Department of Mathematics, London School of Economics

Optimisation in function spaces

Amol Sasane

Introduction

This pamphlet on calculus of variations and optimal control theory contains the most important results in the subject, treated largely in order of urgency.

The notes are elementary assuming no prerequisites beyond knowledge of linear algebra and ordinary calculus (with ϵ-δ arguments). The notes should hence be accessible to a wide spectrum of students.

In ordinary calculus, one dealt with limiting processes in finite-dimensional vector spaces ($\mathbb{R}$ or $\mathbb{R}^n$), but optimisation problems arising in applications require a calculus in spaces of functions (which are infinite-dimensional vector spaces). For instance, we mention the following problem.

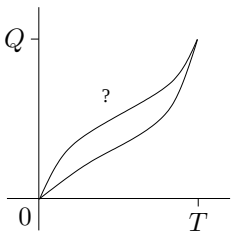
Problem. A copper mining company intends to remove all of the copper ore from a region that contains an estimated Q tons, over a time period of T years. As it is extracted, they will sell it for processing at a net price per ton (at time t) of

$$p(t) = P - ax(t) - bx'(t)$$

where positive constants P, a, and b are known, and where $x(t)$ denotes the total tonnage sold by time t (something that the company decides). If the company wishes to maximize its total profit given by

$$I(x) = \int_0^T [P - ax(t) - bx'(t)]x'(t)dt,$$

where $x(0) = 0$ and $x(T) = Q$, how might it proceed?



The optimal mining operation problem: what shape of the curve x gives the maximum profit?

We observe that this is an optimization problem: to each curve between the points $(0, 0)$ and (T, Q), there is a number (the associated profit), and the problem is to find the shape of the curve that minimizes this function

$$I : \{\text{curves between } (0, 0) \text{ and } (T, Q)\} \to \mathbb{R}.$$

This problem does not fit into the usual framework of calculus, where typically one has a function from some subset of the *finite* dimensional vector space $\mathbb{R}^n$ to $\mathbb{R}$, and one wishes to find a vector

in $\mathbb{R}^n$ that minimizes/maximizes the function, while in the above problem one has a subset of an *infinite* dimensional function space.

In ordinary calculus, given a function $f : \mathbb{R} \to \mathbb{R}$, we solve the optimization problem of finding that $x_0 \in \mathbb{R}$ that has the property that for all $x \in \mathbb{R}$, $f(x) \geq f(x_0)$ based on the following basic fact:

At the point x where $f(x)$ is minimum, the derivative $f(x)$ is zero.

This gives us an alogrithm to solve optimization problems: differentiate the given function, and find all x such that $f'(x) = 0$. These special x's are then candidates which maximize or minimize f. We would like to have a similar algorithm to solve optimization problems when the given function has its domain as a subset of some function space[1].

Thus the need arises for developing calculus in more general spaces than $\mathbb{R}^n$. Although we have only considered one example, optimisation problems requiring calculus in infinite-dimensional vector spaces arise from many applications and from various disciplines such as economics, engineering, physics, and so on. Mathematicians observed that different problems from varied fields often have related features and properties. This fact was used for an effective unifying approach towards such problems, the unification being obtained by the omission of unessential details. Hence the advantage of an *abstract* approach is that it concentrates on the essential facts, so that these facts become clearly visible and one's attention is not disturbed by unimportant details. Moreover, by developing a box of tools in the abstract framework, one is equipped to solve many different problems (that are really the same problem in disguise!). For example, while fishing for various different species of fish (bass, sardines, perch, and so on), one notices that in each of these different algorithms, the basic steps are the same: all one needs is a fishing rod and some bait. Of course, what bait one uses, where and when one fishes, depends on the particular species one wants to catch, but underlying these minor details, the basic technique is the same. So one can come up with an abstract algorithm for fishing, and applying this general algorithm to the particular species at hand, one gets an algorithm for catching that particular species. Such an abstract approach also has the advantage that it helps us to tackle unseen problems. For instance, if we are faced with a hitherto unknown species of fish, all that one has to do in order to catch it is to find out what it eats, and then by applying the general fishing algorithm, one would also be able to catch this new species.

In the abstract approach, one usually starts from a set of elements satisfying certain axioms. The theory then consists of logical consequences which result from the axioms and are derived as theorems once and for all. These general theorems can then later be applied to various concrete special sets satisfying the axioms.

We will develop such an abstract scheme for doing calculus in function spaces and other infinite-dimensional spaces. Having done this, we will be equipped with a box of tools for solving many problems, and in particular, we will return to the optimal mining operation problem again and solve it.

These notes contain many exercises, which form an integral part of the text, as some results relegated to the exercises are used in proving theorems. Some of the exercises are routine, and the harder ones are marked by an asterisk $(*)$.

Most applications of optimisation theory are drawn from the rudiments of the theory, but not all are, and no one can tell what topics will become important. In these notes we have described a few topics from optimisation and control theory which are basic and find widespread use, but by no means is the choice of topics 'complete'. However, equipped with this basic knowledge of the elementary facts, the student can undertake a serious study of a more advanced treatise on the

[1] By a function space, we mean an infinite-dimensional vector space comprising functions on an interval $[a, b]$.

subject, and the bibliography gives a few textbooks which might be suitable for further reading.

I am thankful to Dr. Sara Maad from the University of Surrey, U.K., for several useful discussions.

Amol Sasane
12 August, 2005.

Contents

Chapter 1

Calculus in normed spaces

1.1 Introduction

The derivative is used in solving maximization/minimization problems in the familiar calculus of functions from $\mathbb{R}$ to $\mathbb{R}$. Consider the quadratic function $f(x) = ax^2 + bx + c$. Suppose that one wants to know the points x_0 at which f assumes a maximum or a minimum. We know that if f has a maximum or a minimum at the point x_0, then the derivative of the function must be zero at that point: $f'(x_0) = 0$. See Figure 1.1.

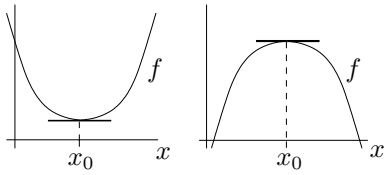


Figure 1.1: Necessary condition for x_0 to be an extremal point for f is that $f'(x_0) = 0$.

So one can then one can proceed as follows. First find the expression for the derivative: $f'(x) = 2ax + b$. Next solve for the unknown x_0 in the equation $f'(x_0) = 0$, that is,

$$2ax_0 + b = 0 \tag{1.1}$$

and so we find that a candidate for the point x_0 which minimizes or maximizes f is $x_0 = -\frac{b}{2a}$, which is obtained by solving the algebraic equation (1.1) above.

We wish to do the above with maps living on function spaces, such as $C[a, b]$, and taking values in $\mathbb{R}$. In order to do this we need a notion of derivative of a map from a function space to $\mathbb{R}$, and an analogue of the fact above concerning the necessity of the vanishing derivative at extremal points. In order to talk about the derivative, we need a notion of limits (so that the derivative can be defined), and in order to have a notion of a limit, we need a notion of 'distance' in the function space. It turns out that vector spaces such as $C[a, b]$ can be equipped with a 'norm', and this provides a 'distance' between two vectors. Having done this, we have the familiar setting of calculus, and we have notions of convergence, continuity and differentiability. This chapter has three sections:

1. In the first section, we will introduce the notion of a normed space. Roughly speaking, a

normed space is simply a vector space in which, using a function (called a norm), we can measure distances between vectors.

2. In the second section, we will discuss continuity of maps between two normed spaces. Continuity is important, since in the context of optimization problems, we would want the function being optimized to be a continuous function, and one which does not have sudden jumps. In this section we will also study those linear transformations that are also continuous, and give a characterization of such maps.

3. In this last section, we will study differentiability of maps between normed spaces. We will define the derivative of a map, and we will also prove a necessary condition for an extremum (derivative vanishes) and a sufficient condition for an extremum.

1.2 Normed spaces

1.2.1 Vector spaces

In this subsection, we recall the definition of a vector space. Roughly speaking it is a set of elements, called "vectors". Any two vectors can be "added", resulting in a new vector, and any vector can be multiplied by an element from $\mathbb{R}$, so as to give a new vector. The precise definition is given below.

Definition. A *vector space* over $\mathbb{R}$, is a set X together with two functions, $+ : X \times X \to X$, called *vector addition*, and $\cdot : \mathbb{R} \times X \to X$, called *scalar multiplication* that satisfy the following:

V1. For all $x_1, x_2, x_3 \in X$, $x_1 + (x_2 + x_3) = (x_1 + x_2) + x_3$.

V2. There exists an element, denoted by 0 (called the *zero vector*) such that for all $x \in X$, $x + 0 = 0 + x = x$.

V3. For every $x \in X$, there exists an element, denoted by $-x$, such that $x + (-x) = (-x) + x = 0$.

V4. For all x_1, x_2 in X, $x_1 + x_2 = x_2 + x_1$.

V5. For all $x \in X$, $1 \cdot x = x$.

V6. For all $x \in X$ and all $\alpha, \beta \in \mathbb{R}$, $\alpha \cdot (\beta \cdot x) = (\alpha\beta) \cdot x$.

V7. For all $x \in X$ and all $\alpha, \beta \in \mathbb{R}$, $(\alpha + \beta) \cdot x = \alpha \cdot x + \beta \cdot x$.

V8. For all $x_1, x_2 \in X$ and all $\alpha \in \mathbb{R}$, $\alpha \cdot (x_1 + x_2) = \alpha \cdot x_1 + \alpha \cdot x_2$.

Examples.

1. $\mathbb{R}$ is a vector space over $\mathbb{R}$, with vector addition being the usual addition of real numbers, and scalar multiplication being the usual multiplication of real numbers.

2. $\mathbb{R}^n$ is a vector space over $\mathbb{R}$, with addition and scalar multiplication defined as follows:

$$\text{if } \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n, \text{ then } \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{bmatrix};$$

$$\text{if } \alpha \in \mathbb{R} \text{ and } \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n, \text{ then } \alpha \cdot \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \alpha x_1 \\ \vdots \\ \alpha x_n \end{bmatrix}.$$

3. The sequence space ℓ^∞. This example and the next one give a first impression of how surprisingly general the concept of a vector space is.

   Let ℓ^∞ denote the vector space of all bounded sequences with values in $\mathbb{R}$, and with addition and scalar multiplication defined as follows:

$$(x_n)_{n\in\mathbb{N}} + (y_n)_{n\in\mathbb{N}} = (x_n + y_n)_{n\in\mathbb{N}}, \quad (x_n)_{n\in\mathbb{N}}, (y_n)_{n\in\mathbb{N}} \in \ell^\infty; \tag{1.2}$$

$$\alpha(x_n)_{n\in\mathbb{N}} = (\alpha x_n)_{n\in\mathbb{N}}, \quad \alpha \in \mathbb{R}, (x_n)_{n\in\mathbb{N}} \in \ell^\infty. \tag{1.3}$$

4. The function space $C[a, b]$. Let $a, b \in \mathbb{R}$ and $a < b$. Consider the vector space comprising functions $f : [a, b] \to \mathbb{R}$ that are continuous on $[a, b]$, with addition and scalar multiplication defined as follows. If $f, g \in C[a, b]$, then $f + g \in C[a, b]$ is the function given by

$$(f + g)(x) = f(x) + g(x), \quad x \in [a, b]. \tag{1.4}$$

   If $\alpha \in \mathbb{R}$ and $f \in C[a, b]$, then $\alpha f \in C[a, b]$ is the function given by

$$(\alpha f)(x) = \alpha f(x), \quad x \in [a, b]. \tag{1.5}$$

   $C[a, b]$ is referred to as a 'function space', since each vector in $C[a, b]$ is a function (from $[a, b]$ to $\mathbb{R}$).

5. Let $C^1[a, b]$ denote the space of continuously differentiable functions on $[a, b]$:

$$C^1[a, b] = \{f : [a, b] \to \mathbb{R} \mid f \text{ is continuously differentiable}\},$$

   (Recall that a function $f : [a, b] \to \mathbb{R}$ is *continuously differentiable* if for every $c \in [a, b]$, the derivative of f at c, namely $f'(c)$, exists, and the map $c \mapsto f'(c) : [a, b] \to \mathbb{R}$ is a continuous function.) Then $C^1[a, b]$ is a vector space with the operations defined by (1.4) and (1.5). $\Diamond$

Exercises.

1. Let $y_a, y_b \in \mathbb{R}$, and let

$$S(y_a, y_b) = \{x \in C^1[a, b] \mid x(a) = y_a \text{ and } x(b) = y_b\}.$$

   For what values of y_a, y_b is $S(y_a, y_b)$ a vector space?

2. Show that $C[0, 1]$ is not a finite dimensional vector space.

   HINT: One can prove this by contradiction. Let $C[0, 1]$ be a finite dimensional vector space with dimension d, say. First show that the set $B = \{x, x^2, \ldots, x^d\}$ is linearly independent. Then B is a basis for $C[0, 1]$, and so the constant function 1 should be a linear combination of the functions from B. Derive a contradiction.

1.2.2 Normed spaces

In order to do 'calculus' (that is, speak about limiting processes, convergence, approximation, continuity) in vector spaces, we need a notion of 'distance' or 'closeness' between the vectors of the vector space. This is provided by the notion of a norm.

Definitions. Let X be a vector space over $\mathbb{R}$ or $\mathbb{C}$. A *norm* on X is a function $\|\cdot\| : X \to [0, +\infty)$ such that:

N1. *(Positive definiteness)* For all $x \in X$, $\|x\| \geq 0$. If $x \in X$, then $\|x\| = 0$ iff $x = 0$.

N2. For all $\alpha \in \mathbb{R}$ (respectively $\mathbb{C}$) and for all $x \in X$, $\|\alpha x\| = |\alpha|\|x\|$.

N3. *(Triangle inequality)* For all $x, y \in X$, $\|x + y\| \leq \|x\| + \|y\|$.

A *normed space* is a vector space X equipped with a norm.

If $x, y \in X$, then the number $\|x - y\|$ provides a notion of closeness of points x and y in X, that is, a 'distance' between them. Thus $\|x\| = \|x - 0\|$ is the distance of x from the zero vector in X.

We now give a few examples of normed spaces.

Examples.

1. $\mathbb{R}$ is a vector space over $\mathbb{R}$, and if we define $\|\cdot\| : \mathbb{R} \to [0, +\infty)$ by

   $$\|x\| = |x|, \quad x \in \mathbb{R},$$

   then it becomes a normed space.

2. $\mathbb{R}^n$ is a vector space over $\mathbb{R}$, and let

   $$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}}, \quad x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n.$$

   Then $\mathbb{R}^n$ is a normed space (see Exercise 5a on page 6).

   This is not the only norm that can be defined on $\mathbb{R}^n$. For example,

   $$\|x\|_1 = \sum_{i=1}^{n} |x_i|, \quad \text{and} \quad \|x\|_\infty = \max\{|x_1|, \ldots, |x_n|\}, \quad x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n,$$

   are also examples of norms (see Exercise 5a on page 6).

   Note that $(\mathbb{R}^n, \|\cdot\|_2)$, $(\mathbb{R}^n, \|\cdot\|_1)$ and $(\mathbb{R}^n, \|\cdot\|_\infty)$ are all *different* normed spaces. This illustrates the important fact that from a given vector space, we can obtain various normed spaces by choosing different norms. What norm is considered depends on the particular application at hand. We illustrate this in the next paragraph.

   Suppose that we are interested in comparing the economic performance of a country from year to year, using certain economic indicators. For example, let the ordered 365-tuple

$x = (x_1, \ldots, x_{365})$ be the record of the daily industrial averages. A measure of differences in yearly performance is given by

$$\|x - y\| = \sum_{i=1}^{365} |x_i - y_i|.$$

Thus the space $(\mathbb{R}^{365}, \|\cdot\|_1)$ arises naturally. We might also be interested in the monthly cost of living index. Let the record of this index for a year be given by 12-tuples $x = (x_1, \ldots, x_{12})$. A measure of differences in yearly performance of the cost of living index is given by

$$\|x - y\| = \max\{|x_1 - y_1|, \ldots, |x_{12} - y_{12}|\},$$

which is the distance between x and y in the normed space $(\mathbb{R}^{12}, \|\cdot\|_\infty)$.

3. The sequence space ℓ^∞. This example and the next one give a first impression of how surprisingly general the concept of a normed space is.

   Let ℓ^∞ denote the vector space of all bounded sequences, with the addition and scalar multiplication defined earlier in (1.2)-(1.3).

   Define

   $$\|(x_n)_{n\in\mathbb{N}}\|_\infty = \sup_{n\in\mathbb{N}} |x_n|, \quad (x_n)_{n\in\mathbb{N}} \in \ell^\infty.$$

   Then it is easy to check that $\|\cdot\|_\infty$ is a norm, and so $(\ell^\infty, \|\cdot\|_\infty)$ is a normed space.

4. The function space $C[a, b]$. Let $a, b \in \mathbb{R}$ and $a < b$. Consider the vector space comprising functions that are continuous on $[a, b]$, with addition and scalar multiplication defined earlier by (1.4)-(1.5).

