with each possible behavior. For a given initial data (t_0, x_0), the behaviors are parameterized by control functions u. Thus, the cost functional assigns a cost value to each admissible control. In this book, cost functionals will be denoted by J and will be of the form

$$J(u) := \int_{t_0}^{t_f} L(t, x(t), u(t))dt + K(t_f, x_f) \tag{1.2}$$

where L and K are given functions (*running cost* and *terminal cost*, respectively), t_f is the *final* (or *terminal*) *time* which is either free or fixed, and $x_f := x(t_f)$ is the *final* (or *terminal*) *state* which is either free or fixed or belongs to some given target set. Note again that u itself is a function of time; this is why we say that J is a *functional* (a real-valued function on a space of functions).

The optimal control problem can then be posed as follows: Find a control u that minimizes $J(u)$ over all admissible controls (or at least over nearby controls). Later we will need to come back to this problem formulation

and fill in some technical details. In particular, we will need to specify what regularity properties should be imposed on the function f and on the admissible controls u to ensure that state trajectories of the control system are well defined. Several versions of the above problem (depending, for example, on the role of the final time and the final state) will be stated more precisely when we are ready to study them. The reader who wishes to preview this material can find it in Section 3.3.

It can be argued that optimality is a universal principle of life, in the sense that many—if not most—processes in nature are governed by solutions to some optimization problems (although we may never know exactly what is being optimized). We will soon see that fundamental laws of mechanics can be cast in an optimization context. From an engineering point of view, optimality provides a very useful design principle, and the cost to be minimized (or the profit to be maximized) is often naturally contained in the problem itself. Some examples of optimal control problems arising in applications include the following:

- Send a rocket to the moon with minimal fuel consumption.
- Produce a given amount of chemical in minimal time and/or with minimal amount of catalyst used (or maximize the amount produced in given time).
- Bring sales of a new product to a desired level while minimizing the amount of money spent on the advertising campaign.
- Maximize throughput or accuracy of information transmission over a communication channel with a given bandwidth/capacity.

The reader will easily think of other examples. Several specific optimal control problems will be examined in detail later in the book. We briefly discuss one simple example here to better illustrate the general problem formulation.

Example 1.1 *Consider a simple model of a car moving on a horizontal line. Let $x \in \mathbb{R}$ be the car's position and let u be the acceleration which acts as the control input. We put a bound on the maximal allowable acceleration by letting the control set U be the bounded interval $[-1, 1]$ (negative acceleration corresponds to braking). The dynamics of the car are $\ddot{x} = u$. In order to arrive at a first-order differential equation model of the form* (1.1)*, let us relabel the car's position x as x_1 and denote its velocity $\dot{x}$ by x_2. This gives the control system $\dot{x}_1 = x_2$, $\dot{x}_2 = u$ with state $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2$. Now, suppose that we want to "park" the car at the origin, i.e., bring it to rest there, in* minimal time. *This objective is captured by the cost functional* (1.2) *with the constant running cost $L \equiv 1$, no terminal cost ($K \equiv 0$), and the fixed final state $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$. We will solve this optimal control problem in Section 4.4.1.*

(The basic form of the optimal control strategy may be intuitively obvious, but obtaining a complete description of the optimal control requires some work.) □

In this book we focus on the *mathematical theory* of optimal control. We will not undertake an in-depth study of any of the applications mentioned above. Instead, we will concentrate on the fundamental aspects common to all of them. After finishing this book, the reader familiar with a specific application domain should have no difficulty reading papers that deal with applications of optimal control theory to that domain, and will be prepared to think creatively about new ways of applying the theory.

We can view the optimal control problem as that of choosing the best *path* among all paths feasible for the system, with respect to the given cost function. In this sense, the problem is *infinite-dimensional*, because the space of paths is an infinite-dimensional function space. This problem is also a *dynamic* optimization problem, in the sense that it involves a dynamical system and time. However, to gain appreciation for this problem, it will be useful to first recall some basic facts about the more standard static finite-dimensional optimization problem, concerned with finding a minimum of a given function $f : \mathbb{R}^n \to \mathbb{R}$. Then, when we get back to infinite-dimensional optimization, we will more clearly see the similarities but also the differences.

The subject studied in this book has a rich and beautiful history; the topics are ordered in such a way as to allow us to trace its chronological development. In particular, we will start with *calculus of variations*, which deals with path optimization but not in the setting of control systems. The optimization problems treated by calculus of variations are infinite-dimensional but not dynamic. We will then make a transition to optimal control theory and develop a truly dynamic framework. This modern treatment is based on two key developments, initially independent but ultimately closely related and complementary to each other: the maximum principle and the principle of dynamic programming.

1.2 SOME BACKGROUND ON FINITE-DIMENSIONAL OPTIMIZATION

Consider a function $f : \mathbb{R}^n \to \mathbb{R}$. Let D be some subset of $\mathbb{R}^n$, which could be the entire $\mathbb{R}^n$. We denote by $|\cdot|$ the standard Euclidean norm on $\mathbb{R}^n$.

A point $x^* \in D$ is a *local minimum* of f over D if there exists an $\varepsilon > 0$ such that for all $x \in D$ satisfying $|x - x^*| < \varepsilon$ we have

$$f(x^*) \leq f(x). \tag{1.3}$$

In other words, x^* is a local minimum if in some ball around it, f does not attain a value smaller than $f(x^*)$. Note that this refers only to points in D;

the behavior of f outside D is irrelevant, and in fact we could have taken the domain of f to be D rather than $\mathbb{R}^n$.

If the inequality in (1.3) is strict for $x \neq x^*$, then we have a *strict* local minimum. If (1.3) holds for *all* $x \in D$, then the minimum is *global* over D. By default, when we say "a minimum" we mean a local minimum. Obviously, a minimum need not be unique unless it is both strict and global.

The notions of a (local, strict, global) *maximum* are defined similarly. If a point is either a maximum or a minimum, it is called an *extremum*. Observe that maxima of f are minima of $-f$, so there is no need to develop separate results for both. We focus on the minima, i.e., we view f as a *cost* function to be minimized (rather than a profit to be maximized).

1.2.1 Unconstrained optimization

The term "unconstrained optimization" usually refers to the situation where all points x sufficiently near x^* in $\mathbb{R}^n$ are in D, i.e., x^* belongs to D together with some $\mathbb{R}^n$-neighborhood. The simplest case is when $D = \mathbb{R}^n$, which is sometimes called the *completely unconstrained* case. However, as far as *local* minimization is concerned, it is enough to assume that x^* is an interior point of D. This is automatically true if D is an open subset of $\mathbb{R}^n$.

First-order necessary condition for optimality

Suppose that f is a $\mathcal{C}^1$ (continuously differentiable) function and x^* is its local minimum. Pick an arbitrary vector $d \in \mathbb{R}^n$. Since we are in the unconstrained case, moving away from x^* in the direction of d or $-d$ cannot immediately take us outside D. In other words, we have $x^* + \alpha d \in D$ for all $\alpha \in \mathbb{R}$ close enough to 0.

For a fixed d, we can consider $f(x^* + \alpha d)$ as a function of the real parameter α, whose domain is some interval containing 0. Let us call this new function g:

$$g(\alpha) := f(x^* + \alpha d). \tag{1.4}$$

Since x^* is a minimum of f, it is clear that 0 is a minimum of g. Passing from f to g is useful because g is a function of a scalar variable and so its minima can be studied using ordinary calculus. In particular, we can write down the first-order Taylor expansion for g around $\alpha = 0$:

$$g(\alpha) = g(0) + g'(0)\alpha + o(\alpha) \tag{1.5}$$

where $o(\alpha)$ represents "higher-order terms" which go to 0 faster than α as α approaches 0, i.e.,

$$\lim_{\alpha \to 0} \frac{o(\alpha)}{\alpha} = 0. \tag{1.6}$$

We claim that

$$g'(0) = 0. \tag{1.7}$$

To show this, suppose that $g'(0) \neq 0$. Then, in view of (1.6), there exists an $\varepsilon > 0$ small enough so that for all nonzero α with $|\alpha| < \varepsilon$, the absolute value of the fraction in (1.6) is less than $|g'(0)|$. We can write this as

$$|\alpha| < \varepsilon,\ \alpha \neq 0 \quad \Rightarrow \quad |o(\alpha)| < |g'(0)\alpha|.$$

For these values of α, (1.5) gives

$$g(\alpha) - g(0) < g'(0)\alpha + |g'(0)\alpha|. \tag{1.8}$$

If we further restrict α to have the opposite sign to $g'(0)$, then the right-hand side of (1.8) becomes 0 and we obtain $g(\alpha) - g(0) < 0$. But this contradicts the fact that g has a minimum at 0. We have thus shown that (1.7) is indeed true.

We now need to re-express this result in terms of the original function f. A simple application of the chain rule from vector calculus yields the formula

$$g'(\alpha) = \nabla f(x^* + \alpha d) \cdot d \tag{1.9}$$

where

$$\nabla f := (f_{x_1}, \ldots, f_{x_n})^T$$

is the *gradient* of f and $\cdot$ denotes inner product.[1] Whenever there is no danger of confusion, we use subscripts as a shorthand notation for partial derivatives: $f_{x_i} := \partial f / \partial x_i$. Setting $\alpha = 0$ in (1.9), we have

$$g'(0) = \nabla f(x^*) \cdot d \tag{1.10}$$

and this equals 0 by (1.7). Since d was arbitrary, we conclude that

$$\boxed{\nabla f(x^*) = 0} \tag{1.11}$$

This is the **first-order necessary condition for optimality**.

A point x^* satisfying this condition is called a *stationary point*. The condition is "first-order" because it is derived using the first-order expansion (1.5). We emphasize that the result is valid when $f \in \mathcal{C}^1$ and x^* is an interior point of D.

[1] There is no consensus in the literature whether the gradient is a column vector or a row vector. Treating it as a row vector would simplify the notation since it often appears in a product with another vector. Geometrically, however, it plays the role of a regular column vector, and for consistency we follow this latter convention everywhere.

Second-order conditions for optimality

We now derive another necessary condition and also a sufficient condition for optimality, under the stronger hypothesis that f is a $\mathcal{C}^2$ function (twice continuously differentiable).

First, we assume as before that x^* is a local minimum and derive a necessary condition. For an arbitrary fixed $d \in \mathbb{R}^n$, let us consider a Taylor expansion of $g(\alpha) = f(x^* + \alpha d)$ again, but this time include *second-order terms*:

$$g(\alpha) = g(0) + g'(0)\alpha + \frac{1}{2}g''(0)\alpha^2 + o(\alpha^2) \tag{1.12}$$

where

$$\lim_{\alpha \to 0} \frac{o(\alpha^2)}{\alpha^2} = 0. \tag{1.13}$$

We know from the derivation of the first-order necessary condition that $g'(0)$ must be 0. We claim that

$$g''(0) \geq 0. \tag{1.14}$$

Indeed, suppose that $g''(0) < 0$. By (1.13), there exists an $\varepsilon > 0$ such that

$$|\alpha| < \varepsilon,\ \alpha \neq 0 \quad \Rightarrow \quad |o(\alpha^2)| < \frac{1}{2}|g''(0)|\alpha^2.$$

For these values of α, (1.12) reduces to $g(\alpha) - g(0) < 0$, contradicting that fact that 0 is a minimum of g. Therefore, (1.14) must hold.

What does this result imply about the original function f? To see what $g''(0)$ is in terms of f, we need to differentiate the formula (1.9). The reader may find it helpful to first rewrite (1.9) more explicitly as

$$g'(\alpha) = \sum_{i=1}^{n} f_{x_i}(x^* + \alpha d)d_i.$$

Differentiating both sides with respect to α, we have

$$g''(\alpha) = \sum_{i,j=1}^{n} f_{x_i x_j}(x^* + \alpha d)d_i d_j$$

where double subscripts are used to denote second-order partial derivatives. For $\alpha = 0$ this gives

$$g''(0) = \sum_{i,j=1}^{n} f_{x_i x_j}(x^*)d_i d_j$$

or, in matrix notation,

$$g''(0) = d^T \nabla^2 f(x^*) d \tag{1.15}$$

where

$$\nabla^2 f := \begin{pmatrix} f_{x_1x_1} & \cdots & f_{x_1x_n} \\ \vdots & \ddots & \vdots \\ f_{x_nx_1} & \cdots & f_{x_nx_n} \end{pmatrix}$$

is the *Hessian* matrix of f. In view of (1.14), (1.15), and the fact that d was arbitrary, we conclude that the matrix $\nabla^2 f(x^*)$ must be positive semidefinite:

$$\boxed{\nabla^2 f(x^*) \geq 0} \quad \text{(positive semidefinite)}$$

This is the **second-order necessary condition for optimality**.

Like the previous first-order necessary condition, this second-order condition only applies to the unconstrained case. But, unlike the first-order condition, it requires f to be $\mathcal{C}^2$ and not just $\mathcal{C}^1$. Another difference with the first-order condition is that the second-order condition distinguishes minima from maxima: at a local maximum, the Hessian must be *negative* semidefinite, while the first-order condition applies to any extremum (a minimum or a maximum).

Strengthening the second-order necessary condition and combining it with the first-order necessary condition, we can obtain the following **second-order sufficient condition for optimality**: *If a $\mathcal{C}^2$ function f satisfies*

$$\nabla f(x^*) = 0 \qquad \text{and} \qquad \nabla^2 f(x^*) > 0 \ \ \text{(positive } \textit{definite}\text{)} \tag{1.16}$$

on an interior point x^* of its domain, then x^* is a strict local minimum of f. To see why this is true, take an arbitrary $d \in \mathbb{R}^n$ and consider again the second-order expansion (1.12) for $g(\alpha) = f(x^* + \alpha d)$. We know that $g'(0)$ is given by (1.10), thus it is 0 because $\nabla f(x^*) = 0$. Next, $g''(0)$ is given by (1.15), and so we have

$$f(x^* + \alpha d) = f(x^*) + \frac{1}{2} d^T \nabla^2 f(x^*) d\alpha^2 + o(\alpha^2). \tag{1.17}$$

The intuition is that since the Hessian $\nabla^2 f(x^*)$ is a positive definite matrix, the second-order term dominates the higher-order term $o(\alpha^2)$. To establish this fact rigorously, note that by the definition of $o(\alpha^2)$ we can pick an $\varepsilon > 0$ small enough so that

$$|\alpha| < \varepsilon, \ \alpha \neq 0 \quad \Rightarrow \quad |o(\alpha^2)| < \frac{1}{2} d^T \nabla^2 f(x^*) d\alpha^2$$

and for these values of α we deduce from (1.17) that $f(x^* + \alpha d) > f(x^*)$.

To conclude that x^* is a (strict) local minimum, one more technical detail is needed. According to the definition of a local minimum (see page 3), we must show that $f(x^*)$ is the lowest value of f in some ball around x^*. But the term $o(\alpha^2)$ and hence the value of ε in the above construction depend on

the choice of the direction d. It is clear that this dependence is continuous, since all the other terms in (1.17) are continuous in d.[2] Also, without loss of generality we can restrict d to be of unit length, and then we can take the minimum of ε over all such d. Since the unit sphere in $\mathbb{R}^n$ is compact, the minimum is well defined (thanks to the Weierstrass Theorem which is discussed below). This minimal value of ε provides the radius of the desired ball around x^* in which the lowest value of f is achieved at x^*.