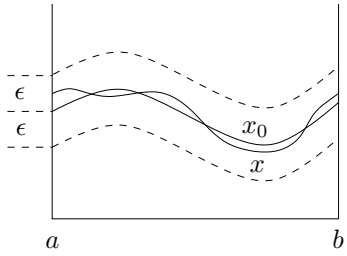


Figure 1.2: The set of all continuous functions x whose graph lies between the two dotted lines is the 'ball' $B(f, \epsilon) = \{x \in C[a, b] \mid \|x - x_0\|_\infty < \epsilon\}$.

   Define

   $$\|x\|_\infty = \sup_{t\in[a,b]} |x(t)|, \quad x \in C[a, b]. \tag{1.6}$$

   Then $\|\cdot\|_\infty$ is a norm on $C[a, b]$. Another norm is given by

   $$\|x\|_1 = \int_a^b |x(t)|dt, \quad x \in C[a, b]. \tag{1.7}$$

5. The function space $C^1[a, b]$. The space $C^1[a, b]$ consists of all functions x defined on $[a, b]$ which are continuous and have a continuous first derivative. The operations of addition and multiplication by scalars are the same as in $C[a, b]$, but we shall use the following norm:

   $$\|x\|_{1,\infty} = \sup_{t\in[a,b]} |x(t)| + \sup_{t\in[a,b]} \left|\frac{dx}{dt}(t)\right|. \tag{1.8}$$

Thus two functions in $C^1[a,b]$ are regarded as close together if both the functions themselves as well as their first derivatives are close together. Indeed, $\|x_1 - x_2\| < \epsilon$ implies that

$$|x_1(t) - x_2(t)| < \epsilon \quad \text{and} \quad \left| \frac{dx_1}{dt}(t) - \frac{dx_2}{dt}(t) \right| < \epsilon \quad \text{for all } t \in [a,b], \tag{1.9}$$

and conversely, (1.9) implies that $\|x_1 - x_2\| < 2\epsilon$. ◊

Exercises.

1. Let $(X, \|\cdot\|)$ be a normed space. Prove that for all $x, y \in X$, $|\|x\| - \|y\|| \leq \|x - y\|$.

2. If $x \in \mathbb{R}$, then let $\|x\| = |x|^2$. Is $\|\cdot\|$ a norm on $\mathbb{R}$?

3. Let $(X, \|\cdot\|)$ be a normed space and $r > 0$. Show that the function $x \mapsto r\|x\|$ defines a norm on X.

   Thus there are infinitely many other norms on any normed space.

4. Let X be a normed space $\|\cdot\|_X$ and Y be a subspace of X. Prove that Y is also a normed space with the norm $\|\cdot\|_Y$ defined simply as the restriction of the norm $\|\cdot\|_X$ to Y. This norm on Y is called the *induced norm*.

5. The *Cauchy-Schwarz inequality* says that if $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ are any real numbers, then

$$\left(\sum_{i=1}^{n} x_i y_i \right)^2 \leq \left(\sum_{i=1}^{n} x_i^2 \right) \left(\sum_{i=1}^{n} y_i^2 \right).$$

   If $n \in \mathbb{N}$, then for

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n,$$

   define

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} \text{ if } p = 1 \text{ or } 2, \quad \text{and} \quad \|x\|_\infty = \max\{|x_1|, \ldots, |x_n|\}. \tag{1.10}$$

   (a) Show that the function $x \mapsto \|x\|_p$ is a norm on $\mathbb{R}^n$.
   HINT: Use Cauchy-Schwarz inequality for the $p = 2$ case.

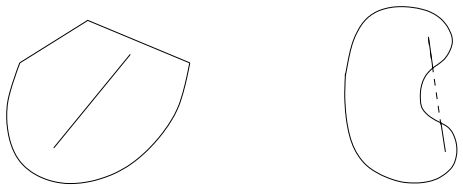
   (b) Let $n = 2$. Depict the following sets pictorially:

$$B_2(0,1) = \{x \in \mathbb{R}^2 \mid \|x\|_2 < 1\},$$
$$B_1(0,1) = \{x \in \mathbb{R}^2 \mid \|x\|_1 < 1\},$$
$$B_\infty(0,1) = \{x \in \mathbb{R}^2 \mid \|x\|_\infty < 1\}.$$

6. A subset C of a vector space X is said to be *convex* if for all $x, y \in C$, and all $\alpha \in [0,1]$, $\alpha x + (1-\alpha)y \in C$; see Figure 1.3.

   (a) Show that the unit ball $B(0,1) = \{x \in X \mid \|x\| < 1\}$ is convex in any normed space $(X, \|\cdot\|)$.

   (b) Sketch the curve $\{(x_1, x_2) \in \mathbb{R}^2 \mid \sqrt{|x_1|} + \sqrt{|x_2|} = 1\}$.



convex not convex

Figure 1.3: Examples of convex and nonconvex sets in $\mathbb{R}^2$.

(c) Prove that

$$\|x\|_{\frac{1}{2}} := \left(\sqrt{|x_1|} + \sqrt{|x_2|}\right)^2, \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2,$$

does not define a norm on $\mathbb{R}^2$.

7. (a) Show that the *polyhedron*

$$P_n = \left\{ \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n \;\middle|\; \forall i \in \{1, \ldots, n\},\ x_i > 0 \text{ and } \sum_{i=1}^{n} x_i = 1 \right\}$$

is convex in $\mathbb{R}^n$. Sketch P_2.

(b) Prove that

$$\text{if } \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in P_n, \text{ then } \sum_{i=1}^{n} \frac{1}{x_i} \geq n^2. \tag{1.11}$$

HINT: Use the Cauchy-Schwarz inequality.

(c) In the financial world, there is a method of investment called *dollar cost averaging*. Roughly speaking, this means that one invests a fixed amount of money regularly instead of a lumpsum. It is claimed that a person using dollar cost averaging should be better off than one who invests all the amount at one time. Suppose a fixed amount A is used to buy shares at prices $p_1, \ldots, p_n$. Then the total number of shares is then $\frac{A}{p_1} + \cdots + \frac{A}{p_n}$. If one invests the amount nA at a time when the share price is the average of $p_1, \ldots, p_n$, then the number of shares which one can purchase is $\frac{n^2 A}{p_1 + \cdots + p_n}$. Using the inequality (1.11), conclude that dollar cost averaging is at least as good as purchasing at the average share price.

8. ($*$) Show that (1.7) defines a norm on $C[a, b]$.

9. ($*$) Let $C^n[a, b]$ denote the space of n times continuously differentiable functions on $[a, b]$:

$$C^n[a, b] = \{f : [a, b] \to \mathbb{R} \mid f \text{ is } n \text{ times continuously differentiable}\},$$

equipped with the norm

$$\|f\|_{n,\infty} = \|f\|_\infty + \|f'\|_\infty + \cdots + \|f^{(n)}\|_\infty, \quad f \in C^n[a, b]. \tag{1.12}$$

Show that (1.12) defines a norm on $C[a, b]$.

1.3 Continuity

In this section, we consider continuous maps from a normed space X to a normed space Y. The spaces X and Y have a notion of distance between vectors (namely the norm of the difference between the two vectors). Hence we can talk about continuity of maps between these normed spaces, just as in the case of ordinary calculus.

Since the normed spaces are also vector spaces, linear maps play an important role. Recall that linear maps are those maps that preserve the vector space operations of addition and scalar multiplication. These are already familiar to the reader from elementary linear algebra, and they are called *linear transformations*.

In the context of normed spaces, it is then natural to focus attention on those linear transformations that are also continuous. These are called *bounded linear operators*. The reason for this terminology will become clear in Theorem 1.3.1.

The set of all bounded linear operators is itself a vector space, with obvious operations of addition and scalar multiplication, and as we shall see, it also has a natural notion of a norm, called the *operator norm*.

1.3.1 Linear transformations

We recall the definition of linear transformations below. Roughly speaking, linear transformations are maps that respect vector space operations.

Definition. Let X and Y be vector spaces over $\mathbb{R}$. A map $T : X \to Y$ is called a *linear transformation* if it satisfies the following:

L1. For all $x_1, x_2 \in X$, $T(x_1 + x_2) = T(x_1) + T(x_2)$.

L2. For all $x \in X$ and all $\alpha \in \mathbb{R}$, $T(\alpha \cdot x) = \alpha \cdot T(x)$.

Examples.

1. Let $m, n \in \mathbb{N}$ and $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$. If

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \in \mathbb{R}^{m \times n},$$

then the function $T_A : \mathbb{R}^n \to \mathbb{R}^m$ defined by

$$T_A \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + \cdots + a_{1n}x_n \\ \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{n} a_{1k}x_k \\ \vdots \\ \sum_{k=1}^{n} a_{mk}x_k \end{bmatrix} \text{ for all } \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n, \quad (1.13)$$

is a linear transformation from the vector space $\mathbb{R}^n$ to the vector space $\mathbb{R}^m$. Indeed,

$$T_A\left(\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}\right) = T_A \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + T_A \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \text{ for all } \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n,$$

and so L1 holds. Moreover,

$$T_A\left(\alpha \cdot \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}\right) = \alpha \cdot T_A \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \text{ for all } \alpha \in \mathbb{R} \text{ and all } \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n,$$

and so L2 holds as well. Hence T_A is a linear transformation.

2. Let $X = Y = \ell^\infty$. Consider the maps R, L from ℓ^2 to ℓ^2, defined as follows: if $(x_n)_{n\in\mathbb{N}} \in \ell^\infty$, then

$$R((x_1, x_2, x_3, \dots)) = (x_2, x_3, a_4, \dots) \quad \text{and} \quad L((x_1, x_2, x_3, \dots)) = (0, x_1, x_2, x_3, \dots).$$

It is easy to see that R and L are linear transformations.

3. The map $T : C[a, b] \to \mathbb{R}$ given by

$$Tf = f\left(\frac{a+b}{2}\right) \quad \text{for all } f \in C[a, b],$$

is a linear transformation from the vector space $C[a, b]$ to the vector space $\mathbb{R}$. Indeed, we have

$$T(f+g) = (f+g)\left(\frac{a+b}{2}\right) = f\left(\frac{a+b}{2}\right) + g\left(\frac{a+b}{2}\right) = T(f) + T(g), \text{ for all } f, g \in C[a, b],$$

and so L1 holds. Furthermore

$$T(\alpha \cdot f) = (\alpha \cdot f)\left(\frac{a+b}{2}\right) = \alpha f\left(\frac{a+b}{2}\right) = \alpha T(f), \text{ for all } \alpha \in \mathbb{R} \text{ and all } f \in C[a, b],$$

and so L2 holds too. Thus T is a linear transformation.

Similarly, the map $I : C[a, b] \to \mathbb{R}$ given by

$$I(f) = \int_a^b f(x)dx \quad \text{for all } f \in C[a, b],$$

is a linear transformation.

Another example of a linear transformation is the operation of differentiation: let $X = C^1[a, b]$ and $Y = C[a, b]$. Define $D : C^1[a, b] \to C[a, b]$ as follows: if $f \in C^1[a, b]$, then

$$(D(f))(x) = \frac{df}{dx}(x), \quad x \in [a, b].$$

It is easy to check that D is a linear transformation from the space of continuously differentiable functions to the space of continuous functions. ◊

Exercises. Let $a, b \in \mathbb{R}$, not both zeros, and consider the two real-valued functions f_1, f_2 defined on $\mathbb{R}$ by

$$f_1(x) = e^{ax}\cos(bx) \quad \text{and} \quad f_2(x) = e^{ax}\sin(bx), \quad x \in \mathbb{R}.$$

f_1 and f_2 are vectors belonging to the infinite-dimensional vector space over $\mathbb{R}$ (denoted by $C^1(\mathbb{R}, \mathbb{R})$), comprising all continuously differentiable functions from $\mathbb{R}$ to $\mathbb{R}$. Denote by $\mathscr{V}$ the span of the two functions f_1 and f_2.

1. Prove that f_1 and f_2 are linearly independent in $C^1(\mathbb{R}, \mathbb{R})$.
2. Show that the differentiation map D, $f \mapsto \frac{df}{dx}$, is a linear transformation from $\mathscr{V}$ to $\mathscr{V}$.
3. What is the matrix of D with respect to the basis $\mathscr{B} = \{f_1, f_2\}$?
4. Prove that D is invertible, and compute the matrix corresponding to this inverse.
5. Using the result above, compute the indefinite integrals

$$\int e^{ax} \cos(bx) dx \quad \text{and} \quad \int e^{ax} \sin(bx) dx.$$

Let X and Y be normed spaces. As there is a notion of distance between pairs of vectors in either space (provided by the norm of the difference of the pair of vectors in each respective space), one can talk about continuity of maps. Within the huge collection of all maps, the class of continuous maps form important subset. Continuous maps play a prominent role in functional analysis since they possess some useful properties.

Before discussing the case of a function between normed spaces, let us first of all recall the notion of continuity of a function $f : \mathbb{R} \to \mathbb{R}$.

1.3.2 Continuity of functions from $\mathbb{R}$ to $\mathbb{R}$

In everyday speech, a 'continuous' process is one that proceeds without gaps of interruptions or sudden changes. What does it mean for a function $f : \mathbb{R} \to \mathbb{R}$ to be continuous? The common informal definition of this concept states that a function f is continuous if one can sketch its graph without lifting the pencil. In other words, the graph of f has no breaks in it. If a break does occur in the graph, then this break will occur at some point. Thus (based on this visual view of continuity), we first give the formal definition of the continuity of a function *at a point* below. Next, if a function is continuous at *each* point, then it will be called continuous.

If a function has a break at a point, say x_0, then even if points x are close to x_0, the points $f(x)$ do not get close to $f(x_0)$. See Figure 1.4.

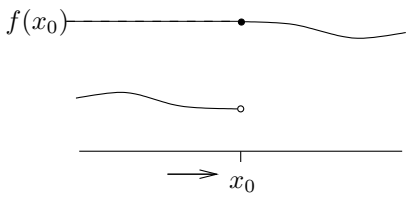


Figure 1.4: A function with a break at x_0. If x lies to the left of x_0, then $f(x)$ is not close to $f(x_0)$, no matter how close x comes to x_0.

This motivates the definition of continuity in calculus, which guarantees that if a function is continuous at a point x_0, then we can make $f(x)$ as close as we like to $f(x_0)$, by choosing x sufficiently close to x_0. See Figure 1.5.

Definitions. A function $f : \mathbb{R} \to \mathbb{R}$ is *continuous at* x_0 if for every $\epsilon > 0$, there exists a $\delta > 0$ such that for all $x \in \mathbb{R}$ satisfying $|x - x_0| < \delta$, $|f(x) - f(x_0)| < \epsilon$.

A function $f : \mathbb{R} \to \mathbb{R}$ is *continuous* if for every $x_0 \in \mathbb{R}$, f is continuous at x_0.

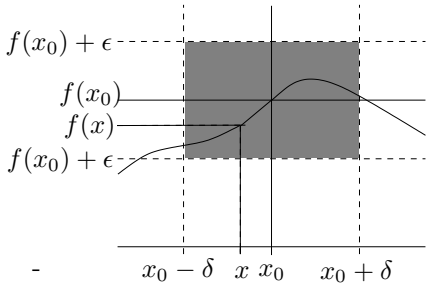


Figure 1.5: The definition of the continuity of a function at point x_0. If the function is continuous at x_0, then given any $\epsilon > 0$ (which determines a strip around the line $y = f(x_0)$ of width 2ϵ), there exists a $\delta > 0$ (which determines an interval of width 2δ around the point x_0) such that whenever x lies in this width (so that x satisfies $|x - x_0| < \delta$) and then $f(x)$ satisfies $|f(x) - f(x_0)| < \epsilon$.

For instance, if $\alpha \in \mathbb{R}$, then the linear map $x \mapsto x$ is continuous. It can be seen that sums and products of continuous functions are also continuous, and so it follows that all polynomial functions belong to the class of continuous functions from $\mathbb{R}$ to $\mathbb{R}$.

1.3.3 Continuity of functions between normed spaces

We now define the set of continuous maps from a normed space X to a normed space Y.

We observe that in the definition of continuity in ordinary calculus, if x, y are real numbers, then $|x - y|$ is a measure of the distance between them, and that the absolute value $|\cdot|$ is a norm in the finite (1) dimensional normed space $\mathbb{R}$.

So it is natural to define continuity in arbitrary normed spaces by simply replacing the absolute values by the corresponding norms, since the norm provides the notion of distance between vectors.

Definitions. Let X and Y be normed spaces over $\mathbb{R}$, and $x_0 \in X$. A map $f : X \to Y$ is said to be *continuous at* x_0 if

$$\forall \epsilon > 0, \quad \exists \delta > 0 \text{ such that } \forall x \in X \text{ satisfying } \|x - x_0\| < \delta, \quad \|f(x) - f(x_0)\| < \epsilon. \tag{1.14}$$

The map $f : X \to Y$ is called *continuous* if for all $x_0 \in X$, f is continuous at x_0.