Feasible directions, global minima, and convex problems

The key fact that we used in the previous developments was that for every $d \in \mathbb{R}^n$, points of the form $x^* + \alpha d$ for α sufficiently close to 0 belong to D. This is no longer the case if D has a boundary (e.g., D is a closed ball in $\mathbb{R}^n$) and x^* is a point on this boundary. Such situations do not fit into the unconstrained optimization scenario as we defined it at the beginning of Section 1.2.1; however, for simple enough sets D and with some extra care, a similar analysis is possible. Let us call a vector $d \in \mathbb{R}^n$ a *feasible direction* (at x^*) if $x^* + \alpha d \in D$ for small enough $\alpha > 0$ (see Figure 1.1). If not all directions d are feasible, then the condition $\nabla f(x^*) = 0$ is no longer necessary for optimality. We can still define the function (1.4) for every feasible direction d, but the proof of (1.7) is no longer valid because α is now nonnegative. We leave it to the reader to modify that argument and show that if x^* is a local minimum, then $\nabla f(x^*) \cdot d \geq 0$ for every feasible direction d. As for the second-order necessary condition, the inequality (1.14) is still true if $g'(0) = 0$, which together with (1.10) and (1.15) implies that we must have $d^T \nabla^2 f(x^*) d \geq 0$ for all feasible directions satisfying $\nabla f(x^*) \cdot d = 0$.

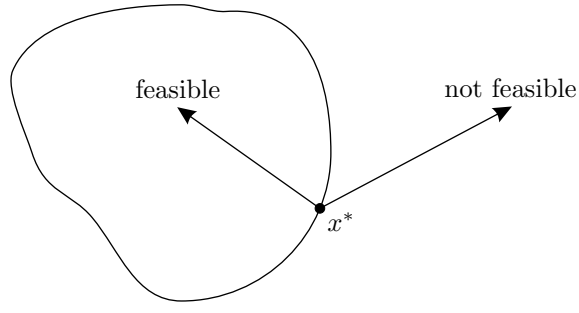

Figure 1.1: Feasible directions

If the set D is *convex*, then the line segment connecting x^* to an arbitrary other point $x \in D$ lies entirely in D. All points on this line segment take the

[2] The term $o(\alpha^2)$ can be described more precisely using Taylor's theorem with remainder, which is a higher-order generalization of the Mean Value Theorem; see, e.g., [Rud76, Theorem 5.15]. We will discuss this issue in more detail later when deriving the corresponding result in calculus of variations (see Section 2.6).

form $x^* + \alpha d$, $\alpha \in [0, \bar{\alpha}]$ for some $d \in \mathbb{R}^n$ and $\bar{\alpha} > 0$. This means that the feasible direction approach is particularly suitable for the case of a convex D. But if D is not convex, then the first-order and second-order necessary conditions in terms of feasible directions are conservative. The next exercise touches on the issue of sufficiency.

Exercise 1.1 *Suppose that f is a $\mathcal{C}^2$ function and x^* is a point of its domain at which we have $\nabla f(x^*) \cdot d \geq 0$ and $d^T \nabla^2 f(x^*) d > 0$ for every nonzero feasible direction d. Is x^* necessarily a local minimum of f? Prove or give a counterexample.* □

When we are not dealing with the completely unconstrained case in which D is the entire $\mathbb{R}^n$, we think of D as the constraint set over which f is being minimized. Particularly important in optimization theory is the case when equality constraints are present, so that D is a lower-dimensional surface in $\mathbb{R}^n$ (see Figure 1.2). In such situations, the above method which utilizes feasible directions represented by straight lines is no longer suitable: there might not be any feasible directions, and then the corresponding necessary conditions are vacuous. We will describe a refined approach to constrained optimization in Section 1.2.2; it essentially replaces straight lines with arbitrary curves.

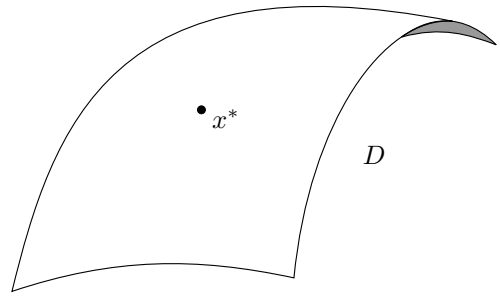

Figure 1.2: A surface

So far we have only discussed local minima. In practice, however, one is typically interested in finding a global minimum over a given domain (or constraint set) D, if such a global minimum exists. We now briefly discuss how conditions for local optimality can be useful for solving global optimization problems as well, provided that these global problems have certain nice features.

The following basic existence result is known as the **Weierstrass Theorem**: *If f is a continuous function and D is a compact set, then there exists a global minimum of f over D.* The reader will recall that for subsets of $\mathbb{R}^n$, compactness can be defined in three equivalent ways:

1) D is compact if it is closed and bounded.

2) D is compact if every open cover of D has a finite subcover.

3) D is compact if every sequence in D has a subsequence converging to some point in D (sequential compactness).

We will revisit compactness and the Weierstrass Theorem in the infinite-dimensional optimization setting.

The necessary conditions for local optimality that we discussed earlier suggest the following procedure for finding a global minimum. First, find all interior points of D satisfying $\nabla f(x^*) = 0$ (the stationary points). If f is not differentiable everywhere, include also points where ∇f does not exist (these points together with the stationary points comprise the *critical points*). Next, find all boundary points satisfying $\nabla f(x^*) \cdot d \geq 0$ for all feasible d. Finally, compare values at all these candidate points and choose the smallest one. If one can afford the computation of second derivatives, then the second-order conditions can be used in combination with the first-order ones.

If D is a convex set and f is a convex function, then the minimization problem is particularly tractable. First, a local minimum is automatically a global one. Second, the first-order necessary condition (for $f \in \mathcal{C}^1$) is also a sufficient condition. Thus if $\nabla f(x^*) \cdot d \geq 0$ for all feasible directions d, or in particular if x^* is an interior point of D and $\nabla f(x^*) = 0$, then x^* is a global minimum. These properties are consequences of the fact (illustrated in Figure 1.3) that the graph of a convex function f lies above that of the linear approximation $x \mapsto f(x^*) + \nabla f(x^*) \cdot (x - x^*)$.

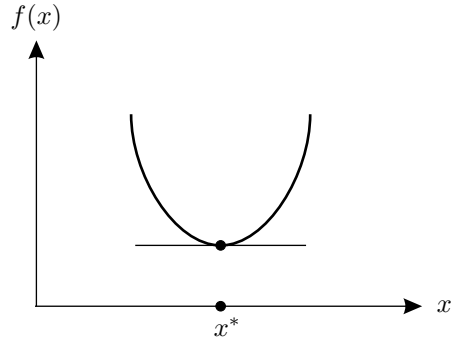

Figure 1.3: A convex function

Efficient numerical algorithms—such as the well-known steepest descent (or gradient) method—exist for converging to points satisfying $\nabla f(x^*) = 0$ (stationary points). For convex problems, these algorithms yield convergence to global minima.

1.2.2 Constrained optimization

Now suppose that D is a surface in $\mathbb{R}^n$ defined by the *equality constraints*

$$h_1(x) = h_2(x) = \cdots = h_m(x) = 0 \tag{1.18}$$

where h_i, $i = 1, \ldots, m$ are $\mathcal{C}^1$ functions from $\mathbb{R}^n$ to $\mathbb{R}$. We assume that f is a $\mathcal{C}^1$ function and study its minima over D.

First-order necessary condition (Lagrange multipliers)

Let $x^* \in D$ be a local minimum of f over D. We assume that x^* is a *regular point* of D in the sense that the gradients ∇h_i, $i = 1, \ldots, m$ are linearly independent at x^*. This is a technical assumption needed to rule out degenerate situations; see Exercise 1.2 below.