We will see in the next section that the examples of the linear transformations given in the previous section are all continuous maps, if the vector spaces are equipped with their usual norms. Here we give an example of a *nonlinear* map which is continuous.

Example. Consider the squaring map $S : C[a, b] \to C[a, b]$ defined as follows:

$$(S(u))(t) = (u(t))^2, \quad t \in [a, b], \quad u \in C[a, b]. \tag{1.15}$$

The map is not linear (why?), but it is continuous. Indeed, let $u_0 \in C[a, b]$. Let

$$M = \max\{|u(t)| \mid t \in [a, b]\}$$

(extreme value theorem). Given any $\epsilon > 0$, let

$$\delta = \min\left\{1, \frac{\epsilon}{2M + 1}\right\}.$$

Then for any $u \in C[a,b]$, such that $\|u - u_0\| < \delta$, we have for all $t \in [a,b]$

$$\begin{aligned} |(u(t))^2 - (u_0(t))^2| &= |u(t) - u_0(t)||u(t) + u_0(t)| \\ &< \delta(|u(t) - u_0(t) + 2u_0(t)|) \\ &\leq \delta(|u(t) - u_0(t)| + 2|u_0(t)|) \\ &\leq \delta(\|u - u_0\| + 2M) \\ &< \delta(\delta + 2M) \\ &\leq \delta(1 + 2M) \\ &\leq \epsilon. \end{aligned}$$

Hence for all $u \in C[a,b]$ satisfying $\|u - u_0\| < \delta$, we have

$$\|S(u) - S(u_0)\| = \sup_{t\in[a,b]} |(u(t))^2 - (u_0(t))^2| \leq \epsilon.$$

So S is continuous at u_0. As the choice of $u_0 \in C[a,b]$ was arbitrary, it follows that S is continuous on $C[a,b]$. ◊

Exercises.

1. Show that the map $S : C[a,b] \to C[a,b]$ given by (1.15) is not a linear transformation.
2. Let $(X, \|\cdot\|)$ be a normed space. Show that the norm $\|\cdot\| : X \to \mathbb{R}$ is a continuous map.
3. (∗) Let $(z_n)_{n\in\mathbb{N}}$ be a sequence in a normed space Z and let $z \in Z$. The sequence $(z_n)_{n\in\mathbb{N}}$ *converges to* z if for all $\epsilon > 0$, there exists an $N \in \mathbb{N}$ such that for all $n \in \mathbb{N}$ satisfying $n \geq N$, $\|z_n - z\| < \epsilon$.

   Let X, Y be normed spaces and suppose that $f : X \to Y$ is a map. Prove that f is continuous at $x_0 \in X$ iff

   $$\boxed{\begin{array}{l}\text{for every convergent sequence } (x_n)_{n\in\mathbb{N}} \text{ contained in } X \text{ with limit } x_0, \\ (f(x_n))_{n\in\mathbb{N}} \text{ is convergent and } \lim_{n\to\infty} f(x_n) = f(x_0).\end{array}} \tag{1.16}$$

4. (∗) This exercise concerns the norm on $C^1[a,b]$ we have chosen to use. Since we want to be able to use ordinary analytic operations such as passage to the limit, then, given a function $I : C^1[a,b] \to \mathbb{R}$, it is reasonable to choose a norm such that I is continuous.

   (a) It might seem that induced norm from the space $C[a,b]$ (of which $C^1[a,b]$ as a subspace) would be adequate for the study of variational problems. However, this is not true. In fact the function

   $$I(x) = \int_a^b F\left(x(t), \frac{dx}{dt}(t), t\right) dt$$

   may not be continuous if we use the norm induced by $C[a,b]$. For example, show that the arc length function $L : C^1[0,1] \to \mathbb{R}$ given by

   $$L(x) = \int_0^1 \sqrt{1 + (x'(t))^2} dt$$

   is not continuous if we equip $C^1[0,1]$ with the norm

   $$\|x\| = \sup_{t\in[0,1]} |x(t)|.$$

   HINT: For every curve, we can find another curve arbitrarily close to the first in the sense of the norm of $C[a,b]$, whose length differs from that of the first curve by a factor of 10, say.

(b) Show that the arc length function L is continuous if we equip $C^1[a,b]$ with the norm given by (1.12).

5. Consider the function $I : C^1[a,b] \to \mathbb{R}$ defined by

$$I(x) = \int_a^b \left(x(t) + t\frac{dx}{dt}(t) \right) dt, \quad x \in C^1[a,b].$$

Is I linear? Is it continuous? Let $S(y_a, y_b) = \{x \in C^1[a,b] \mid x(a) = y_a \text{ and } x(b) = y_b\}$. Prove that I is constant on $S(y_a, y_b)$. What is the value of I on $S(y_a, y_b)$?

1.3.4 The normed space $\mathscr{L}(X,Y)$

In this section we study those linear transformations from a normed space X to a normed space Y that are also continuous. We begin by giving a characterization of continuous linear transformations.

Theorem 1.3.1 *Let X and Y be normed spaces over $\mathbb{R}$. Let $T : X \to Y$ be a linear transformation. Then the following properties of T are equivalent:*

1. *T is continuous.*
2. *T is continuous at 0.*
3. *There exists a number M such that for all $x \in X$, $\|Tx\| \leq M\|x\|$.*

Proof

$1 \Rightarrow 2$. Evident.

$2 \Rightarrow 3$. For every $\epsilon > 0$, for example $\epsilon = 1$, there exists a $\delta > 0$ such that $\|x\| \leq \delta$ implies $\|Tx\| \leq 1$. This yields:

$$\|Tx\| \leq \frac{1}{\delta}\|x\| \quad \text{for all } x \in X. \tag{1.17}$$

This is true if $\|x\| = \delta$. But if (1.17) holds for some x, then owing to the homogeneity of T and of the norm, it also holds for αx, for any arbitrary $\alpha \in \mathbb{R}$. Since every x can be written in the form $x = \alpha y$ with $\|y\| = \delta$ (take $\alpha = \frac{\|x\|}{\delta}$), (1.17) is valid for all x. Thus we have that for all $x \in X$,

$$\|Tx\| \leq M\|x\|$$

with $M = \frac{1}{\delta}$.

$3 \Rightarrow 1$. From linearity, we have:

$$\|Tx - Ty\| = \|T(x-y)\| \leq M\|x-y\|$$

for all $x, y \in X$. The continuity follows immediately. ■

Owing to the characterization of continuous linear transformations by the existence of a bound as in item 3 above, they are called *bounded* linear operators.

Theorem 1.3.2 *Let X and Y be normed spaces over $\mathbb{R}$.*

1. *Let $T : X \to Y$ be a linear operator. Of all the constants M possible in 3 of Theorem 1.3.1, there is a smallest one, and this is given by:*

$$\|T\| = \sup_{\|x\|\leq 1} \|Tx\|. \tag{1.18}$$

2. *The set $\mathscr{L}(X,Y)$ of bounded linear operators from X to Y with addition and scalar multiplication defined by:*

$$(T+S)x = Tx + Sx, \quad x \in X, \tag{1.19}$$
$$(\alpha T)x = \alpha Tx, \quad x \in X, \quad \alpha \in \mathbb{R}, \tag{1.20}$$

is a vector space. The map $T \mapsto \|T\|$ is a norm on this space.

Proof 1. From item 3 of Theorem 1.3.1, it follows immediately that $\|T\| \leq M$. Conversely we have, by the definition of $\|T\|$, that $\|x\| \leq 1 \;\Rightarrow\; \|Tx\| \leq \|T\|$. Owing to the homogeneity of T and of the norm, it again follows from this that:

$$\|Tx\| \leq \|T\|\|x\| \quad \text{for all } x \in X \tag{1.21}$$

which means that $\|T\|$ is the smallest constant M that can occur in item 3 of Theorem 1.3.1.

2. We already know from linear algebra that the space of *all* linear transformations from a vector space X to a vector space Y, equipped with the operations of addition and scalar multiplication given by (1.19) and (1.20), forms a vector space. We now prove that the subset $\mathscr{L}(X,Y)$ comprising *bounded* linear transformations is a subspace of this vector space, and consequently it is itself a vector space.

We first prove that if T, S are in bounded linear transformations, then so are $T+S$ and αT. It is clear that $T+S$ and αT are linear transformations. Moreover, there holds that

$$\|(T+S)x\| \leq \|Tx\| + \|Sx\| \leq (\|T\| + \|S\|)\|x\|, \quad x \in X, \tag{1.22}$$

from which it follows that $T+S$ is bounded. Also there holds:

$$\|\alpha T\| = \sup_{\|x\|\leq 1} \|\alpha Tx\| = \sup_{\|x\|\leq 1} |\alpha|\|Tx\| = |\alpha| \sup_{\|x\|\leq 1} \|Tx\| = |\alpha|\|T\|. \tag{1.23}$$

Finally, the 0 operator, is bounded and so it belongs to $\mathscr{L}(X,Y)$.

Furthermore, $\mathscr{L}(X,Y)$ is a normed space. Indeed, from (1.22), it follows that

$$\|T+S\| \leq \|T\| + \|S\|,$$

and so N3 holds. Also, from (1.23) we see that N2 holds. We have $\|T\| \geq 0$; from (1.21) it follows that if $\|T\| = 0$, then $Tx = 0$ for all $x \in X$, that is, $T = 0$, the operator 0, which is the zero vector of the space $\mathscr{L}(X,Y)$. This shows that N1 holds. ∎

So we have shown that the space of all continuous linear transformations (which we also call the space of bounded linear operators), $\mathscr{L}(X,Y)$, can be equipped with the operator norm given by (1.18), so that $\mathscr{L}(X,Y)$ becomes a normed space.

Remark. The space $\mathscr{L}(X,\mathbb{R})$ is denoted by X' (sometimes X^*) and is called the *dual space*. Elements of the dual space are called *functionals*.

We give a few examples of bounded linear operators below:

Examples.

1. Let $X = \mathbb{R}^n$, $Y = \mathbb{R}^m$, and let

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \in \mathbb{R}^{m\times n}.$$

We equip X and Y with the Euclidean norm. From the Cauchy-Schwarz inequality, it follows that

$$\left(\sum_{j=1}^{n} a_{ij}x_j\right)^2 \leq \left(\sum_{j=1}^{n} a_{ij}^2\right)\|x\|^2,$$

for each $i \in \{1, \dots m\}$. This yields $\|T_A x\| \leq \|A\|_2\|x\|$ where

$$\|A\|_2 = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2\right)^{\frac{1}{2}}. \tag{1.24}$$

Thus we see that *all* linear transformations in finite dimensional spaces are continuous, and that if X and Y are equipped with the Euclidean norm, then the operator norm is majorized by the Euclidean norm of the matrix:

$$\|A\| \leq \|A\|_2.$$

2. We take $X = C[a, b]$, and $Y = \mathbb{R}$. Consider the operator $I : C[a, b] \to \mathbb{R}$ given by

$$I(f) = \int_a^b f(x)dx. \tag{1.25}$$

The map $f \mapsto I(f)$ is clearly linear. Moreover,

$$|I(f)| \leq \int_a^b |f(x)|dx \leq \int_a^b \|f\|_\infty dx = (b-a)\|f\|_\infty.$$

Thus it follows that I is bounded, and that $\|I\| \leq b - a$. ◊

Exercises.

1. Let X, Y be normed spaces, and $T \in \mathscr{L}(X, Y)$. Show that

$$\|T\| = \sup\{\|Tx\| \mid x \in X \text{ and } \|x\| = 1\}.$$

2. Let $(\lambda_n)_{n\in\mathbb{N}}$ be a bounded sequence of scalars, and consider the *diagonal operator* $D : \ell^\infty \to \ell^\infty$ defined as follows:

$$D(x_1, x_2, x_3, \dots) = (\lambda_1 x_1, \lambda_2 x_2, \lambda_3 x_3, \dots), \quad (x_n)_{n\in\mathbb{N}} \in \ell^\infty. \tag{1.26}$$

Prove that $D \in \mathscr{L}(\ell^\infty)$ and that

$$\|D\| = \sup_{n\in\mathbb{N}} |\lambda_n|.$$

3. An analogue of the diagonal operator in the context of function spaces is the multiplication operator. Let l be a continuous function on $[0, 1]$. Define the *multiplication operator* $M : C[0, 1] \to C[0, 1]$ as follows:

$$(Mf)(x) = l(x)f(x), \quad x \in [0, 1], \quad f \in C[0, 1].$$

Is M a bounded linear operator?

4. Prove that the *averaging operator* $A : \ell^\infty \to \ell^\infty$, defined by

$$A(x_1, x_2, x_3, \dots) = \left(x_1, \frac{x_1 + x_2}{2}, \frac{x_1 + x_2 + x_3}{3}, \dots \right), \tag{1.27}$$

is a bounded linear operator.

5. Consider the subspace c of ℓ^∞ comprising convergent sequences. Prove that the limit map $l : c \to \mathbb{R}$ given by

$$l(x_n)_{n\in\mathbb{N}} = \lim_{n\to\infty} x_n, \quad (x_n)_{n\in\mathbb{N}} \in c, \tag{1.28}$$

is an element in the dual space $\mathscr{L}(c, \mathbb{R})$ of c, when c is equipped with the induced norm from ℓ^∞.

1.4 Differentiation

In the last section we studied continuity of operators from a normed space X to a normed space Y. In this section, we will study differentiation: we will define the (Frechet) derivative of a map $F : X \to Y$ at a point $x_0 \in X$. Roughly speaking, the derivative of a nonlinear map at a point is a local approximation by means of a continuous linear transformation. Thus the derivative at a point will be a bounded linear operator.

Next, we will use the derivative in solving optimization problems in normed spaces. If I is a differentiable map from the normed space X to $\mathbb{R}$, we prove Theorem 1.4.2, which says that this derivative must vanish at local maximum/minimum of the map I.

Finally, we apply Theorem 1.4.2 to the problem mentioned in the introduction. This is the concrete case when X comprises continuously differentiable functions on the interval $[0, T]$, and I is the map

$$I(x) = \int_0^T (P - ax(t) - bx'(t))dt. \tag{1.29}$$

Setting the derivative of such a functional to zero, a necessary condition (in the form of a *differential* equation) for an extremal curve x_0 is obtained. The solution x_0 of this differential equation is the candidate which maximizes/minimizes the function I.

1.4.1 The derivative

Recall that for a function $f : \mathbb{R} \to \mathbb{R}$, the derivative at a point x_0 is the approximation of f around x_0 by a straight line. See Figure 1.6.

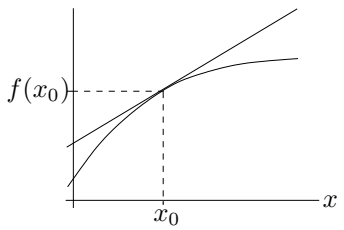


Figure 1.6: The derivative of f at x_0.

The derivative $f'(x_0)$ gives the slope of the line which is tangent to the function f at the point x_0:

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}.$$

In other words,

$$\lim_{x \to x_0} \left| \frac{f(x) - f(x_0) - f'(x_0)(x - x_0)}{x - x_0} \right| = 0,$$

that is,

$$\forall \epsilon > 0,\ \exists \delta > 0 \text{ such that } \forall x \in \mathbb{R} \backslash \{x_0\} \text{ satisfying } |x - x_0| < \delta,\ \frac{|f(x) - f(x_0) - f'(x_0)(x - x_0)|}{|x - x_0|} < \epsilon.$$

Observe that every real number α gives rise to a linear transformation from $\mathbb{R}$ to $\mathbb{R}$: the operator in question is simply multiplication by α, that is the map $x \mapsto \alpha x$. We can therefore think of $(f'(x_0))(x - x_0)$ as the action of the linear transformation $L : \mathbb{R} \to \mathbb{R}$ on the vector $x - x_0$, where L is given by

$$L(h) = f'(x_0)h, \quad h \in \mathbb{R}.$$

Hence the derivative $f'(x_0)$ is simply a linear map from $\mathbb{R}$ to $\mathbb{R}$. In the same manner, in the definition of a derivative of a map $F : X \to Y$ between normed spaces X and Y, the derivative of F at a point x_0 will be defined to be a linear transformation from X to Y.

A linear map $L : \mathbb{R} \to \mathbb{R}$ is automatically continuous[1]. But this is not true in general if $\mathbb{R}$ is replaced by infinite dimensional normed spaces! And we would expect that the derivative (being the approximation of the map at a point) to have the same property as the function itself at that point. Of course, a differentiable function should first of all be continuous (so that this situation matches with the case of functions from $\mathbb{R}$ to $\mathbb{R}$ from ordinary calculus), and so we expect the derivative to be a continuous linear transformation, that is, it should be a bounded linear operator. So while generalizing the notion of the derivative from ordinary calculus to the case of a map $F : X \to Y$ between normed spaces X and Y, we now specify *continuity* of the derivative as well. Thus, in the definition of a derivative of a map F, the derivative of F at a point x_0 will be defined to be a *bounded* linear transformation from X to Y, that is, an element of $\mathscr{L}(X, Y)$.

This motivates the following definition.

Definition. Let X, Y be normed spaces. If $F : X \to Y$ be a map and $x_0 \in X$, then F is said to be *differentiable at* x_0 if there exists a bounded linear operator $L \in \mathscr{L}(X, Y)$ such that

$$\forall \epsilon > 0,\ \exists \delta > 0 \text{ such that } \forall x \in X \backslash \{x_0\} \text{ satisfying } \|x - x_0\| < \delta,\ \frac{\|F(x) - F(x_0) - L(x - x_0)\|}{\|x - x_0\|} < \epsilon. \tag{1.30}$$

The operator L is called a *derivative of* F *at* x_0. If F is differentiable at every point $x \in X$, then it is simply said to be *differentiable*.

We now prove that if F is differentiable at x_0, then its derivative is unique.

Theorem 1.4.1 *Let X, Y be normed spaces. If $F : X \to Y$ is differentiable at $x_0 \in X$, then the derivative of F at x_0 is unique.*

[1]Indeed, every linear map $L : \mathbb{R} \to \mathbb{R}$ is simply given by multiplication, since $L(x) = L(x \cdot 1) = xL(1)$. Consequently $|L(x) - L(y)| = |L(1)||x - y|$, and so L is continuous!

Proof Suppose that $L_1, L_2 \in \mathscr{L}(X,Y)$ are derivatives of F at x_0. Given $\epsilon > 0$, choose a δ such that (1.30) holds with L_1 and L_2 instead of L. Consequently

$$\forall x \in X \setminus \{x_0\} \text{ satisfying } \|x - x_0\| < \delta, \ \frac{\|L_2(x - x_0) - L_1(x - x_0)\|}{\|x - x_0\|} < 2\epsilon. \tag{1.31}$$

Given any $h \in X$ such that $h \neq 0$, define $x = x_0 + \frac{\delta}{2\|h\|}h$. Then $\|x - x_0\| = \frac{\delta}{2} < \delta$ and so (1.31) yields

$$\|(L_2 - L_1)h\| \leq 2\epsilon\|h\|. \tag{1.32}$$

Hence $\|L_2 - L_1\| \leq 2\epsilon$, and since the choice of $\epsilon > 0$ was arbitrary, we obtain $\|L_2 - L_1\| = 0$. So $L_2 = L_1$, and this completes the proof. ■

Notation. We denote the derivative of F at x_0 by $DF(x_0)$.

We now calculate the derivative in the case of a few simple examples.

Examples.

1. Consider the nonlinear squaring map S from the example on page 11. We had seen that S is continuous. We now see that $S : C[a,b] \to C[a,b]$ is in fact differentiable. We note that

$$(Su - Su_0)(t) = u(t)^2 - u_0(t)^2 = \underbrace{(u(t) + u_0(t))}(u(t) - u_0(t)). \tag{1.33}$$

   As u approaches u_0 in $C[a,b]$, the term $u(t)+u_0(t)$ above approaches $2u_0(t)$. So from (1.33), we suspect that $(DS)(u_0)$ would be the multiplication map M by $2u_0$:

$$(Mu)(t) = 2u_0(t)u(t), \quad t \in [a,b].$$

   Let us prove this. Let $\epsilon > 0$. We have

$$\begin{aligned} |(Su - Su_0 - M(u - u_0))(t)| &= |u(t)^2 - u_0(t)^2 - 2u_0(t)(u(t) - u_0(t))| \\ &= |u(t)^2 + u_0(t)^2 - 2u_0(t)u(t)| \\ &= |u(t) - u_0(t)|^2 \\ &\leq \|u - u_0\|^2. \end{aligned}$$

   Hence if $\delta := \epsilon > 0$, then for all $u \in C[a,b] \setminus \{u_0\}$ satisfying $\|u - u_0\| < \delta$, we have

$$\frac{\|Su - Su_0 - M(u - u_0)\|}{\|u - u_0\|} \leq \|u - u_0\| < \delta = \epsilon.$$

   Thus $DS(u_0) = M$.

2. Let X, Y be normed spaces and let $T \in \mathscr{L}(X,Y)$. Is T differentiable, and if so, what is its derivative?

   Recall that the derivative at a point is the linear transformation that approximates the map at that point. If the map is itself linear, then we expect the derivative to equal the given linear map! We claim that $(DT)(x_0) = T$, and we prove this below.

   Given $\epsilon > 0$, choose any $\delta > 0$. Then for all $x \in X$ satisfying $\|x - x_0\| < \delta$, we have

$$\|Tx - Tx_0 - T(x - x_0)\| = \|Tx - Tx_0 - Tx + Tx_0\| = 0 < \epsilon.$$

   Consequently $(DT)(x_0) = T$.

   In particular, if $X = \mathbb{R}^n$, $Y = \mathbb{R}^m$, and $T = T_A$, where $A \in \mathbb{R}^{m\times n}$, then $(DT_A)(x_0) = T_A$. ◊

Exercises.

1. Let X, Y be normed spaces. Prove that if $F : X \to Y$ is differentiable at x_0, then it F continuous at x_0.

2. Consider the functional $I : C[a, b] \to \mathbb{R}$ given by

$$I(x) = \int_a^b x(t)dt.$$

   Prove that I is differentiable, and find its derivative at $x_0 \in C[a, b]$.

3. $(*)$ Prove that the square of a differentiable functional $I : X \to \mathbb{R}$ is differentiable, and find an expression for its derivative at $x \in X$.
   HINT: $(I(x))^2 - (I(x_0))^2 = (I(x) + I(x_0))(I(x) - I(x_0)) \approx 2I(x_0)DI(x_0)(x - x_0)$ if $x \approx x_0$.

4. (a) Given x_1, x_2 in a normed space X, define

$$\varphi(t) = tx_1 + (1 - t)x_2.$$

   Prove that if $I : X \to \mathbb{R}$ is differentiable, then $I \circ \varphi : [0, 1] \to \mathbb{R}$ is differentiable and

$$\frac{d}{dt}(I \circ \varphi)(t) = [DI(\varphi(t))](x_1 - x_2).$$

   (b) Prove that if $I_1, I_2 : X \to \mathbb{R}$ are differentiable and their derivatives are equal at every $x \in X$, then I_1 and I_2 differ by a constant.

1.4.2 Optimization: necessity of vanishing derivative

In this section we take the normed space $Y = \mathbb{R}$, and consider maps $I : X \to \mathbb{R}$. We wish to find points $x_0 \in X$ that maximize/minimize I.

In elementary analysis, a necessary condition for a differentiable function $f : \mathbb{R} \to \mathbb{R}$ to have a local extremum (local maximum or local minimum) at $x_0 \in \mathbb{R}$ is that $f'(x_0) = 0$. We will prove a similar necessary condition for a differentiable function $I : X \to \mathbb{R}$.

First we specify what exactly we mean by a 'local maximum/minimum' (collectively termed 'local extremum'). Roughly speaking, a point $x_0 \in X$ is a local maximum/minimum for I if for all points x in some neighbourhood of that point, the values $I(x)$ are all less (respectively greater) than $I(x_0)$. Since in general the functions I might be defined only on some subset S of a normed space X, we give the following general definition.

Definition. Let X be a normed space and $S \subset X$. A function $I : S \to \mathbb{R}$ is said to have a *local extremum* at x_0 $(\in S)$ if there exists a $\delta > 0$ such that

$$\forall x \in S \text{ satisfying } \|x - x_0\| < \delta, \quad I(x) \geq I(x_0) \quad \text{(local minimum)}$$

or

$$\forall x \in S \text{ satisfying } \|x - x_0\| < \delta, \quad I(x) \leq I(x_0) \quad \text{(local maximum)}.$$

Theorem 1.4.2 *Let X be a normed space, and let $I : X \to \mathbb{R}$ be a function that is differentiable at $x_0 \in X$. If I has a local extremum at x_0, then $(DI)(x_0) = 0$.*

Proof We prove the statement in the case that I has a local minimum at x_0. (If instead I has a local maximum at x_0, then the function $-I$ has a local minimum at x_0, and so $(DI)(x_0) = -(D(-I))(x_0) = 0$.)

For notational simplicity, we denote $(DI)(x_0)$ by L. Suppose that $Lh \neq 0$ for some $h \in X$. Let $\epsilon > 0$ be given. Choose a δ such that for all $x \in X$ satisfying $\|x - x_0\| < \delta$, $I(x) \geq I(x_0)$, and moreover if $x \neq x_0$, then

$$\frac{|I(x) - I(x_0) - L(x - x_0)|}{\|x - x_0\|} < \epsilon.$$

Define the sequence

$$x_n = x_0 - \frac{1}{n}\frac{Lh}{|Lh|}h, \quad n \in \mathbb{N}.$$

We note that $\|x_n - x_0\| = \frac{\|h\|}{n}$, and so with N chosen large enough, we have $\|x_n - x_0\| < \delta$ for all $n > N$. It follows that for all $n > N$,

$$0 \leq \frac{I(x_n) - I(x_0)}{\|x_n - x_0\|} < \frac{L(x_n - x_0)}{\|x_n - x_0\|} + \epsilon = -\frac{|Lh|}{\|h\|} + \epsilon.$$

Since the choice of $\epsilon > 0$ was arbitrary, we obtain $|Lh| \leq 0$, and so $Lh = 0$, a contradiction. ■

Remark. Note that this is a *necessary* condition for the existence of a local extremum. Thus the vanishing of a derivative at some point x_0 doesn't imply local extremality at x_0! This is analogous to the case of $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^3$, for which $f'(0) = 0$, although f clearly does not have a local minimum or maximum at 0. In the next section we study an important class of functions $I : X \to \mathbb{R}$, called *convex* functions, for which a vanishing derivative implies the function has a global minimum at that point!

1.4.3 Optimization: sufficiency in the convex case

In this section, we will show that if $I : X \to \mathbb{R}$ is a convex function, then a vanishing derivative is enough to conclude that the function has a global minimum at that point. We begin by giving the definition of a convex function.

Definition. Let X be a normed space. A function $I : X \to \mathbb{R}$ is *convex* if for all $x_1, x_2 \in X$ and all $\alpha \in [0, 1]$,

$$I(\alpha x_1 + (1 - \alpha)x_2) \leq \alpha I(x_1) + (1 - \alpha)I(x_2). \tag{1.34}$$

We now consider a few examples of convex functions.

Examples.

1. If $X = \mathbb{R}$, then the function $f(x) = x^2$, $x \in \mathbb{R}$, is convex. This is visually obvious from Figure 1.8, since we see that the point B lies above the point A:

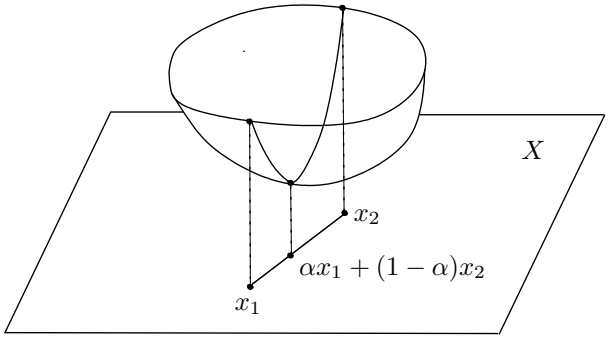


Figure 1.7: Convex function.

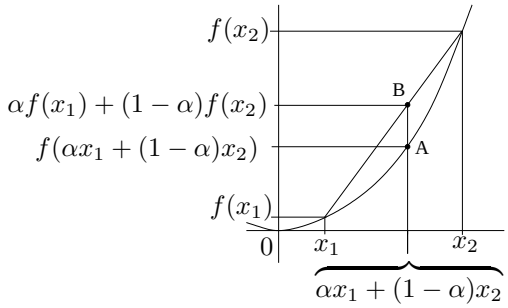


Figure 1.8: The convex function $x \mapsto x^2$.

But one can prove this as follows: for all $x_1, x_2 \in \mathbb{R}$ and all $\alpha \in [0,1]$, we have

$$\begin{aligned} f(\alpha x_1 + (1-\alpha)x_2) &= (\alpha x_1 + (1-\alpha)x_2)^2 = \alpha^2 x_1^2 + 2\alpha(1-\alpha)x_1x_2 + (1-\alpha)^2 x_2^2 \\ &= \alpha x_1^2 + (1-\alpha)x_2^2 + (\alpha^2 - \alpha)x_1^2 + (\alpha^2 - \alpha)x_2^2 + 2\alpha(1-\alpha)x_1x_2 \\ &= \alpha x_1^2 + (1-\alpha)x_2^2 - \alpha(1-\alpha)(x_1^2 + x_2^2 - 2x_1x_2) \\ &= \alpha x_1^2 + (1-\alpha)x_2^2 - \alpha(1-\alpha)(x_1 - x_2)^2 \\ &\leq \alpha x_1^2 + (1-\alpha)x_2^2 = \alpha f(x_1) + (1-\alpha)f(x_2). \end{aligned}$$

A slick way of proving convexity of smooth functions from $\mathbb{R}$ to $\mathbb{R}$ is to check if f'' is nonnegative; see Exercise 1 below.

2. Consider $I : C[0,1] \to \mathbb{R}$ given by

$$I(f) = \int_0^1 (f(x))^2 dx, \quad f \in C[0,1].$$

Then I is convex, since for all $f_1, f_2 \in C[0,1]$ and all $\alpha \in [0,1]$, we see that

$$\begin{aligned} I(\alpha f_1 + (1-\alpha)f_2) &= \int_0^1 (\alpha f_1(x) + (1-\alpha)f_2(x))^2 dx \\ &\leq \int_0^1 \alpha(f_1(x))^2 + (1-\alpha)(f_2(x))^2 dx \text{ (using the convexity of } y \mapsto y^2) \\ &= \alpha \int_0^1 (f_1(x))^2 dx + (1-\alpha)\int_0^1 (f_2(x))^2 dx \\ &= \alpha I(f_1) + (1-\alpha)I(f_2). \end{aligned}$$

Thus I is convex. ◊

In order to prove the theorem on the sufficiency of the vanishing derivative in the case of a convex function, we will need the following result, which says that if a differentiable function f is convex, then its derivative f' is an increasing function, that is, if $x \leq y$, then $f'(x) \leq f'(y)$. (In Exercise 1 below, we will also prove a converse.)

Lemma 1.4.3 *If $f : \mathbb{R} \to \mathbb{R}$ is convex and differentiable, then f' is an increasing function.*

Proof Let $x < u < y$. If $\alpha = \frac{u-x}{y-1}$, then $\alpha \in (0,1)$, and $1 - \alpha = \frac{y-u}{y-x}$. From the convexity of f, we obtain

$$\frac{u-x}{y-x}f(y) + \frac{y-u}{y-x}f(x) \geq f\left(\frac{u-x}{y-x}y + \frac{y-u}{y-x}x\right) = f(u)$$

that is,

$$(y-x)f(u) \leq (u-x)f(y) + (y-u)f(x). \tag{1.35}$$

From (1.35), we obtain $(y-x)f(u) \leq (u-x)f(y) + (y-x+x-u)f(x)$, that is,

$$(y-x)f(u) - (y-x)f(x) \leq (u-x)f(y) - (u-x)f(x),$$

and so

$$\frac{f(u)-f(x)}{u-x} \leq \frac{f(y)-f(x)}{y-x}. \tag{1.36}$$

From (1.35), we also have $(y-x)f(u) \leq (u-y+y-x)f(y) + (y-u)f(x)$, that is,

$$(y-x)f(u) - (y-x)f(y) \leq (u-y)f(y) - (u-y)f(x),$$

and so

$$\frac{f(y)-f(x)}{y-x} \leq \frac{f(y)-f(u)}{y-u}. \tag{1.37}$$

Combining (1.36) and (1.37),

$$\frac{f(u)-f(x)}{u-x} \leq \frac{f(y)-f(x)}{y-x} \leq \frac{f(y)-f(u)}{y-u}.$$

Passing the limit as $u \searrow x$ and $u \nearrow y$, we obtain $f'(x) \leq \dfrac{f(y)-f(x)}{y-x} \leq f'(y)$, and so f' is increasing. ■

We are now ready to prove the result on the existence of global minima. First of all, we mention that if I is a function from a normed space X to $\mathbb{R}$, then I is said to have a *global minimum* at the point $x_0 \in X$ if for all $x \in X$, $I(x) \geq I(x_0)$. Similarly if $I(x) \leq I(x_0)$ for all x, then I is said to have a *global maximum* at x_0. We also note that the problem of finding a maximizer for a map I can always be converted to a minimization problem by considering $-I$ instead of I. We now prove the following.