Instead of line segments containing x^* which we used in the unconstrained case, we now consider *curves* in D passing through x^*. Such a curve is a family of points $x(\alpha) \in D$ parameterized by $\alpha \in \mathbb{R}$, with $x(0) = x^*$. We require the function $x(\cdot)$ to be $\mathcal{C}^1$, at least for α near 0. Given an arbitrary curve of this kind, we can consider the function

$$g(\alpha) := f(x(\alpha)).$$

Note that when there are no equality constraints, functions of the form (1.4) considered previously can be viewed as special cases of this more general construction. From the fact that 0 is a minimum of g, we derive exactly as before that (1.7) holds, i.e., $g'(0) = 0$. To interpret this result in terms of f, note that

$$g'(\alpha) = \nabla f(x(\alpha)) \cdot x'(\alpha)$$

which for $\alpha = 0$ gives

$$g'(0) = \nabla f(x^*) \cdot x'(0) = 0. \tag{1.19}$$

The vector $x'(0) \in \mathbb{R}^n$ is an important object for us here. From the first-order Taylor expansion $x(\alpha) = x^* + x'(0)\alpha + o(\alpha)$ we see that $x'(0)$ defines a linear approximation of $x(\cdot)$ at x^*. Geometrically, it specifies the infinitesimal direction of the curve (see Figure 1.4). The vector $x'(0)$ is a *tangent vector* to D at x^*. It lives in the *tangent space* to D at x^*, which is denoted by $T_{x^*}D$. (We can think of this space as having its origin at x^*.)

We want to have a more explicit characterization of the tangent space $T_{x^*}D$, which will help us understand it better. Since D was defined as the set of points satisfying the equalities (1.18), and since the points $x(\alpha)$ lie in D by construction, we must have $h_i(x(\alpha)) = 0$ for all α and all $i \in \{1, \ldots, m\}$. Differentiating this formula gives

$$0 = \frac{d}{d\alpha} h_i(x(\alpha)) = \nabla h_i(x(\alpha)) \cdot x'(\alpha), \qquad i = 1, \ldots, m$$

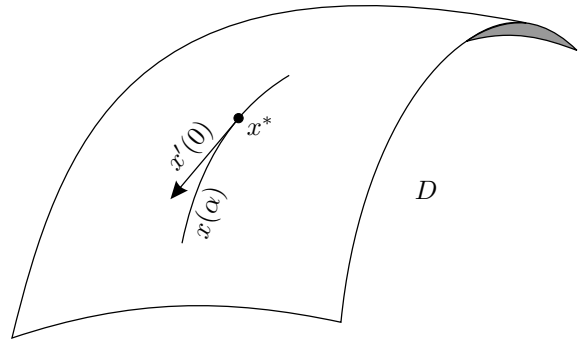

Figure 1.4: A tangent vector

for all α (close enough to 0). Setting $\alpha = 0$ and remembering that $x(0) = x^*$, we obtain

$$0 = \frac{d}{d\alpha}\bigg|_{\alpha=0} h_i(x(\alpha)) = \nabla h_i(x^*) \cdot x'(0), \qquad i = 1, \ldots, m.$$

We have shown that for an arbitrary $\mathcal{C}^1$ curve $x(\cdot)$ in D with $x(0) = x^*$, its tangent vector $x'(0)$ must satisfy $\nabla h_i(x^*) \cdot x'(0) = 0$ for each i. Actually, one can show that the converse is also true, namely, every vector $d \in \mathbb{R}^n$ satisfying

$$\nabla h_i(x^*) \cdot d = 0, \qquad i = 1, \ldots, m \tag{1.20}$$

is a tangent vector to D at x^* corresponding to some curve. (We do not give a proof of this fact but note that it relies on x^* being a regular point of D.) In other words, the tangent vectors to D at x^* are exactly the vectors d for which (1.20) holds. This is the characterization of the tangent space $T_{x^*}D$ that we were looking for. It is clear from (1.20) that $T_{x^*}D$ is a subspace of $\mathbb{R}^n$; in particular, if d is a tangent vector, then so is $-d$ (going from $x'(0)$ to $-x'(0)$ corresponds to reversing the direction of the curve).

Now let us go back to (1.19), which tells us that $\nabla f(x^*) \cdot d = 0$ for all $d \in T_{x^*}D$ (since the curve $x(\cdot)$ and thus the tangent vector $x'(0)$ were arbitrary). In view of the characterization of $T_{x^*}D$ given by (1.20), we can rewrite this condition as follows:

$$\nabla f(x^*) \cdot d = 0 \qquad \forall\, d \;\text{ such that }\; \nabla h_i(x^*) \cdot d = 0,\; i = 1, \ldots, m. \tag{1.21}$$

The relation between $\nabla f(x^*)$ and $\nabla h_i(x^*)$ expressed by (1.21) looks somewhat clumsy, since checking it involves a search over d. Can we eliminate d from this relation and make it more explicit? A careful look at (1.21) should quickly lead the reader to the following statement.

Claim: The gradient of f at x^* is a linear combination of the gradients of the constraint functions $h_1, \ldots, h_m$ at x^*:

$$\nabla f(x^*) \in \operatorname{span}\{\nabla h_i(x^*),\, i = 1, \ldots, m\}. \tag{1.22}$$

Indeed, if the claim were not true, then $\nabla f(x^*)$ would have a component orthogonal to $\text{span}\{\nabla h_i(x^*)\}$, i.e, there would exist a $d \neq 0$ satisfying (1.20) such that $\nabla f(x^*)$ can be written in the form

$$\nabla f(x^*) = d - \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*) \tag{1.23}$$

for some $\lambda_1^*, \ldots, \lambda_m^* \in \mathbb{R}$. Taking the inner product with d on both sides of (1.23) and using (1.20) gives

$$\nabla f(x^*) \cdot d = d \cdot d \neq 0$$

and we reach a contradiction with (1.21).

Geometrically, the claim says that $\nabla f(x^*)$ is normal to D at x^*. This situation is illustrated in Figure 1.5 for the case of two constraints in $\mathbb{R}^3$. Note that if there is only one constraint, say $h_1(x) = 0$, then D is a two-dimensional surface and $\nabla f(x^*)$ must be proportional to $\nabla h_1(x^*)$, the normal direction to D at x^*. When the second constraint $h_2(x) = 0$ is added, D becomes a curve (the thick curve in the figure) and $\nabla f(x^*)$ is allowed to live in the plane spanned by $\nabla h_1(x^*)$ and $\nabla h_2(x^*)$, i.e., the normal plane to D at x^*. In general, the intuition behind the claim is that unless $\nabla f(x^*)$ is normal to D, there are curves in D passing through x^* whose tangent vectors at x^* make both positive and negative inner products with $\nabla f(x^*)$, hence in particular f can be decreased by moving away from x^* while staying in D.

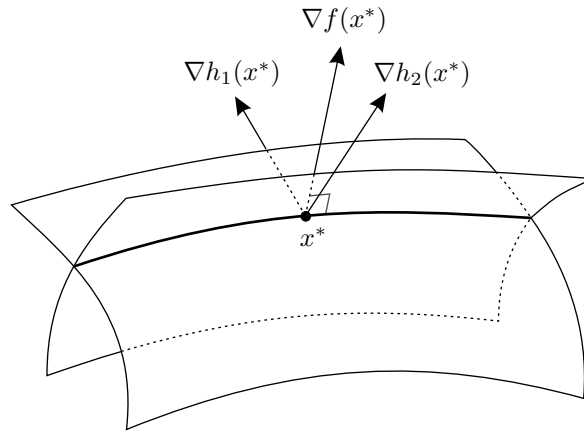

Figure 1.5: Gradient vectors and constrained optimality

The condition (1.22) means that there exist real numbers $\lambda_1^*, \ldots, \lambda_m^*$ such that

$$\boxed{\nabla f(x^*) + \lambda_1^* \nabla h_1(x^*) + \cdots + \lambda_m^* \nabla h_m(x^*) = 0} \tag{1.24}$$

This is the **first-order necessary condition for constrained optimality**. The coefficients λ_i^*, $i = 1, \ldots, m$ are called *Lagrange multipliers.*

Exercise 1.2 *Give an example where a local minimum x^* is not a regular point and the above necessary condition is false (be sure to justify both of these claims).* □

The above proof of the first-order necessary condition for constrained optimality involves geometric concepts. We also left a gap in it because we did not prove the converse implication in the equivalent characterization of the tangent space given by (1.20). We now give a shorter alternative proof which is purely analytic, and which will be useful when we study problems with constraints in calculus of variations. However, the geometric intuition behind the previous proof will be helpful for us later as well. We invite the reader to study both proofs as a way of testing the mathematical background knowledge that will be required in the subsequent chapters.

Let us start again by assuming that x^* is a local minimum of f over D, where D is a surface in $\mathbb{R}^n$ defined by the equality constraints (1.18) and x^* is a regular point of D. Our goal is to rederive the necessary condition expressed by (1.24). For simplicity, we only give the argument for the case of a single constraint $h(x) = 0$, i.e., $m = 1$; the extension to $m > 1$ is straightforward (see Exercise 1.3 below). Given two arbitrary vectors $d_1, d_2 \in \mathbb{R}^n$, we can consider the following map from $\mathbb{R} \times \mathbb{R}$ to itself:

$$F : (\alpha_1, \alpha_2) \mapsto \big(f(x^* + \alpha_1 d_1 + \alpha_2 d_2), h(x^* + \alpha_1 d_1 + \alpha_2 d_2)\big).$$

The Jacobian matrix of F at $(0, 0)$ is

$$\begin{pmatrix} \nabla f(x^*) \cdot d_1 & \nabla f(x^*) \cdot d_2 \\ \nabla h(x^*) \cdot d_1 & \nabla h(x^*) \cdot d_2 \end{pmatrix}. \tag{1.25}$$

If this Jacobian matrix were nonsingular, then we could apply the Inverse Function Theorem (see, e.g., [Rud76, Theorem 9.24]) and conclude that there are neighborhoods of $(0, 0)$ and $F(0, 0) = (f(x^*), 0)$ on which the map F is a bijection (has an inverse). This would imply, in particular, that there are points x arbitrarily close to x^* such that $h(x) = 0$ and $f(x) < f(x^*)$; such points would be obtained by taking preimages of points on the ray directed to the left from $F(0, 0)$ in Figure 1.6. But this cannot be true, since $h(x) = 0$ means that $x \in D$ and we know that x^* is a local minimum of f over D. Therefore, the matrix (1.25) is singular.

Regularity of x^* in the present case just means that the gradient $\nabla h(x^*)$ is nonzero. Choose a d_1 such that $\nabla h(x^*) \cdot d_1 \neq 0$. With this d_1 fixed, let $\lambda^* := -(\nabla f(x^*) \cdot d_1)/(\nabla h(x^*) \cdot d_1)$, so that $\nabla f(x^*) \cdot d_1 = -\lambda^* \nabla h(x^*) \cdot d_1$. Since the matrix (1.25) must be singular for all choices of d_2, its first row must be a constant multiple of its second row (the second row being nonzero

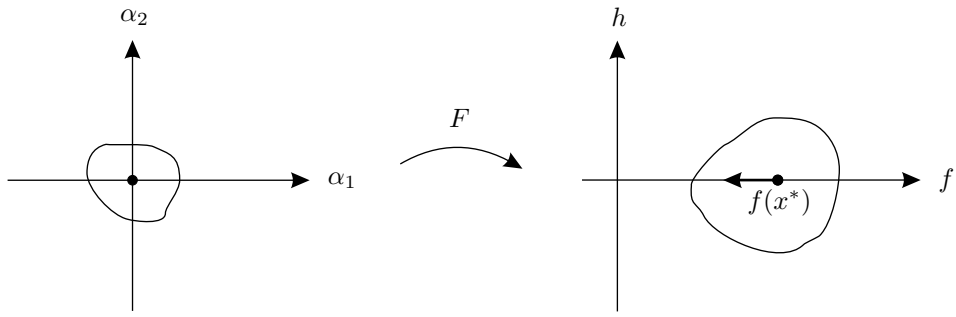

Figure 1.6: Illustrating the alternative proof

by our choice of d_1). Thus we have $\nabla f(x^*) \cdot d_2 = -\lambda^* \nabla h(x^*) \cdot d_2$, or $(\nabla f(x^*) + \lambda^* \nabla h(x^*)) \cdot d_2 = 0$, and this must be true for all $d_2 \in \mathbb{R}^n$. It follows that $\nabla f(x^*) + \lambda^* \nabla h(x^*) = 0$, which proves (1.24) for the case when $m = 1$.

Exercise 1.3 *Generalize the previous argument to an arbitrary number $m \geq 1$ of equality constraints (still assuming that x^* is a regular point).* □

The first-order necessary condition for constrained optimality generalizes the corresponding result we derived earlier for the unconstrained case. The condition (1.24) together with the constraints (1.18) is a system of $n + m$ equations in $n + m$ unknowns: n components of x^* plus m components of the Lagrange multiplier vector $\lambda^* = (\lambda_1^*, \ldots, \lambda_m^*)^T$. For $m = 0$, we recover the condition (1.11) which consists of n equations in n unknowns. To make this relation even more explicit, consider the function $\ell : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ defined by

$$\ell(x, \lambda) := f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) \tag{1.26}$$

which we call the *augmented cost* function. If x^* is a local constrained minimum of f and λ^* is the corresponding vector of Lagrange multipliers for which (1.24) holds, then the gradient of ℓ at (x^*, λ^*) satisfies

$$\nabla \ell(x^*, \lambda^*) = \begin{pmatrix} \ell_x(x^*, \lambda^*) \\ \ell_\lambda(x^*, \lambda^*) \end{pmatrix} = \begin{pmatrix} \nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*) \\ h(x^*) \end{pmatrix} = 0 \tag{1.27}$$

where ℓ_x, ℓ_λ are the vectors of partial derivatives of ℓ with respect to the components of x and λ, respectively, and $h = (h_1, \ldots, h_m)^T$ is the vector of constraint functions. We conclude that (x^*, λ^*) is a usual (unconstrained) stationary point of the augmented cost ℓ. Loosely speaking, adding Lagrange multipliers converts a constrained problem into an unconstrained one, and the first-order necessary condition (1.24) for constrained optimality is recovered from the first-order necessary condition for unconstrained optimality applied to ℓ.

The idea of passing from constrained minimization of the original cost function to unconstrained minimization of the augmented cost function is due to Lagrange. If (x^*, λ^*) is a minimum of ℓ, then we must have $h(x^*) = 0$ (because otherwise we could decrease ℓ by changing λ^*), and subject to these constraints x^* should minimize f (because otherwise it would not minimize ℓ). Also, it is clear that (1.27) must hold. However, it does *not* follow that (1.27) is a necessary condition for x^* to be a constrained minimum of f. Unfortunately, there is no quick way to derive the first-order necessary condition for constrained optimality by working with the augmented cost—something that Lagrange originally attempted to do. Nevertheless, the basic form of the augmented cost function (1.26) is fundamental in constrained optimization theory, and will reappear in various forms several times in this book.

Even though the condition in terms of Lagrange multipliers is only necessary and not sufficient for constrained optimality, it is very useful for narrowing down candidates for local extrema. The next exercise illustrates this point for a well-known optimization problem arising in optics.