Theorem 1.4.4 *Let X be a normed space and $I : X \to \mathbb{R}$ be differentiable. Suppose that I is convex. If $x_0 \in X$ is such that $(DI)(x_0) = 0$, then I has a global minimum at x_0.*

Proof Suppose that $x_1 \in X$ and $I(x_1) < I(x_0)$. Define $f : \mathbb{R} \to \mathbb{R}$ by

$$f(\alpha) = I(\alpha x_1 + (1-\alpha)x_0), \quad \alpha \in \mathbb{R}.$$

The function f is convex, since if $r \in [0,1]$ and $\alpha, \beta \in \mathbb{R}$, then we have

$$\begin{aligned} f(r\alpha + (1-r)\beta) &= I((r\alpha + (1-r)\beta)x_1 + (1 - r\alpha - (1-r)\beta)x_0) \\ &= I(r(\alpha x_1 + (1-\alpha)x_0) + (1-r)(\beta x_1 + (1-\beta)x_0)) \\ &\leq rI(\alpha x_1 + (1-\alpha)x_0) + (1-r)I(\beta x_1 + (1-\beta)x_0) \\ &= rf(\alpha) + (1-r)f(\beta). \end{aligned}$$

From Exercise 4a on page 19, it follows that f is differentiable on $[0,1]$, and

$$f'(0) = ((DI)(x_0))(x_1 - x_0) = 0.$$

Since $f(1) = I(x_1) < I(x_0) = f(0)$, by the mean value theorem[2], there exists a $c \in (0,1)$ such that

$$f'(c) = \frac{f(1) - f(0)}{1 - 0} < 0 = f'(0).$$

This contradicts the convexity of f (see Lemma 1.4.3 above), and so $I(x_1) \geq I(x_0)$. Hence I has a global minimum at x_0. ■

Exercises.

1. Prove that if $f : \mathbb{R} \to \mathbb{R}$ is twice continuously differentiable and $f''(x) > 0$ for all $x \in \mathbb{R}$, then f is convex.
2. Let X be a normed space, and $f \in \mathscr{L}(X, \mathbb{R})$. Show that f is convex.
3. If X is a normed space, then prove that the norm function, $x \mapsto \|x\| : X \to \mathbb{R}$, is a convex.
4. Let X be a normed space, and let $f : X \to \mathbb{R}$ be a function. Define the *epigraph* of f by

   $$U(f) = \bigcup_{x \in X} \{x\} \times (f(x), +\infty) \subset X \times \mathbb{R}.$$

   This is the 'region above the graph of f'. Show that if f is convex, then $U(f)$ is a convex subset of $X \times \mathbb{R}$. (See Exercise 6 on page 6 for the definition of a convex set).
5. $(*)$ Show that if $f : \mathbb{R} \to \mathbb{R}$ is convex, then for all $n \in \mathbb{N}$ and all $x_1, \ldots, x_n \in \mathbb{R}$, there holds that

   $$f\left(\frac{x_1 + \cdots + x_n}{n}\right) \leq \frac{f(x_1) + \cdots + f(x_n)}{n}.$$

1.4.4 An example of optimization in a function space

Example. A copper mining company intends to remove all of the copper ore from a region that contains an estimated Q tons, over a time period of T years. As it is extracted, they will sell it for processing at a net price per ton of

$$p(x(t), x'(t)) = P - ax(t) - bx'(t)$$

for positive constants P, a, and b, where $x(t)$ denotes the total tonnage sold by time t. (This pricing model allows the cost of mining to increase with the extent of the mined region and speed of production.)

[2]The mean value theorem says that if $f : [a,b] \to \mathbb{R}$ is a continuous function that is differentiable in (a,b), then there exists a $c \in (a,b)$ such that $\frac{f(b)-f(a)}{b-a} = f'(c)$.

If the company wishes to maximize its total profit given by

$$I(x) = \int_0^T p(x(t), x'(t))x'(t)dt, \tag{1.38}$$

where $x(0) = 0$ and $x(T) = Q$, how might it proceed?

STEP 1. First of all we note that the set of curves in $C^1[0,T]$ satisfying $x(a) = 0$ and $x(T) = Q$ do not form a linear space! So Theorem 1.4.2 is not applicable directly. Hence we introduce a new linear space X, and consider a new function $\tilde{I} : X \to \mathbb{R}$ which is defined in terms of the old function I.

Introduce the linear space $X = \{x \in C^1[0,T] \mid x(0) = x(T) = 0\}$, with the $C^1[0,T]$-norm:

$$\|x\| = \sup_{t\in[0,T]} |x(t)| + \sup_{t\in[0,T]} |x'(t)|.$$

Then for all $h \in X$, x_0+h satisfies $(x_0+h)(0) = 0$ and $(x_0+h)(T) = Q$. Defining $\tilde{I}(h) = I(x_0+h)$, we note that $\tilde{I} : X \to \mathbb{R}$ has an extremum at 0. It follows from Theorem 1.4.2 that $(D\tilde{I})(0) = 0$. Note that by the 0 in the right hand side of the equality, we mean the zero functional, namely the continuous linear map from X to $\mathbb{R}$, which is defined by $h \mapsto 0$ for all $h \in X$.

STEP 2. We now calculate $\tilde{I}'(0)$. We have

$$\begin{aligned}\tilde{I}(h) - \tilde{I}(0) &= \int_0^T P - a(x_0(t) + h(t)) - b(x_0'(t) + h'(t))dt - \int_0^T P - ax_0(t) - bx_0(t)dt \\ &= \int_0^T P - ax_0(t) - 2bx_0'(t))h'(t) - ax_0'(t)h(t)dt + \int_0^T -ah(t)h'(t) - bh'(t)h'(t)dt.\end{aligned}$$

Since the map

$$h \mapsto \int_0^T (P - ax_0(t) - 2bx_0'(t))h'(t) - ax_0'(t)h(t)dt$$

is a functional from X to $\mathbb{R}$ and since

$$\left| \int_0^T -ah(t)h'(t) - bh'(t)h'(t)dt \right| \leq T(a+b)\|h\|^2,$$

it follows that

$$[(D\tilde{I})(0)](h) = \int_0^T (P - ax_0(t) - 2bx_0'(t))h'(t) - ax_0'(t)h(t)dt = \int_0^T (P - 2bx_0'(t))h'(t)dt,$$

where the last equality follows using partial integration:

$$\int_0^T ax_0'(t)h(t)dt = -\int_0^T ax_0(t)h'(t)dt + ax_0(t)h(t)|_{t=0}^T = -\int_0^T ax_0(t)h'(t)dt.$$

STEP 3. Since $(D\tilde{I})(0) = 0$, it follows that

$$\int_0^T \left(P - ax_0(t) - 2bx_0'(t) - a\int_0^t x_0'(\tau)d\tau \right) h'(t)dt = 0$$

for all $h \in C^1[0,T]$ with $h(0) = h(T) = 0$. We now prove the following.

Lemma 1.4.5 *If $k \in C[a,b]$ and*

$$\int_a^b k(t)h'(t)dt = 0$$

for all $h \in C^1[a,b]$ with $h(a) = h(b) = 0$, then there exists a constant c such that $k(t) = c$ for all $t \in [a,b]$.

Proof Define the constant c and the function h via

$$\int_a^b (k(t) - c)dt = 0 \quad \text{and} \quad h(t) = \int_a^t (k(\tau) - c)d\tau.$$

Then $h \in C^1[a,b]$ and it satisfies $h(a) = h(b) = 0$. Furthermore,

$$\int_a^b (k(t) - c)^2 dt = \int_a^b (k(t) - c)h'(t)dt = \int_a^b k(t)h'(t)dt - c(h(b) - h(a)) = 0.$$

Thus $k(t) - c = 0$ for all $t \in [a,b]$. ■

Step 4. The above result implies in our case that

$$\forall t \in [0,T], \quad P - 2bx_0'(t) = c. \tag{1.39}$$

Integrating, we obtain $x_0(t) = At + B$, $t \in [0,T]$, for some constants A and B. Using $x_0(0) = 0$ and $x_0(T) = Q$, we obtain $x_0(t) = \frac{t}{T}Q$, $t \in [0,T]$. This is the optimal mining operation.

Step 5. Finally we show that this is the optimal mining operation, that is $I(x_0) \geq I(x)$ for all x such that $x(0) = 0$ and $x(T) = Q$. We prove this by showing $-\tilde{I}$ is convex, and so by Theorem 1.4.4, $-\tilde{I}$ in fact has a global minimum at 0.

Let $h_1, h_2 \in X$, and $\alpha \in [0,1]$, and define $x_1 = x_0 + h_1$, $x_2 = x_0 + h_2$. Then we have

$$\int_0^T (\alpha x_1'(t) + (1-\alpha)x_2'(t))^2 dt \leq \int_0^T \alpha(x_1'(t))^2 + (1-\alpha)(x_2'(t))^2 dt, \tag{1.40}$$

using the convexity of $y \mapsto y^2$. Furthermore, $x_1(0) = 0 = x_2(0)$ and $x_1(T) = Q = x_2(T)$, and so

$$\begin{aligned}
& \int_0^T (\alpha x_1'(t) + (1-\alpha)x_2'(t))(\alpha x_1(t) + (1-\alpha)x_2(t))dt \\
= \; & \frac{1}{2}\int_0^T \frac{d}{dt}(\alpha x_1(t) + (1-\alpha)x_2(t))^2 dt \\
= \; & \frac{1}{2}Q^2 = \alpha\frac{1}{2}Q^2 + (1-\alpha)\frac{1}{2}Q^2 \\
= \; & \alpha\int_0^T x_1'(t)x_1(t)dt + (1-\alpha)\int_0^T x_2'(t)x_2(t)dt.
\end{aligned}$$

Hence

$$\begin{aligned}
-\tilde{I}(\alpha h_1 + (1-\alpha)h_2) &= -I(x_0 + \alpha h_1 + (1-\alpha)h_2) \\
&= -I(\alpha x_0 + (1-\alpha)x_0 + \alpha h_1 + (1-\alpha)h_2) \\
&= -I(\alpha x_1 + (1-\alpha)x_2) \\
&= b\int_0^T (\alpha x_1'(t) + (1-\alpha)x_2'(t))^2 dt \\
&\quad + a\int_0^T (\alpha x_1'(t) + (1-\alpha)x_2'(t))(\alpha x_1(t) + (1-\alpha)x_2(t))dt \\
&\quad - P\int_0^T (\alpha x_1'(t) + (1-\alpha)x_2'(t))dt \\
&\leq \alpha\int_0^T (x_1'(t))^2 dt + (1-\alpha)\int_0^T (x_2'(t))^2 dt \\
&\quad + \alpha\int_0^T x_1'(t)x_1(t)dt + (1-\alpha)\int_0^T x_2'(t)x_2(t)dt \\
&\quad - \alpha P\int_0^T x_1'(t)dt - (1-\alpha)P\int_0^T x_2'(t)dt \\
&= \alpha\left(\int_0^T x_1'(t)(bx_1'(t) + ax_1(t) - P)dt\right) \\
&\quad + (1-\alpha)\left(\int_0^T x_2'(t)(bx_2'(t) + ax_2(t) - P)dt\right) \\
&= \alpha(-I(x_1)) + (1-\alpha)(-I(x_2)) = \alpha(-\tilde{I}(h_1)) + (1-\alpha)(-\tilde{I}(h_2)).
\end{aligned}$$

Hence $-\tilde{I}$ is convex. ◊

The above optimization problem is a special case of the following general problem, which we will consider in the next chapter.

Let I be a function of the form

$$I(x) = \int_a^b F\left(x(t), \frac{dx}{dt}(t), t\right) dt,$$

where $F(\alpha, \beta, \gamma)$ is a 'nice' function and $x \in C^1[a,b]$ is such that $x(a) = y_a$ and $x(b) = y_b$. Then proceeding in a similar manner as above, it can be shown that if I has an extremum at x_0, then x_0 satisfies the *Euler-Lagrange equation*:

$$\frac{\partial F}{\partial \alpha}\left(x_0(t), \frac{dx_0}{dt}(t), t\right) - \frac{d}{dt}\left(\frac{\partial F}{\partial \beta}\left(x_0(t), \frac{dx_0}{dt}(t), t\right)\right) = 0, \quad t \in [a,b]. \tag{1.41}$$

(This equation is abbreviated by $F_x - \frac{d}{dt}F_{x'} = 0$.)

Chapter 2

The Euler-Lagrange equation

In this chapter, we will give necessary conditions for an extremum of a function of the type

$$I(x) = \int_a^b F\left(x(t), x'(t), t\right) dt,$$

with various types of boundary conditions. The necessary condition is in the form of a differential equation that the extremal curve should satisfy, and this differential equation is called the *Euler-Lagrange* equation.

We begin with the simplest type of boundary conditions, where the curves are allowed to vary between two fixed points.

2.1 The simplest optimisation problem

The *simplest optimisation problem* can be formulated as follows:

Let $F(\alpha, \beta, \gamma)$ be a function with continuous first and second partial derivatives with respect to (α, β, γ). Then find $x \in C^1[a, b]$ such that $x(a) = y_a$ and $x(b) = y_b$, and which is an extremum for the function

$$I(x) = \int_a^b F\left(x(t), x'(t), t\right) dt. \tag{2.1}$$

In other words, the simplest optimisation problem consists of finding an extremum of a function of the form (2.5), where the class of admissible curves comprises all smooth curves joining two *fixed* points; see Figure 2.1. We will apply the necessary condition for an extremum (established in Theorem 1.4.2) to the solve the simplest optimisation problem described above.

Theorem 2.1.1 *Let $S = \{x \in C^1[a, b] \mid x(a) = y_a$ and $x(b) = y_b\}$, and let $I : S \to \mathbb{R}$ be a function of the form*

$$I(x) = \int_a^b F\left(x(t), x'(t), t\right) dt.$$

If I has an extremum at $x_0 \in S$, then x_0 satisfies the Euler-Lagrange equation*:*

$$\frac{\partial F}{\partial \alpha}\left(x_0(t), x_0'(t), t\right) - \frac{d}{dt}\left(\frac{\partial F}{\partial \beta}\left(x_0(t), x_0'(t), t\right)\right) = 0, \quad t \in [a, b]. \tag{2.2}$$

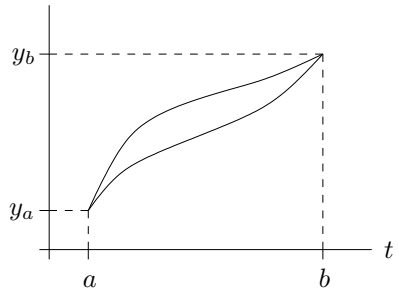


Figure 2.1: Possible paths joining the two fixed points (a, y_a) and (b, y_b).

Proof The proof is long and so we divide it into several steps.

STEP 1. First of all we note that the set S is not a vector space (unless $y_a = 0 = y_b$)! So Theorem 1.4.2 is not applicable directly. Hence we introduce a new linear space X, and consider a new function $\tilde{I} : X \to \mathbb{R}$ which is defined in terms of the old function I.

Introduce the linear space

$$X = \{x \in C^1[a, b] \mid x(a) = x(b) = 0\},$$

with the induced norm from $C^1[a, b]$. Then for all $h \in X$, $x_0 + h$ satisfies $(x_0 + h)(a) = y_a$ and $(x_0 + h)(b) = y_b$. Defining $\tilde{I}(h) = I(x_0 + h)$, for $h \in X$, we note that $\tilde{I} : X \to \mathbb{R}$ has a local extremum at 0. It follows from Theorem 1.4.2 that[1] $D\tilde{I}(0) = 0$.