Exercise 1.4 *Consider a curve D in the plane described by the equation $h(x) = 0$, where $h : \mathbb{R}^2 \to \mathbb{R}$ is a $\mathcal{C}^1$ function. Let y and z be two fixed points in the plane, lying on the same side with respect to D (but not on D itself). Suppose that a ray of light emanates from y, gets reflected off D at some point $x^* \in D$, and arrives at z. Consider the following two statements: (i) x^* must be such that the total Euclidean distance traveled by light to go from y to z is minimized over all nearby candidate reflection points $x \in D$ (Fermat's principle); (ii) the angles that the light ray makes with the line normal to D at x^* before and after the reflection must be the same (the law of reflection). Accepting the first statement as a hypothesis, prove that the second statement follows from it, with the help of the first-order necessary condition for constrained optimality* (1.24). □

Second-order conditions

For the sake of completeness, we quickly state the second-order conditions for constrained optimality; they will not be used in the sequel. For the necessary condition, suppose that x^* is a regular point of D and a local minimum of f over D, where D is defined by the equality constraints (1.18) as before. We let λ^* be the vector of Lagrange multipliers provided by the first-order necessary condition, and define the augmented cost ℓ as in (1.26). We also assume that f is $\mathcal{C}^2$. Consider the Hessian of ℓ with respect to x evaluated at (x^*, λ^*):

$$\ell_{xx}(x^*, \lambda^*) = \nabla^2 f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 h_i(x^*).$$

The second-order necessary condition says that this Hessian matrix must be positive semidefinite on the tangent space to D at x^*, i.e., we must have $d^T \ell_{xx}(x^*, \lambda^*) d \geq 0$ for all $d \in T_{x^*} D$. Note that this is weaker than asking the above Hessian matrix to be positive semidefinite in the usual sense (on the entire $\mathbb{R}^n$).

The second-order sufficient condition says that a point $x^* \in D$ is a strict constrained local minimum of f if the first-order necessary condition for constrained optimality (1.24) holds and, in addition, we have

$$d^T \ell_{xx}(x^*, \lambda^*) d > 0 \qquad \forall\, d \text{ such that } \nabla h_i(x^*) \cdot d = 0,\ i = 1, \ldots, m. \quad (1.28)$$

Again, here λ^* is the vector of Lagrange multipliers and ℓ is the corresponding augmented cost. Note that regularity of x^* is not needed for this sufficient condition to be true. If x^* is in fact a regular point, then we know (from our derivation of the first-order necessary condition for constrained optimality) that the condition imposed on d in (1.28) describes exactly the tangent vectors to D at x^*. In other words, in this case (1.28) is equivalent to saying that $\ell_{xx}(x^*, \lambda^*)$ is positive definite on the tangent space $T_{x^*} D$.

1.3 PREVIEW OF INFINITE-DIMENSIONAL OPTIMIZATION

In Section 1.2 we considered the problem of minimizing a function $f : \mathbb{R}^n \to \mathbb{R}$. Now, instead of $\mathbb{R}^n$ we want to allow a general vector space V, and in fact we are interested in the case when this vector space V is infinite-dimensional. Specifically, V will itself be a space of functions. Let us denote a generic function in V by y, reserving the letter x for the argument of y. (This x will typically be a scalar, and has no relation with $x \in \mathbb{R}^n$ from the previous section.) The function to be minimized is a real-valued function on V, which we now denote by J. Since J is a function on a space of functions, it is called a *functional*. To summarize, we are minimizing a functional $J : V \to \mathbb{R}$.

Unlike in the case of $\mathbb{R}^n$, there does not exist a "universal" function space. Many different choices for V are possible, and specifying the desired space V is part of the problem formulation. Another issue is that in order to define *local* minima of J over V, we need to specify what it means for two functions in V to be close to each other. Recall that in the definition of a local minimum in Section 1.2, a ball of radius ε with respect to the standard Euclidean norm on $\mathbb{R}^n$ was used to define the notion of closeness. In the present case we will again employ ε-balls, but we need to specify which norm we are going to use. While in $\mathbb{R}^n$ all norms are equivalent (i.e., are within a constant multiple of one another), in function spaces different choices of a norm lead to drastically different notions of closeness. Thus, the first thing we need to do is become more familiar with function spaces and norms on them.

1.3.1 Function spaces, norms, and local minima

Typical function spaces that we will consider are spaces of functions from some interval $[a, b]$ to $\mathbb{R}^n$ (for some $n \geq 1$). Different spaces result from placing different requirements on the regularity of these functions. For example, we will frequently work with the function space $\mathcal{C}^k([a, b], \mathbb{R}^n)$, whose elements are k-times continuously differentiable (here $k \geq 0$ is an integer; for $k = 0$ the functions are just continuous). Relaxing the $\mathcal{C}^k$ assumption, we can arrive at the spaces of piecewise continuous functions or even measurable functions (we will define these more precisely later when we need them). On the other hand, stronger regularity assumptions lead us to $\mathcal{C}^\infty$ (smooth, or infinitely many times differentiable) functions or to real analytic functions (the latter are $\mathcal{C}^\infty$ functions that agree with their Taylor series around every point).

We regard these function spaces as linear vector spaces over $\mathbb{R}$. Why are they infinite-dimensional? One way to see this is to observe that the monomials $1, x, x^2, x^3, \ldots$ are linearly independent. Another example of an infinite set of linearly independent functions is provided by the (trigonometric) Fourier basis.

As we already mentioned, we also need to equip our function space V with a *norm* $\|\cdot\|$. This is a real-valued function on V which is positive definite ($\|y\| > 0$ if $y \not\equiv 0$), homogeneous ($\|\lambda y\| = |\lambda| \cdot \|y\|$ for all $\lambda \in \mathbb{R}$, $y \in V$), and satisfies the triangle inequality ($\|y + z\| \leq \|y\| + \|z\|$). The norm gives us the notion of a *distance*, or *metric*, $d(y, z) := \|y - z\|$. This allows us to define local minima and enables us to talk about topological concepts such as convergence and continuity (more on this in Section 1.3.4 below). We will see how the norm plays a crucial role in the subsequent developments.

On the space $\mathcal{C}^0([a, b], \mathbb{R}^n)$, a commonly used norm is

$$\|y\|_0 := \max_{a \leq x \leq b} |y(x)| \tag{1.29}$$

where $|\cdot|$ is the standard Euclidean norm on $\mathbb{R}^n$ as before. Replacing the maximum by a supremum, we can extend the 0-norm (1.29) to functions that are defined over an infinite interval or are not necessarily continuous. On $\mathcal{C}^1([a, b], \mathbb{R}^n)$, another natural candidate for a norm is obtained by adding the 0-norms of y and its first derivative:

$$\|y\|_1 := \max_{a \leq x \leq b} |y(x)| + \max_{a \leq x \leq b} |y'(x)|. \tag{1.30}$$

This construction can be continued in the obvious way to yield the k-norm on $\mathcal{C}^k([a, b], \mathbb{R}^n)$ for each k. The k-norm can also be used on $\mathcal{C}^\ell([a, b], \mathbb{R}^n)$ for all $\ell \geq k$. There exist many other norms, such as for example the $\mathcal{L}_p$

norm

$$\|y\|_{\mathcal{L}_p} := \left(\int_a^b |y(x)|^p dx \right)^{1/p} \tag{1.31}$$

where p is a positive integer. In fact, the 0-norm (1.29) is also known as the $\mathcal{L}_\infty$ norm.