STEP 2. We now calculate $D\tilde{I}(0)$. We have

$$\begin{aligned}\tilde{I}(h) - \tilde{I}(0) &= \int_a^b F\left((x_0 + h)(t), (x_0 + h)'(t), t\right) dt - \int_a^b F\left(x_0(t), x_0'(t), t\right) dt \\ &= \int_a^b \left[F\left(x_0(t) + h(t), x_0'(t) + h'(t), t\right) dt - F\left(x_0(t), x_0'(t), t\right)\right] dt.\end{aligned}$$

Recall that from Taylor's theorem, if F possesses partial derivatives of order 2 in a ball B of radius r around the point $(\alpha_0, \beta_0, \gamma_0)$ in $\mathbb{R}^3$, then for all $(\alpha, \beta, \gamma) \in B$, there exists a $\Theta \in [0, 1]$ such that

$$\begin{aligned}F(\alpha, \beta, \gamma) &= F(\alpha_0, \beta_0, \gamma_0) + \left((\alpha - \alpha_0)\frac{\partial}{\partial \alpha} + (\beta - \beta_0)\frac{\partial}{\partial \beta} + (\gamma - \gamma_0)\frac{\partial}{\partial \gamma}\right) F\Big|_{(\alpha_0, \beta_0, \gamma_0)} + \\ &\quad \frac{1}{2!}\left((\alpha - \alpha_0)\frac{\partial}{\partial \alpha} + (\beta - \beta_0)\frac{\partial}{\partial \beta} + (\gamma - \gamma_0)\frac{\partial}{\partial \gamma}\right)^2 F\Big|_{(\alpha_0, \beta_0, \gamma_0) + \Theta((\alpha, \beta, \gamma) - (\alpha_0, \beta_0, \gamma_0))}.\end{aligned}$$

Hence for $h \in X$ such that $\|h\|$ is small enough,

$$\begin{aligned}\tilde{I}(h) - \tilde{I}(0) &= \int_a^b \left[\frac{\partial F}{\partial \alpha}\left(x_0(t), x_0'(t), t\right) h(t) + \frac{\partial F}{\partial \beta}\left(x_0(t), x_0'(t), t\right) h'(t)\right] dt + \\ &\quad \frac{1}{2!}\int_a^b \left(h(t)\frac{\partial}{\partial \alpha} + h'(t)\frac{\partial}{\partial \beta}\right)^2 F\Big|_{(x_0(t) + \Theta(t)h(t), x_0'(t) + \Theta(t)h'(t), t)} dt.\end{aligned}$$

It can be checked that there exists a $M > 0$ such that

$$\left|\frac{1}{2!}\int_a^b \left(h(t)\frac{\partial}{\partial \alpha} + h'(t)\frac{\partial}{\partial \beta}\right)^2 F\Big|_{(x_0(t) + \Theta(t)h(t), x_0'(t) + \Theta(t)h'(t), t)} dt\right| \leq M\|h\|^2,$$

[1] Note that by the 0 in the right hand side of the equality, we mean the zero map, namely the continuous linear map from X to $\mathbb{R}$, which is defined by $h \mapsto 0$ for all $h \in X$.

and so $D\tilde{I}(0)$ is the map

$$h \mapsto \int_a^b \left[\frac{\partial F}{\partial \alpha}\left(x_0(t), x_0'(t), t\right) h(t) + \frac{\partial F}{\partial \beta}\left(x_0(t), x_0'(t), t\right) h'(t) \right] dt. \tag{2.3}$$

Step 3. Next we show that if the map in (2.3) is the zero map, then this implies that (2.2) holds. Define

$$A(t) = \int_a^t \frac{\partial F}{\partial \alpha}\left(x_0(\tau), x_0'(\tau), \tau\right) d\tau.$$

Integrating by parts, we find that

$$\int_a^b \frac{\partial F}{\partial \alpha}\left(x_0(t), x_0'(t), t\right) h(t) dt = -\int_a^b A(t) h'(t) dt,$$

and so from (2.3), it follows that $D\tilde{I}(0) = 0$ implies that

$$\int_a^b \left[-A(t) + \frac{\partial F}{\partial \beta}\left(x_0(t), x_0'(t), t\right) \right] h'(t) dt = 0 \text{ for all } h \in X.$$

Step 4. Finally, using Lemma 1.4.5, we obtain

$$-A(t) + \frac{\partial F}{\partial \beta}\left(x_0(t), x_0'(t), t\right) = k \text{ for all } t \in [a, b].$$

Differentiating with respect to t, we obtain (2.3). This completes the proof of Theorem 2.1.1. ■

Note that the Euler-Lagrange equation is only a *necessary* condition for the existence of an extremum (see the remark following Theorem 1.4.2). However, in many cases, the Euler-Lagrange equation by itself is enough to give a complete solution of the problem. In fact, the existence of an extremum is sometimes clear from the context of the problem. If in such scenarios, there exists only one solution to the Euler-Lagrange equation, then this solution must a fortiori be the point for which the extremum is achieved.

Example. Let $S = \{x \in C^1[0,1] \mid x(0) = 0 \text{ and } x(1) = 1\}$. Consider the function $I : S \to \mathbb{R}$ given by

$$I(x) = \int_0^1 \left(\frac{d}{dt} x(t) - 1 \right)^2 dt.$$

We wish to find $x_0 \in S$ that minimizes I. We proceed as follows:

Step 1. We have $F(\alpha, \beta, \gamma) = (\beta - 1)^2$, and so $\dfrac{\partial F}{\partial \alpha} = 0$ and $\dfrac{\partial F}{\partial \beta} = 2(\beta - 1)$.

Step 2. The Euler-Lagrange equation (2.2) is now given by

$$0 - \frac{d}{dt}(2(x_0'(t) - 1)) = 0 \quad \text{for all } t \in [0, 1].$$

Step 3. Integrating , we obtain $2(x_0'(t) - 1) = C$, for some constant C, and so $x_0' = \frac{C}{2} + 1 =: A$. Integrating again, we have $x_0(t) = At + B$, where A and B are suitable constants.

STEP 4. The constants A and B can be determined by using that fact that $x_0 \in S$, and so $x_0(0) = 0$ and $x_0(a) = 1$. Thus we have

$$\begin{aligned} A0 + B &= 0, \\ A1 + B &= 1, \end{aligned}$$

which yield $A = 1$ and $B = 0$.

So the unique solution x_0 of the Euler-Lagrange equation in S is $x_0(t) = t$, $t \in [0, 1]$; see Figure 2.2.

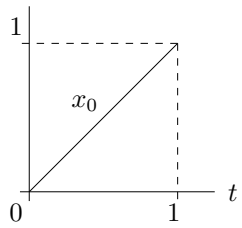


Figure 2.2: Minimizer for I.

Now we argue that the solution x_0 indeed minimizes I. Since $(x'(t) - 1)^2 \geq 0$ for all $t \in [0, 1]$, it follows that $I(x) \geq 0$ for all $x \in C^1[0, 1]$. But

$$I(x_0) = \int_0^1 (x_0'(t) - 1)^2 dt = \int_0^1 (1 - 1)^2 dt = \int_0^1 0 dt = 0.$$

As $I(x) \geq 0 = I(x_0)$ for all $x \in S$, it follows that x_0 minimizes I. ◊

Definition. The solutions of the Euler-Lagrange equation (2.3) are called *critical curves*.

The Euler-Lagrange equation is in general a second order differential equation, but in some special cases, it can be reduced to a first order differential equation or where its solution can be obtained entirely by evaluating integrals. We indicate some special cases in Exercise 3 on page 31, where in each instance, F is independent of one of its arguments.

Exercises.

1. Let $S = \{x \in C^1[0, 1] \mid x(0) = 0 = x(1)\}$. Consider the map $I : S \to \mathbb{R}$ given by

   $$I(x) = \int_0^1 (x(t))^3 dt, \quad x \in S.$$

   Using Theorem 2.1.1, find the critical curve $x_0 \in S$ for I. Does I have a local extremum at x_0?

2. Write the Euler-Lagrange equation when F is given by

   (a) $F(\alpha, \beta, \gamma) = \sin \beta$,
   (b) $F(\alpha, \beta, \gamma) = \alpha^3 \beta^3$,
   (c) $F(\alpha, \beta, \gamma) = \alpha^2 - \beta^2$,
   (d) $F(\alpha, \beta, \gamma) = 2\gamma\beta - \beta^2 + 3\beta\alpha^2$.

3. Prove that:

   (a) If $F(\alpha, \beta, \gamma)$ does not depend on α, then the Euler-Lagrange equation becomes

   $$\frac{\partial F}{\partial \beta}(x_0(t), x_0'(t), t) = C,$$

   where C is a constant.

   (b) If F does not depend on β, then the Euler-Lagrange equation becomes

   $$\frac{\partial F}{\partial \alpha}(x_0(t), x_0'(t), t) = 0.$$

   (c) If F does not depend on γ and if x_0 is twice-differentiable in $[a, b]$, then the Euler-Lagrange equation becomes

   $$F(x_0(t), x_0'(t), t) - x_0'(t)\frac{\partial F}{\partial \beta}(x_0(t), x_0'(t), t) = C,$$

   where C is a constant.
   Hint: What is $\frac{d}{dt}\left(F(x_0(t), x_0'(t), t) - x_0'(t)\frac{\partial F}{\partial \beta}(x_0(t), x_0'(t), t)\right)$?

4. Find the curve which has minimum length between $(0, 0)$ and $(1, 1)$.

5. Let $S = \{x \in C^1[0, 1] \mid x(0) = 0 \text{ and } x(1) = 1\}$. Find critical curves in S for the functions $I : S \to \mathbb{R}$, where I is given by:

   (a) $I(x) = \int_0^1 x'(t)dt$

   (b) $I(x) = \int_0^1 x(t)x'(t)dt$

   (c) $I(x) = \int_0^1 (x(t) + tx'(t))dt$

   for $x \in S$.

6. Find critical curves for the function

   $$I(x) = \int_1^2 t^3(x'(t))^2 dt$$

   where $x \in C^1[1, 2]$ with $x(1) = 5$ and $x(2) = 2$.

7. Find critical curves for the function

   $$I(x) = \int_1^2 \frac{(x'(t))^3}{t^2} dt$$

   where $x \in C^1[1, 2]$ with $x(1) = 1$ and $x(2) = 7$.

8. Find critical curves for the function

   $$I(x) = \int_0^1 \left[2tx(t) - (x'(t))^2 + 3x'(t)(x(t))^2\right] dt$$

   where $x \in C^1[0, 1]$ with $x(0) = 0$ and $x(1) = -1$.

9. Find critical curves for the function

$$I(x) = \int_0^1 \left[2(x(t))^3 + 3t^2 x'(t)\right] dt$$

where $x \in C^1[0,1]$ with $x(0) = 0$ and $x(1) = 1$. What if $x(0) = 0$ and $x(1) = 2$?

10. Consider the copper mining company mentioned in the introduction. If future money is discounted continuously at a constant rate δ, then we can assess the present value of profits from this mining operation by introducing a factor of $e^{-\delta t}$ in the integrand of (1.38). Suppose that $a = 4$, $b = 1$, $\delta = 1$ and $P = 2$. Find a critical mining operation x_0 such that $x_0(0) = 0$ and $x_0(T) = Q$.

11. Consider the quadratic function $q : \mathbb{R} \to \mathbb{R}$ given by $q(r) = ar^2 + br + c$ $(r \in \mathbb{R})$ with $a > 0$. It is easy to see that the minimum value of q is $\dfrac{4ac - b^2}{4a}$.

Let x_1, x_2 be fixed functions in $C[0,1]$. Regarding the left hand side of the obvious inequality

$$\int_0^1 (x_1(t) + r x_2(t))^2 dt \geq 0$$

as a quadratic function q of r, with

$$a = \int_0^1 (x_2(t))^2 dt, \quad b = 2\int_0^1 x_1(t)x_2(t)dt, \quad c = \int_0^1 (x_1(t))^2 dt,$$

it follows that $\dfrac{4ac - b^2}{4a} \geq 0$, that is, the *Cauchy-Schwarz inequality* holds:

$$\left(\int_0^1 (x_1(t))^2 dt\right)\left(\int_0^1 (x_2(t))^2 dt\right) \geq \left(\int_0^1 x_1(t)x_2(t)dt\right)^2.$$

(a) Let $S = \{x \in C^1[0,1] \mid x(0) = 0 \text{ and } x(1) = 1\}$. Consider the function $I : S \to \mathbb{R}$ defined by

$$I(x) = \int_0^1 e^{-2t} \left(x'(t)\right)^2 dt, \quad x \in S.$$

Using the Cauchy-Schwarz inequality, show that

$$I(x) \geq \frac{2}{e^2 - 1}.$$

HINT: Take $x_1(t) = e^t$ and $x_2(t) = e^{-t}x'(t)$ for $t \in [0,1]$.

(b) Using the Euler-Lagrange equation, find a critical curve x_0 for I.

(c) Find $I(x_0)$, where x_0 denotes the critical curve found in part 11b. Using part 11a show that x_0 indeed minimizes the function I.

2.2 Calculus of variations: some classical problems

Optimisation problems of the type considered in the previous section were studied in various special cases by many leading mathematicians in the past. These were often solved by various techniques, and these gave rise to the branch of mathematics known as the 'calculus of variations'. The name comes from the fact that often the procedure involved the calculation of the 'variation' in the function I when its argument (which was typically a curve) was changed, and then passing limits. In this section, we mention two classical problems, and indicate how these can be solved using the Euler-Lagrange equation.

2.2.1 The brachistochrone problem

The calculus of variations originated from a problem posed by the Swiss mathematician Johann Bernoulli (1667-1748). He required the form of the curve joining two fixed points A and B in a vertical plane such that a body sliding down the curve (under gravity and no friction) travels from A to B in minimum time. This problem does not have a trivial solution; the straight line from A to B is not the solution (this is also intuitively clear, since if the slope is high at the beginning, the body picks up a high velocity and so its plausible that the travel time could be reduced) and it can be verified experimentally by sliding beads down wires in various shapes.

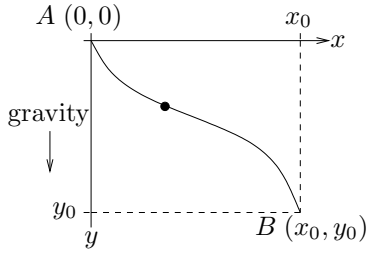


Figure 2.3: The brachistochrone problem.

To pose the problem in mathematical terms, we introduce coordinates as shown in Figure 2.3, so that A is the point $(0,0)$, and B corresponds to (x_0, y_0). Assuming that the particle is released from rest at A, conservation of energy gives $\frac{1}{2}mv^2 - mgy = 0$, where we have taken the zero potential energy level at $y = 0$, and where v denotes the speed of the particle. Thus the speed is given by $v = \frac{ds}{dt} = \sqrt{2gy}$, where s denotes arc length along the curve. From Figure 2.4, we see that an element of arc length, δs is given approximately by $((\delta x)^2 + (\delta y)^2)^{\frac{1}{2}}$.

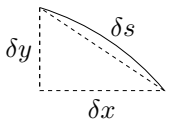


Figure 2.4: Element of arc length.

Hence the time of descent is given by

$$T = \int_{\text{curve}} \frac{ds}{\sqrt{2gy}} = \frac{1}{\sqrt{2g}} \int_0^{y_0} \sqrt{\frac{1 + \left(\frac{dx}{dy}\right)^2}{y}} dy.$$

Our problem is to find the path $\{x(y),\ y \in [0, y_0]\}$, satisfying $x(0) = 0$ and $x(y_0) = x_0$, which minimizes T, that is, to determine the minimizer for the function $I : S \to \mathbb{R}$, where

$$I(x) = \frac{1}{\sqrt{2g}} \int_0^{y_0} \left(\frac{1 + (x'(y))^2}{y}\right)^{\frac{1}{2}} dy, \quad x \in S,$$

and $S = \{x \in C^1[0, y_0] \mid x(0) = 0 \text{ and } x(y_0) = x_0\}$. Here[2] $F(\alpha, \beta, \gamma) = \sqrt{\frac{1+\beta^2}{\gamma}}$ is independent of α, and so the Euler-Lagrange equation becomes

$$\frac{d}{dy}\left(\frac{x'(y)}{\sqrt{1 + (x'(y))^2}} \frac{1}{\sqrt{y}}\right) = 0.$$

[2]Strictly speaking, the F here does *not* satisfy the demands made in Theorem 2.1.1. Notwithstanding this fact, with some additional argument, the solution given here can be fully justified.

Integrating with respect to y, we obtain

$$\frac{x'(y)}{\sqrt{1+(x'(y))^2}}\frac{1}{\sqrt{y}} = C,$$

where C is a constant. It can be shown that the general solution of this differential equation is given by

$$x(\Theta) = \frac{1}{2C^2}(\Theta - \sin\Theta) + \tilde{C}, \quad y(\Theta) = \frac{1}{2C^2}(1-\cos\Theta),$$

where $\tilde{C}$ is another constant. The constants are chosen so that the curve passes through the points $(0,0)$ and (x_0, y_0).

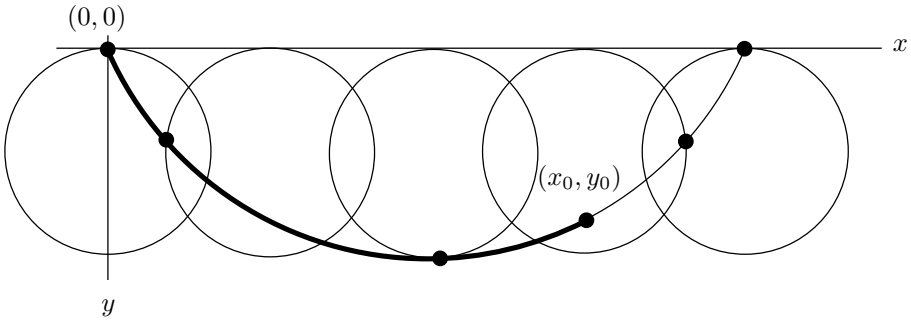


Figure 2.5: The cycloid through $(0,0)$ and (x_0, y_0).

This curve is known as a *cycloid*, and in fact it is the curve described by a point P in a circle that rolls without slipping on the x axis, in such a way that P passes through (x_0, y_0); see Figure 2.5.

Exercise. The new London mayor, devoted to schemes for energy saving, wishes to design a fuel-less transport system driven by gravity. The proposal is that carriages should travel in frictionless underground tunnels being released from rest at their point of departure A (Waterloo) and then allowed to run freely until arriving at destination B (Paddington). Assuming gravity is uniform, show that the minimum travel time between two points (which are at the same level) and distance l apart is $\sqrt{\frac{2\pi l}{g}}$.

2.2.2 Minimum surface area of revolution

The problem of minimum surface area of revolution is to find among all the curves joining two given points (x_0, y_0) and (x_1, y_1), the one which generates the surface of minimum area when rotated about the x axis.

The area of the surface of revolution generated by rotating the curve y about the x axis is

$$S(y) = 2\pi \int_{x_0}^{x_1} y(x)\sqrt{1+(y'(x))^2}\,dx.$$

Since the integrand does not depend explicitly on x, the Euler-Lagrange equation is

$$F(y(x), y'(x), x) - y'(x)\frac{\partial F}{\partial \beta}(y(x), y'(x), x) = C,$$

where C is a constant, that is,

$$y\sqrt{1+(y')^2} - y\frac{(y')^2}{\sqrt{1+(y')^2}} = C.$$

Thus $y = C\sqrt{1+(y')^2}$, and it can be shown that this differential equation has the general solution

$$y(x) = C\cosh\left(\frac{x+C_1}{C}\right). \tag{2.4}$$

This curve is called a *catenary*. The values of the arbitrary constants C and C_1 are determined by the conditions $y(x_0) = y_0$ and $y(x_1) = y_1$. It can be shown that the following three cases are possible, depending on the positions of the points (x_0, y_0) and (x_1, y_1):

1. If a single curve of the form (2.4) passes through the points (x_0, y_0) and (x_1, y_1), then this curve is the solution of the problem; see Figure 2.6.

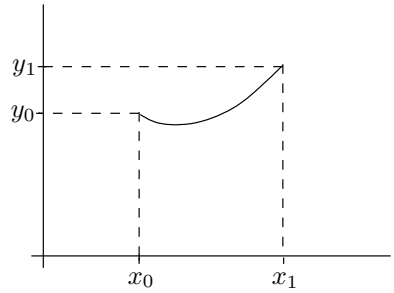


Figure 2.6: The catenary through (x_0, y_0) and (x_1, y_1).

2. If two critical curves can be drawn through the points (x_0, y_0) and (x_1, y_1), then one of the curves actually corresponds to the surface of revolution if minimum area, and the other does not.

3. If there does not exist a curve of the form (2.4) passing through the points (x_0, y_0) and (x_1, y_1), then there is no surface in the class of smooth surfaces of revolution which achieves the minimum area. In fact, if the location of the two points is such that the distance between them is sufficiently large compared to their distances from the x axis, then the area of the surface consisting of two circles of radius y_0 and y_1 will be less than the area of any surface of revolution generated by a smooth curve passing through the points; see Figure 2.7.

Figure 2.7: The polygonal curve $(x_0, y_0) - (x_0, 0) - (x_1, 0) - (x_1, y_1)$.

This is intuitively expected: imagine a soap bubble between concentric rings which are being pulled apart. Initially we get a soap bubble between these rings, but if the distance separating the rings becomes too large, then the soap bubble breaks, leaving soap films on each of the two rings. This example shows that a critical curve need not always exist in the class of curves under consideration.

2.3 Free boundary conditions

Besides the simplest optimisation problem considered in the previous section, we now consider the optimisation problem with *free boundary conditions* (see Figure 2.8).

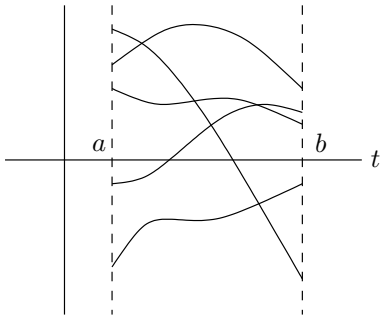


Figure 2.8: Free boundary conditions.

Let $F(\alpha, \beta, \gamma)$ be a function with continuous first and second partial derivatives with respect to (α, β, γ). Then find $x \in C^1[a, b]$ which is an extremum for the function

$$I(x) = \int_a^b F\left(x(t), \frac{dx}{dt}(t), t\right) dt. \tag{2.5}$$

Theorem 2.3.1 *Let $I : C^1[a, b] \to \mathbb{R}$ be a function of the form*

$$I(x) = \int_a^b F\left(x(t), \frac{dx}{dt}(t), t\right) dt, \quad x \in C^1[a, b],$$

where $F(\alpha, \beta, \gamma)$ is a function with continuous first and second partial derivatives with respect to (α, β, γ). If I has a local extremum at x_0, then x_0 satisfies the Euler-Lagrange equation:

$$\frac{\partial F}{\partial \alpha}\left(x_0(t), \frac{dx_0}{dt}(t), t\right) - \frac{d}{dt}\left(\frac{\partial F}{\partial \beta}\left(x_0(t), \frac{dx_0}{dt}(t), t\right)\right) = 0, \quad t \in [a, b], \tag{2.6}$$

together with the transversality conditions

$$\frac{\partial F}{\partial \beta}\left(x_0(t), \frac{dx_0}{dt}(t), t\right)\Big|_{t=a} = 0 \quad \textit{and} \quad \frac{\partial F}{\partial \beta}\left(x_0(t), \frac{dx_0}{dt}(t), t\right)\Big|_{t=b} = 0. \tag{2.7}$$

Proof

Step 1. We take $X = C^1[a, b]$ and compute $DI(x_0)$. Proceeding as in the proof of Theorem 2.1.1, it is easy to see that

$$DI(x_0)(h) = \int_a^b \left[\frac{\partial F}{\partial \alpha}\left(x_0(t), x_0'(t), t\right) h(t) + \frac{\partial F}{\partial \beta}\left(x_0(t), x_0'(t), t\right) h'(t)\right] dt,$$

$h \in C^1[a, b]$. Theorem 1.4.2 implies that this linear functional must be the zero map, that is, $(DI(x_0))(h) = 0$ for all $h \in C^1[a, b]$. In particular, it is also zero for all h in $C^1[a, b]$ such that $h(a) = h(b) = 0$. But recall that in Step 3 and Step 4 of the proof of Theorem 2.1.1, we proved that if

$$\int_a^b \left[\frac{\partial F}{\partial \alpha}\left(x_0(t), x_0'(t), t\right) h(t) + \frac{\partial F}{\partial \beta}\left(x_0(t), x_0'(t), t\right) h'(t)\right] dt = 0 \tag{2.8}$$

for all h in $C^1[a,b]$ such that $h(a) = h(b) = 0$, then this implies that he Euler-Lagrange equation (2.6) holds.

STEP 2. Integration by parts in (2.8) now gives

$$\begin{aligned} DI(x_0)(h) &= \int_a^b \left[\frac{\partial F}{\partial \alpha}(x_0(t), x_0'(t), t) - \frac{d}{dt}\left(\frac{\partial F}{\partial \beta}(x_0(t), x_0'(t), t)\right)\right] h(t)dt + \qquad (2.9) \\ &\quad \frac{\partial F}{\partial \beta}(x_0(t), x_0'(t), t)\, h(t)\Big|_{t=a}^{t=b} \qquad (2.10) \\ &= 0 + \frac{\partial F}{\partial \beta}(x_0(t), x_0'(t), t)\Big|_{t=b} h(b) - \frac{\partial F}{\partial \beta}(x_0(t), x_0'(t), t)\Big|_{t=a} h(a). \end{aligned}$$

The integral in (2.9) vanishes since we have shown in STEP 1 above that (2.6) holds. Thus the condition $DI(x_0) = 0$ now takes the form

$$\frac{\partial F}{\partial \beta}(x_0(t), x_0'(t), t)\Big|_{t=b} h(b) - \frac{\partial F}{\partial \beta}(x_0(t), x_0'(t), t)\Big|_{t=a} h(a) = 0,$$

from which (2.7) follows, since h is arbitrary. This completes the proof. ■

Exercises.

1. Find all curves $y = y(x)$ which have minimum length between the lines $x = 0$ and the line $x = 1$.
2. Find critical curves for the following function, when the values of x are free at the endpoints:

$$I(x) = \int_0^1 \frac{1}{2}\left[(x'(t))^2 + x(t)x'(t) + x'(t) + x(t)\right] dt.$$

Similarly, we can also consider the *mixed* case (see Figure 2.9), when one end of the curve is fixed, say $x(a) = y_a$, and the other end is free. Then it can be shown that the curve x satisfies the Euler-Lagrange equation, the transversality condition

$$\frac{\partial F}{\partial \beta}(x_0(t), x_0'(t), t)\Big|_{t=a} h(a) = 0$$

at the free end point, and $x(a) = y_a$ serves as the other boundary condition.

We can summarize the results by the following: critical curves for (2.5) satisfy the Euler-Lagrange equation (2.6) and moreover there holds

$$\frac{\partial F}{\partial \beta}(x_0(t), x_0'(t), t) = 0 \text{ at the free end point.}$$

Exercises.

1. Find the curve $y = y(x)$ which has minimum length between $(0,0)$ and the line $x = 1$.
2. The cost of a manufacturing process in an industry is described by the function I given by

$$I(x) = \int_0^1 \left[\frac{1}{2}(x'(t))^2 + x(t)\right] dt,$$

for $x \in C^1[0,1]$ with $x(0) = 1$.

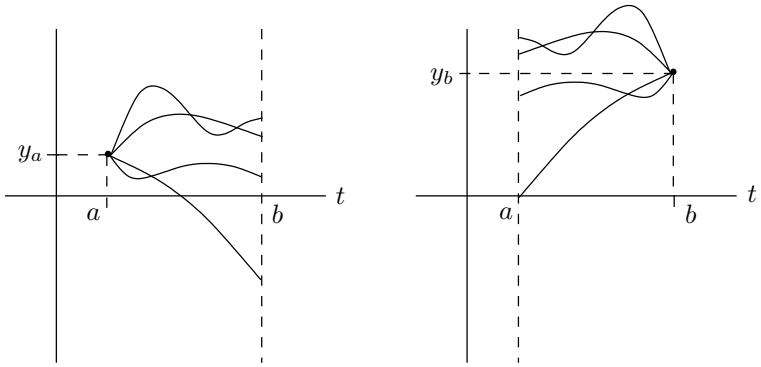


Figure 2.9: Mixed cases.

(a) If also $x(1) = 0$, find a critical curve x_* for I, and find $I(x_*)$.

(b) If $x(1)$ is not specified, find a critical curve x_{**} for I, and find $I(x_{**})$.

(c) Which of the values $I(x_*)$ and $I(x_{**})$ found in parts above is larger? Explain why you would expect this, assuming that x_* and x_{**} in fact minimize I on the respective domains specified above.

3. Find critical curves for the following functions:

(a) $I(x) = \int_0^{\frac{\pi}{2}} \left[(x(t))^2 - (x'(t))^2\right] dt$, $x(0) = 0$ and $x\left(\frac{\pi}{2}\right)$ is free.

(b) $I(x) = \int_0^{\frac{\pi}{2}} \left[(x(t))^2 - (x'(t))^2\right] dt$, $x(0) = 1$ and $x\left(\frac{\pi}{2}\right)$ is free.

4. Determine the curves that maximize the function $I : S \to \mathbb{R}$, where $I(x) = \int_0^1 \cos(x'(t))dt$, $x \in S$ and $S = \{x \in C^1[0,1] \mid x(0) = 0\}$. What are the curves that minimize I?

2.4 Generalization

The results in this chapter can be generalized to the case when the integrand F is a function of more than one independent variable: if we wish to find extremum values of the function

$$I(x_1, \ldots, x_n) = \int_a^b F\left(x_1(t), \ldots, x_n(t), \frac{dx_1}{dt}(t), \ldots, \frac{dx_n}{dt}(t), t\right) dt,$$

where $F(\alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_n, \gamma)$ is a function with continuous partial derivatives of order ≤ 2, and $x_1, \ldots, x_n$ are continuously differentiable functions of the variable t, then following a similar analysis as before, we obtain n Euler-Lagrange equations to be satisfied by the optimal curve, that is,

$$\frac{\partial F}{\partial \alpha_k}(x_{1*}(t), \ldots, x_{n*}(t), x'_{1*}(t), \ldots, x'_{n*}(t), t) - \frac{d}{dt}\left(\frac{\partial F}{\partial \beta_k}(x_{1*}(t), \ldots, x_{n*}(t), x'_{1*}(t), \ldots, x'_{n*}(t), t)\right) = 0,$$

for $t \in [a, b]$, $k \in \{1, \ldots, n\}$. Also at any end point where x_k is free,

$$\frac{\partial F}{\partial \beta_k}\left(x_{1*}(t), \ldots, x_{n*}(t), \frac{dx_{1*}}{dt}(t), \ldots, \frac{dx_{n*}}{dt}(t), t\right) = 0.$$

Exercise. Find critical curves of the function

$$I(x_1, x_2) = \int_0^{\frac{\pi}{2}} \left[(x_1'(t))^2 + (x_2'(t))^2 + 2x_1(t)x_2(t)\right] dt$$

such that $x_1(0) = 0$, $x_1\left(\frac{\pi}{2}\right) = 1$, $x_2(0) = 0$, $x_2\left(\frac{\pi}{2}\right) = 1$.

Remark. Note that with the above result, we can also solve the problem of finding extremal curves for a function of the type

$$I(x) : \int_a^b F\left(x(t), \frac{dx}{dt}(t), \ldots, \frac{d^n x}{dt^n}(t), t\right) dt,$$

for over all (sufficiently differentiable) curves x defined on an interval $[a, b]$, taking values in $\mathbb{R}$. Indeed, we may introduce the auxiliary functions

$$x_1(t) = x(t), \quad x_2(t) = \frac{dx}{dt}(t), \quad \ldots, \quad x_n(t) = \frac{d^n x}{dt^n}(t), \quad t \in [a, b],$$

and consider the problem of finding extremal curves for the new function $\tilde{I}$ defined by

$$\tilde{I}(x_1, \ldots, x_n) = \int_a^b F(x_1(t), x_2(t), \ldots, x_n(t), t)dt.$$

Using the result mentioned in this section, we can then solve this problem. Note that we eliminated *high* order derivatives at the price of converting the *scalar* function into a *vector*-valued function. Since we can always do this, this is one of the reasons in fact for considering functions of the type (1.29) where no high order derivatives occur.

2.5 Optimisation subject to a scalar-valued constraint

In ordinary calculus, one can solve optimisation problems subject to constraints using Lagrange multipliers. We recall this algorithm by considering the following example.

Example. Suppose we want to find the shape of a triangle with a given fixed perimeter P and base a such that the area A is maximized. If the other two side lengths are b, c and $s := \frac{a+b+c}{2}$, then the problem is that of maximizing

$$A(b, c) = \sqrt{s(s-a)(s-b)(s-c)}$$

subject to

$$b + c = P - a.$$

Following the method of Lagrange multipliers, we consider the auxiliary function

$$g(b, c, \lambda) = \sqrt{s(s-a)(s-b)(s-c)} + \lambda[a + b + c - P],$$

where λ is called a *Lagrange multiplier*, and we find extremal points of g. Thus we seek (b_0, c_0, λ_0) such that

$$\frac{\partial g}{\partial b}(b_0, c_0, \lambda_0) = 0, \quad \frac{\partial g}{\partial c}(b_0, c_0, \lambda_0) = 0, \quad \frac{\partial g}{\partial \lambda}(b_0, c_0, \lambda_0) = 0.$$

So we obtain the system

$$\lambda_0 = \frac{A_0}{2(s - b_0)} = \frac{A_0}{2(s - c_0)} \quad \text{and} \quad a + b_0 + c_0 = P,$$

where $A_0 := A(a, b_0, c_0)$, and this has the unique solution

$$b_0 = c_0 = \frac{P-a}{2}.$$

Hence the area is maximized by an isosceles triangle. ◊

Analogous to this method of Lagrange multipliers in finite dimensions, we state the following result for constrained optimization in a normed space, but we do not study the proof.

Theorem 2.5.1 *Suppose that F and G are functions having continuous first and second partial derivatives. Let $L : C^1[a,b] \to \mathbb{R}$ be given by*

$$L(x) = \int_a^b G(x(t), x'(t), t)dt, \quad x \in C^1[a,b].$$

Let $S(y_a, y_b, c) = \{x \in C^1[a,b] \mid x(a) = y_a,\ x(b) = y_b,\ L(x) = c\}$, and $I : S(y_a, y_b, c) \to \mathbb{R}$ be given by

$$I(x) = \int_a^b F(x(t), x'(t), t)dt, \quad x \in S(y_a, y_b, c).$$

If $x_0 \in S(y_a, y_b, c)$ is a local extremum for I and $DL(x_0) \neq 0$, then there exists a $\lambda \in \mathbb{R}$ such that

$$\frac{\partial(F+\lambda G)}{\partial\alpha}(x_0(t), x_0'(t), t) - \frac{d}{dt}\left(\frac{\partial(F+\lambda G)}{\partial\beta}(x_0(t), x_0'(t), t)\right) = 0, \quad t \in [a,b].$$

As an illustration of the above theorem, we solve the following isoperimetric problem, this time not restricting ourselves to triangles.

Example. Among all curves of length l in the in the upper half plane passing through the points $(-r, 0)$ and $(r, 0)$, find one, which together with the interval $[-r, r]$ encloses the largest area.