We are now ready to formally define local minima of a functional. Let V be a vector space of functions equipped with a norm $\|\cdot\|$, let A be a subset of V, and let J be a real-valued functional defined on V (or just on A). A function $y^* \in A$ is a *local minimum* of J over A if there exists an $\varepsilon > 0$ such that for all $y \in A$ satisfying $\|y - y^*\| < \varepsilon$ we have

$$J(y^*) \leq J(y).$$

Note that this definition of a local minimum is completely analogous to the one in the previous section, modulo the change of notation $x \mapsto y$, $D \mapsto A$, $f \mapsto J$, $|\cdot| \mapsto \|\cdot\|$ (also, implicitly, $\mathbb{R}^n \mapsto V$). Strict minima, global minima, and the corresponding notions of maxima are defined in the same way as before. We will continue to refer to minima and maxima collectively as *extrema*.

For the norm $\|\cdot\|$, we will typically use either the 0-norm (1.29) or the 1-norm (1.30), with V being $\mathcal{C}^0([a,b],\mathbb{R}^n)$ or $\mathcal{C}^1([a,b],\mathbb{R}^n)$, respectively. In the remainder of this section we discuss some general conditions for optimality which apply to both of these norms. However, when we develop more specific results later in calculus of variations, our findings for these two cases will be quite different.

1.3.2 First variation and first-order necessary condition

To develop the first-order necessary condition for optimality, we need a notion of derivative for functionals. Let $J : V \to \mathbb{R}$ be a functional on a function space V, and consider some function $y \in V$. The derivative of J at y, which will now be called the first variation, will also be a functional on V, and in fact this functional will be linear. To define it, we consider functions in V of the form $y + \alpha\eta$, where $\eta \in V$ and α is a real parameter (which can be restricted to some interval around 0). The reader will recognize these functions as infinite-dimensional analogs of the points $x^* + \alpha d$ around a given point $x^* \in \mathbb{R}^n$, which we utilized earlier.

A linear functional $\delta J|_y : V \to \mathbb{R}$ is called the *first variation* of J at y if for all η and all α we have

$$J(y + \alpha\eta) = J(y) + \delta J|_y(\eta)\alpha + o(\alpha) \tag{1.32}$$

where $o(\alpha)$ satisfies (1.6). The somewhat cumbersome notation $\delta J|_y(\eta)$ is meant to emphasize that the linear term in α in the expansion (1.32)

depends on both y and η. The requirement that $\delta J|_y$ must be a linear functional is understood in the usual sense: $\delta J|_y(\alpha_1\eta_1 + \alpha_2\eta_2) = \alpha_1\,\delta J|_y(\eta_1) + \alpha_2\,\delta J|_y(\eta_2)$ for all $\eta_1, \eta_2 \in V$ and $\alpha_1, \alpha_2 \in \mathbb{R}$.

The first variation as defined above corresponds to the so-called Gateaux derivative of J, which is just the usual derivative of $J(y + \alpha\eta)$ with respect to α (for fixed y and η) evaluated at $\alpha = 0$:

$$\delta J|_y(\eta) = \lim_{\alpha \to 0} \frac{J(y + \alpha\eta) - J(y)}{\alpha}. \tag{1.33}$$

In other words, if we define

$$g(\alpha) := J(y + \alpha\eta) \tag{1.34}$$

then

$$\delta J|_y(\eta) = g'(0) \tag{1.35}$$

and (1.32) reduces exactly to our earlier first-order expansion (1.5).

Now, suppose that y^* is a local minimum of J over some subset A of V. We call a perturbation[3] $\eta \in V$ *admissible* (with respect to the subset A) if $y^* + \alpha\eta \in A$ for all α sufficiently close to 0. It follows from our definitions of a local minimum and an admissible perturbation that $J(y^* + \alpha\eta)$ as a function of α has a local minimum at $\alpha = 0$ for each admissible η. Let us assume that the first variation $\delta J|_{y^*}$ exists (which is of course not always the case) so that we have (1.32). Applying the same reasoning that we used to derive the necessary condition (1.7) on the basis of (1.5), we quickly arrive at the **first-order necessary condition for optimality**: *For all admissible perturbations η, we must have*

$$\boxed{\delta J|_{y^*}(\eta) = 0} \tag{1.36}$$

As in the finite-dimensional case, the first-order necessary condition applies to both minima and maxima.

When we were studying a minimum x^* of $f : \mathbb{R}^n \to \mathbb{R}$ with the help of the function $g(\alpha) := f(x^* + \alpha d)$, it was easy to translate the equality $g'(0) = 0$ via the formula (1.10) into the necessary condition $\nabla f(x^*) = 0$. The necessary condition (1.36), while conceptually very similar, is much less constructive. To be able to apply it, we need to learn how to compute the first variation of some useful functionals. This subject will be further discussed in the next chapter; for now, we offer an example for the reader to work out.

[3]With a slight abuse of terminology, we call η a perturbation even though the actual perturbation is $\alpha\eta$.

Exercise 1.5 *Consider the space* $V = \mathcal{C}^0([0,1],\mathbb{R})$*, let* $\varphi : \mathbb{R} \to \mathbb{R}$ *be a* $\mathcal{C}^1$ *function, and define the functional* J *on* V *by* $J(y) = \int_0^1 \varphi(y(x))dx$*. Show that its first variation exists and is given by the formula* $\delta J|_y(\eta) = \int_0^1 \varphi'(y(x))\eta(x)dx$*.* □

Our notion of the first variation, defined via the expansion (1.32), is independent of the choice of the norm on V. This means that the first-order necessary condition (1.36) is valid for every norm. To obtain a necessary condition better tailored to a particular norm, we could define $\delta J|_y$ differently, by using the following expansion instead of (1.32):

$$J(y+\eta) = J(y) + \delta J|_y(\eta) + o(\|\eta\|). \tag{1.37}$$

The difference with our original formulation is subtle but substantial. The earlier expansion (1.32) describes how the value of J changes with α for each fixed η. In (1.37), the higher-order term is a function of $\|\eta\|$ and so the expansion captures the effect of all η at once (while α is no longer needed). We remark that the first variation defined via (1.37) corresponds to the so-called Fréchet derivative of J, which is a stronger differentiability notion than the Gateaux derivative (1.33). In fact, (1.37) suggests constructing more general perturbations: instead of working with functions of the form $y + \alpha\eta$, where η is fixed and α is a scalar parameter, we can consider perturbed functions $y + \eta$ which can approach y in a more arbitrary manner as $\|\eta\|$ tends to 0 (multiplying η by a vanishing parameter is just one possibility). This generalization is conceptually similar to that of passing from the lines $x^* + \alpha d$ used in Section 1.2.1 to the curves $x(\alpha)$ utilized in Section 1.2.2. We will start seeing perturbations of this kind in Chapter 3.

In what follows, we retain our original definition of the first variation in terms of (1.32). It is somewhat simpler to work with and is adequate for our needs (at least through Chapter 2). While the norm-dependent formulation could potentially provide sharper conditions for optimality, it takes more work to verify (1.37) for all η compared to verifying (1.32) for a fixed η. Besides, we will eventually abandon the analysis based on the first variation altogether in favor of more powerful tools. However, it is useful to be aware of the alternative formulation (1.37), and we will occasionally make some side remarks related to it. This issue will resurface in Chapter 3 where, although the alternative definition (1.37) of the first variation will not be specifically needed, we will use more general perturbations along the lines of the preceding discussion.

1.3.3 Second variation and second-order conditions

A real-valued functional B on $V \times V$ is called *bilinear* if it is linear in each argument (when the other one is fixed). Setting $Q(y) := B(y,y)$ we then

obtain a *quadratic functional*, or *quadratic form*, on V. This is a direct generalization of the corresponding familiar concepts for finite-dimensional vector spaces.

A quadratic form $\delta^2 J\big|_y : V \to \mathbb{R}$ is called the *second variation* of J at y if for all $\eta \in V$ and all α we have

$$J(y + \alpha\eta) = J(y) + \delta J\big|_y(\eta)\alpha + \delta^2 J\big|_y(\eta)\alpha^2 + o(\alpha^2). \tag{1.38}$$

This exactly corresponds to our previous second-order expansion (1.12) for the function g given by (1.34). Repeating the same argument we used earlier to prove (1.14), we easily establish the following **second-order necessary condition for optimality**: *If y^* is a local minimum of J over $A \subset V$, then for all admissible perturbations η we have*

$$\boxed{\delta^2 J\big|_{y^*}(\eta) \geq 0} \tag{1.39}$$

In other words, the second variation of J at y^* must be positive semidefinite on the space of admissible perturbations. For local maxima, the inequality in (1.39) is reversed. Of course, the usefulness of the condition will depend on our ability to compute the second variation of the functionals that we will want to study.

Exercise 1.6 *Consider the same functional J as in Exercise 1.5, but assume now that φ is $\mathcal{C}^2$. Derive a formula for the second variation of J (make sure that it is indeed a quadratic form).* □

What about a second-order *sufficient* condition for optimality? By analogy with the second-order sufficient condition (1.16) which we derived for the finite-dimensional case, we may guess that we need to combine the first-order necessary condition (1.36) with the strict-inequality counterpart of the second-order necessary condition (1.39), i.e.,

$$\delta^2 J\big|_{y^*}(\eta) > 0 \tag{1.40}$$

(this should again hold for all admissible perturbations η with respect to a subset A of V over which we want y^* to be a minimum). We would then hope to show that for $y = y^*$ the second-order term in (1.38) dominates the higher-order term $o(\alpha^2)$, which would imply that y^* is a strict local minimum (since the first-order term is 0). Our earlier proof of sufficiency of (1.16) followed the same idea. However, examining that proof more closely, the reader will discover that in the present case the argument does not go through.

We know that there exists an $\varepsilon > 0$ such that for all nonzero α with $|\alpha| < \varepsilon$ we have $|o(\alpha^2)| < \delta^2 J\big|_{y^*}(\eta)\alpha^2$. Using this inequality and (1.36), we obtain from (1.38) that $J(y^* + \alpha\eta) > J(y^*)$. Note that this does not yet

prove that y^* is a (strict) local minimum of J. According to the definition of a local minimum, we must show that $J(y^*)$ is the lowest value of J in some ball around y^* with respect to the selected norm $\|\cdot\|$ on V. The problem is that the term $o(\alpha^2)$ and hence the above ε depend on the choice of the perturbation η. In the finite-dimensional case we took the minimum of ε over all perturbations of unit length, but we cannot do that here because the unit sphere in the infinite-dimensional space V is not compact and the Weierstrass Theorem does not apply to it (see Section 1.3.4 below).

One way to resolve the above difficulty would be as follows. The first step is to strengthen the condition (1.40) to

$$\delta^2 J\big|_{y^*}(\eta) \geq \lambda \|\eta\|^2 \tag{1.41}$$

for some number $\lambda > 0$. The property (1.41) does not automatically follow from (1.40), again because we are in an infinite-dimensional space. (Quadratic forms satisfying (1.41) are sometimes called uniformly positive definite.) The second step is to modify the definitions of the first and second variations by explicitly requiring that the higher-order terms decay uniformly with respect to $\|\eta\|$. We already mentioned such an alternative definition of the first variation via the expansion (1.37). Similarly, we could define $\delta^2 J\big|_y$ via the following expansion in place of (1.38):

$$J(y+\eta) = J(y) + \delta J\big|_y(\eta) + \delta^2 J\big|_y(\eta) + o(\|\eta\|^2). \tag{1.42}$$

Adopting these alternative definitions and assuming that (1.36) and (1.41) hold, we could easily prove optimality by noting that $|o(\|\eta\|^2)| < \lambda\|\eta\|^2$ when $\|\eta\|$ is small enough.

With our current definitions of the first and second variations in terms of (1.32) and (1.38), we do not have a general second-order sufficient condition for optimality. However, in variational problems that we are going to study, the functional J to be minimized will take a specific form. This additional structure will allow us to derive conditions under which second-order terms dominate higher-order terms, resulting in optimality. The above discussion was given mainly for illustrative purposes, and will not be directly used in the sequel.

1.3.4 Global minima and convex problems

Regarding global minima of J over a set $A \subset V$, much of the discussion on global minima given at the end of Section 1.2.1 carries over to the present case. In particular, the Weierstrass Theorem is still valid, provided that compactness of A is understood in the sense of the second or third definition given on page 10 (existence of finite subcovers or sequential compactness). These two definitions of compactness are equivalent for linear vector

spaces equipped with a norm (or, more generally, a metric). On the other hand, closed and bounded subsets of an infinite-dimensional vector space are not necessarily compact—we already mentioned noncompactness of the unit sphere—and the Weierstrass Theorem does not apply to them; see the next exercise. We note that since our function space V has a norm, the notions of continuity of J and convergence, closedness, boundedness, and openness in V with respect to this norm are defined exactly as their familiar counterparts in $\mathbb{R}^n$. We leave it to the reader to write down precise definitions or consult the references given at the end of this chapter.

Exercise 1.7 *Give an example of a function space V, a norm on V, a closed and bounded subset A of V, and a continuous functional J on V such that a global minimum of J over A does not exist. (Be sure to demonstrate that all the requested properties hold.)* □

If J is a convex functional and $A \subset V$ is a convex set, then the optimization problem enjoys the same properties as the ones mentioned at the end of Section 1.2.1 for finite-dimensional convex problems. Namely, a local minimum is automatically a global one, and the first-order necessary condition is also a sufficient condition for a minimum. (Convexity of a functional and convexity of a subset of an infinite-dimensional linear vector space are defined exactly as the corresponding standard notions in the finite-dimensional case.) However, imposing extra assumptions to ensure convexity of J would severely restrict the classes of problems that we want to study. In this book, we focus on general theory that applies to not necessarily convex problems; we will not directly use results from convex optimization. Nevertheless, some basic concepts from (finite-dimensional) convex analysis will be important for us later, particularly when we derive the maximum principle.

1.4 NOTES AND REFERENCES FOR CHAPTER 1

Success stories of optimal control theory in various applications are too numerous to be listed here; see [CEHS87, Cla10, ST05, Swa84] for some examples from engineering and well beyond. The reader interested in applications will easily find many other references.

The material in Section 1.2 can be found in standard texts on optimization, such as [Lue84] or [Ber99]. See also Sections 5.2-5.4 of the book [AF66], which will be one of our main references for the optimal control chapters. Complete proofs of the results presented in Section 1.2.2, including the fact that the condition (1.20) is sufficient for d to be a tangent vector, are given in [Lue84, Chapter 10]. The alternative argument based on the inverse function theorem is adopted from [Mac05, Section 1.4]. The necessary condition in terms of Lagrange multipliers can also be derived from a cone separation

property (via Farkas's lemma) as shown, e.g., in [Ber99, Section 3.3.6]; we will see this type of reasoning in the proof of the maximum principle in Chapter 4.

Section 1.3 is largely based on [GF63], which will be our main reference for calculus of variations; function spaces, functionals, and the first variation are introduced in the first several sections of that book, while the second variation is discussed later in Chapter 5. Essentially the same material but in condensed form can be found in [AF66, Section 5.5]. In [GF63] as well as in [Mac05] the first and second variations are defined via (1.37) and (1.42), while other sources such as [You80] follow the approach based on (1.32) and (1.38).

For further background on function spaces and relevant topological concepts, the reader can consult [Rud76] or [Sut75] (the latter text is somewhat more advanced). Another recommended reference on these topics is [Lue69], where Gateaux and Fréchet derivatives and their role in functional minimization are also thoroughly discussed. A general treatment of convex functionals and their minimization is given in [Lue69, Chapter 7]; for convexity-based results more specific to calculus of variations and optimal control, see the monograph [Roc74], the more recent papers [RW00] and [GS07], and the references therein.