We seek a function x for which the integral

$$I(x) = \int_{-r}^{r} x(t)dt$$

takes the largest value, subject to the conditions

$$x(-r) = 0 = x(r), \quad \int_{-r}^{r} \sqrt{1 + (x'(t))^2}dt = l.$$

In light of the theorem above, we seek x_0 and λ such that

$$1 - \lambda\frac{d}{dt}\left(\frac{x_0'(t)}{\sqrt{1 + (x_0'(t))^2}}\right) = 0,$$

which implies

$$\lambda\frac{x_0'}{\sqrt{1 + (x_0')^2}} = t + C,$$

which yields

$$(x_0(t) + D)^2 + (t + C)^2 = \lambda^2,$$

representing a family of circles. The values of C_1, C_2 and λ satisfy the conditions

$$x_0(-r) = 0 = x(r), \quad \int_{-r}^{r} \sqrt{1 + (x_0'(t))^2}dt = l.$$

Then it can be seen that the curve x_0 is a part of a circular arc, with center $(0, -D)$ and radius $\sqrt{D^2 + r^2}$, and the constant D is a solution of the transcendental equation

$$2 \arctan\left(\frac{r}{D}\right) = \frac{l}{\sqrt{D^2 + r^2}}.$$

We note that if $l = \pi r$, then $D = 0$, and so the curve becomes a semicircle. $\Diamond$

Exercise. Find the extremals of the function

$$I(x) = \int_0^1 [(x'(t))^2 + t^2]dt.$$

where $x \in C^1[0, 1]$ with $x(0) = 0$, $x(1) = 0$, and

$$\int_0^1 (x(t))^2 dt = 2.$$

2.6 Optimisation in function spaces versus that in $\mathbb{R}^n$

To understand the basic 'infinite-dimensionality' in the problems of optimisation in function spaces, it is interesting to see how they are related to the problems of the study of functions of n real variables. Thus, consider a function of the form

$$I(x) = \int_a^b F\left(x(t), \frac{dx}{dt}(t), t\right) dt, \quad x(a) = y_a, \quad x(b) = y_b.$$

Here each curve x is assigned a certain number. To find a related function of the sort considered in classical analysis, we may proceed as follows. Using the points

$$a = t_0, t_1, \ldots, t_n, t_{n+1} = b,$$

we divide the interval $[a, b]$ into $n + 1$ equal parts. Then we replace the curve $\{x(t),\ t \in [a, b]\}$ by the polygonal line joining the points

$$(t_0, y_a), (t_1, x(t_1)), \ldots, (t_n, x(t_n)), (t_{n+1}, y_b),$$

and we approximate the function I at x by the sum

$$I_n(y_1, \ldots, y_n) = \sum_{k=1}^{n} F\left(y_k, \frac{y_k - y_{k-1}}{h_k}, t_k\right) h_k, \tag{2.11}$$

where $y_k = x(t_k)$ and $h_k = t_k - t_{k-1}$. Each polygonal line is uniquely determined by the ordinates $y_1, \ldots, y_n$ of its vertices (recall that $y_0 = y_a$ and $y_{n+1} = y_b$ are fixed), and the sum (2.11) is therefore a function of the n variables $y_1, \ldots, y_n$. Thus as an approximation, we can regard the optimisation problem as the problem of finding the extrema of the function $I_n(y_1, \ldots, y_n)$.

In solving optimisation problems in function spaces, Euler made extensive use of this 'method of finite differences'. By replacing smooth curves by polygonal lines, he reduced the problem of finding extrema of a function to the problem of finding extrema of a function of n variables, and then he obtained exact solutions by passing to the limit as $n \to \infty$. In this sense, functions can be regarded as 'functions of infinitely many variables' (that is, the infinitely many values of $x(t)$ at different points), and the calculus of variations can be regarded as the corresponding analog of differential calculus of functions of n real variables.

Example. Consider the problem of finding the curve $x \in C^1[0,1]$ such that $x(0) = 0$ and $x(1) = 1$ that minimizes

$$I(x) = \int_0^1 \sqrt{1 + (x'(t))^2}dt.$$

Discretization yields the auxiliary finite dimensional problem of determining the points $y_1, \ldots, y_{n-1}$ that minimizes

$$I_n(y_1, \ldots, y_{n-1}) = \frac{1}{n}\sum_{k=0}^{n-1} \sqrt{1 + \left(\frac{y_{k+1} - y_k}{1/n}\right)^2},$$

where $y_0 := 0$ and $y_n := 1$.

It is easy to see that the function $x \mapsto \sqrt{1 + x^2}$ is convex (What is its second derivative?). Using Exercise 5 on page 23, we obtain that

$$\frac{\sqrt{1 + a_1^2} + \cdots + \sqrt{1 + a_n^2}}{n} \geq \sqrt{1 + \left(\frac{a_1 + \cdots + a_n}{n}\right)^2}$$

and equality holds if $a_1 = \cdots = a_n$.

With $a_k := \frac{y_{k+1} - y_k}{1/n}$, we obtain that

$$I_n(y_1, \ldots, y_{n-1}) \geq \sqrt{1 + \left(\frac{1}{n}\sum_{k=0}^{n-1} \frac{y_{k+1} - y_k}{1/n}\right)^2} = \sqrt{2},$$

and there is equality if

$$y_k = \frac{k}{n}, \quad k \in \{1, \ldots, n-1\}.$$

The corresponding points clearly lie on the straight line $x(t) = t$, $t \in [0,1]$, which we have already seen is the curve that minimizes I. $\Diamond$

Chapter 3

Control theory

In the following two chapters, we will study optimisation problems in a function space with a 'differential equation constraint'. The simplest example of such a problem is the following. Consider the map $I : C[0,T] \to \mathbb{R}$, defined as follows: if $u \in C[0,T]$, then

$$I(u) = \int_0^T F(x(t), u(t), t)dt,$$

where x is denotes the unique solution to the differential equation

$$x'(t) = f(x(t), u(t)), \quad t \in [0,T], \quad x(0) = x_0. \tag{3.1}$$

This can be roughly viewed as the optimisation problem for I over all $(x,u) \in C^1[0,T] \times C[0,T]$ satisfying the 'constraint' (3.1). Such problems arise quite naturally in optimal control theory, and we begin with a discussion of control theory.

3.1 Control theory

Control theory is application-oriented mathematics that deals with the basic principles underlying the analysis and design of (control) systems. *Systems* can be engineering systems (air conditioner, aircraft, CD player etcetera), economic systems, biological systems and so on. To *control* means that one has to influence the behaviour of the system in a desirable way: for example, in the case of an air conditioner, the aim is to control the temperature of a room and maintain it at a desired level, while in the case of an aircraft, we wish to control its altitude at each point of time so that it follows a desired trajectory.

3.2 Objects of study in control theory

The basic objects of study in control theory are *underdetermined* differential equations. This means that there is some *freeness* in the variables satisfying the differential equation. An example of an underdetermined *algebraic* equation is $x + u = 10$, where x, u are positive integers. There is freedom in choosing, say u, and once u is chosen, then x is uniquely determined. In the same manner, consider the *differential* equation

$$\frac{dx}{dt}(t) = f(x(t), u(t)), \quad x(0) = x_0, \quad t \geq 0, \tag{3.2}$$

$x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$. So if written out, equation (3.2) is the set of equations

$$\begin{aligned} \frac{dx_1}{dt}(t) &= f_1(x_1(t), \ldots, x_n(t), u_1(t), \ldots, u_m(t)), \quad x_1(0) = x_{0,1} \\ &\vdots \\ \frac{dx_n}{dt}(t) &= f_n(x_1(t), \ldots, x_n(t), u_1(t), \ldots, u_m(t)), \quad x_n(0) = x_{0,n}, \end{aligned}$$

where $f_1, \ldots, f_n$ denote the components of f. In (3.2), u is the free variable, called the *input*, which is assumed to be continuous.

Under some 'smoothness' conditions on the function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, there exists a unique solution to the differential equation (3.2) for every initial condition $x_0 \in \mathbb{R}^n$ and every continuous input u:

Theorem 3.2.1 *Suppose that f is continuous in both variables. If there exist $K > 0$, $r > 0$ such that*

$$\|f(x_2, u(t)) - f(x_1, u(t))\| \leq K\|x_2 - x_1\| \tag{3.3}$$

for all $x_1, x_2 \in \overline{B(x_0, r)} = \{x \in \mathbb{R}^n \mid \|x - x_0\| \leq r\}$ and for all $t > 0$, then (3.2) has a unique solution x in the interval $[0, T]$, for some $T > 0$. Furthermore, this solution depends continuously on x_0 for fixed t and u.

Remarks.

1. Continuous dependence on the initial condition is very important, since some inaccuracy is always present in practical situations. We need to know that if the initial conditions are slightly changed, then the solution of the differential equation will change only slightly. Otherwise, small inaccuracies could yield very different solutions.
2. x is called the *state* and (3.2) is called the *state equation*.
3. Condition (3.3) is called the *Lipschitz condition*.

The above theorem guarantees that a solution exists and that it is unique, but it does not give any insight into the size of the time interval on which the solutions exist. The following theorem sheds some light on this.

Theorem 3.2.2 *Let $r > 0$ and define $B_r = \{u \in C[0, T]^m \mid \|u(t)\| \leq r \text{ for all } t\}$. Suppose that f is continuously differentiable in both variables. For every $x_0 \in \mathbb{R}^n$, there exists a unique $t_*(x_0) \in (0, +\infty]$ such that for every $u \in B_r$, (3.2) has a unique solution $x(\cdot)$ in $[0, t_*(x_0))$.*

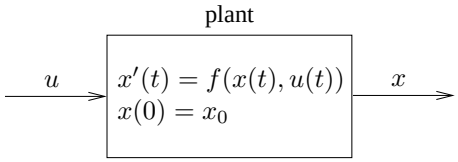


Figure 3.1: A control system.

For our purposes, a *control system* is an equation of the type (3.2), with input u and state x. Once the input u and the initial state $x(0) = x_0$ are specified, the state x is determined. So one can think of a control system as a box, which given the input u and initial state $x(0) = x_0$, manufactures the state according to the law (3.2); see Figure 3.1.

Example. Suppose the population $x(t)$ at time t of fish in a lake evolves according to the differential equation:

$$x'(t) = f(x(t)),$$

where f is some complicated function which is known to model the situation reasonable accurately. A typical example is the Verhulst model, where

$$f(x) = rx\left(1 - \frac{x}{M}\right).$$

(This model makes sense, since first of all the rate of increase in the population should increase with more numbers of fish–the more there are fis, the more they reproduce, and larger is the poulation. However, if there are too many fish, there is competition for the limited food resourse, and then the population starts declining, which is captured by the term $1 - \frac{x}{M}$.)

Now suppose that we harvest the fish at a harvesting rate h. Then the population evolution is described by

$$x'(t) = f(x(t)) - h(t).$$

But the harvesting rate depends on the harvesting effort u:

$$h(t) = x(t)u(t).$$

(The harvesting effort can be thought in terms of the amount of time used for fishing, or the number of fishing nets used, and so on. Then the above equation makes sense, as the harvesting rate is clearly proportional to the number of fish–the more the fish in the lake, the better the catch.)

Hence we arrive at the underdetermined differential equation

$$x'(t) = f(x(t)) - x(t)u(t).$$

This equation is underdetermined, since the u can be decided by the fisherman. This is the input, and once this has been chosen, then the population evolution is determined by the above equation, given some initial population level x_0 of the fish. ◊

If the function f is linear, that is, if $f(\xi, \upsilon) = A\xi + B\upsilon$ for some $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$, then the control system is said to be *linear*. We will study this important class of systems in the rest of this chapter. But we begin with a discussion of the exponential of a matrix, since the solution to a linear control system $x'(t) = Ax(t) + Bu(t)$ can be described in terms of the exponential of A.

3.3 The exponential of a matrix

In this section we introduce the exponential of a matrix, which is useful for obtaining explicit solutions to the linear control system (3.10). We begin with a few preliminaries concerning vector-valued functions.

With a slight abuse of notation, a *vector-valued function* $x(t)$ is a vector whose entries are functions of t. Similarly, a *matrix-valued function* $A(t)$ is a matrix whose entries are functions:

$$\begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{bmatrix}, \quad A(t) = \begin{bmatrix} a_{11}(t) & \dots & a_{1n}(t) \\ \vdots & & \vdots \\ a_{m1}(t) & \dots & a_{mn}(t) \end{bmatrix}.$$

The calculus operations of taking limits, differentiating, and so on are extended to vector-valued and matrix-valued functions by performing the operations on each entry separately. Thus by definition,

$$\lim_{t \to t_0} x(t) = \begin{bmatrix} \lim_{t \to t_0} x_1(t) \\ \vdots \\ \lim_{t \to t_0} x_n(t) \end{bmatrix}.$$

So this limit exists iff $\lim_{t \to t_0} x_i(t)$ exists for all $i \in \{1, \dots, n\}$. Similarly, the derivative of a vector-valued or matrix-valued function is the function obtained by differentiating each entry separately:

$$\frac{dx}{dt}(t) = \begin{bmatrix} x_1'(t) \\ \vdots \\ x_n'(t) \end{bmatrix}, \quad \frac{dA}{dt}(t) = \begin{bmatrix} a_{11}'(t) & \dots & a_{1n}'(t) \\ \vdots & & \vdots \\ a_{m1}'(t) & \dots & a_{mn}'(t) \end{bmatrix},$$

where $x_i'(t)$ is the derivative of $x_i(t)$, and so on. So $\frac{dx}{dt}$ is defined iff each of the functions $x_i(t)$ is differentiable. The derivative can also be described in vector notation, as

$$\frac{dx}{dt}(t) = \lim_{h \to 0} \frac{x(t+h) - x(t)}{h}. \tag{3.4}$$

Here $x(t+h) - x(t)$ is computed by vector addition and the h in the denominator stands for scalar multiplication by h^{-1}. The limit is obtained by evaluating the limit of each entry separately, as above. So the entries of (3.4) are the derivatives $x_i(t)$. The same is true for matrix-valued functions.

A system of homogeneous, first-order, linear constant-coefficient differential equations is a matrix equation of the form

$$\frac{dx}{dt}(t) = Ax(t), \tag{3.5}$$

where A is a $n \times n$ real matrix and $x(t)$ is an n dimensional vector-valued function. Writing out such a system, we obtain a system of n differential equations, of the form

$$\begin{aligned} \frac{dx_1}{dt}(t) &= a_{11}x_1(t) + \cdots + a_{1n}x_n(t) \\ &\vdots \\ \frac{dx_n}{dt}(t) &= a_{n1}x_1(t) + \cdots + a_{nn}x_n(t). \end{aligned}$$

The $x_i(t)$ are unknown functions, and the a_{ij} are scalars. For example, if we substitute the matrix

$$\begin{bmatrix} 3 & -2 \\ 1 & 4 \end{bmatrix}$$

for A, (3.5) becomes a system of two equations in two unknowns:

$$\begin{aligned}\frac{dx_1}{dt}(t) &= 3x_1(t) - 2x_2(t)\\ \frac{dx_2}{dt}(t) &= x_1(t) + 4x_2(t).\end{aligned}$$

Now consider the case when the matrix A is simply a scalar. We learn in calculus that the solutions to the first-order scalar linear differential equation

$$\frac{dx}{dt}(t) = ax(t)$$

are $x(t) = ce^{ta}$, c being an arbitrary constant. Indeed, ce^{ta} obviously solves this equation. To show that every solution has this form, let $x(t)$ be an arbitrary differentiable function which is a solution. We differentiate $e^{-ta}x(t)$ using the product rule:

$$\frac{d}{dt}(e^{-ta}x(t)) = -ae^{-ta}x(t) + e^{-ta}ax(t) = 0.$$

Thus $e^{-ta}x(t)$ is a constant, say c, and $x(t) = ce^{ta}$. Now suppose that analogous to

$$e^a = 1 + a + \frac{a^2}{2!} + \frac{a^3}{3!} + \ldots, \quad a \in \mathbb{R},$$

we define

$$e^A = I + A + \frac{1}{2!}A^2 + \frac{1}{3!}A^3 + \ldots, \quad A \in \mathbb{R}^{n\times n}. \tag{3.6}$$

Later in this section, we study this matrix exponential, and use the matrix-valued function

$$e^{tA} = I + tA + \frac{t^2}{2!}A^2 + \frac{t^3}{3!}A^2 + \ldots$$

(where t is a variable scalar) to solve (3.5). We begin by stating the following result, which shows that the series in (3.6) converges for any given square matrix A.

Theorem 3.3.1 *The series (3.6) converges for any given square matrix A.*

We have collected the proofs together at the end of this section in order to not break up the discussion.

Since matrix multiplication is relatively complicated, it isn't easy to write down the matrix entries of e^A directly. In particular, the entries of e^A are usually *not* obtained by exponentiating the entries of A. However, one case in which the exponential is easily computed, is when A is a diagonal matrix, say with diagonal entries λ_i. Inspection of the series shows that e^A is also diagonal in this case and that its diagonal entries are e^{λ_i}.

The exponential of a matrix A can also be determined when A is *diagonalizable* , that is, whenever we know a matrix P such that $P^{-1}AP$ is a diagonal matrix D. Then $A = PDP^{-1}$, and

using $(PDP^{-1})^k = PD^kP^{-1}$, we obtain

$$\begin{aligned} e^A &= I + A + \frac{1}{2!}A^2 + \frac{1}{3!}A^3 + \dots \\ &= I + PDP^{-1} + \frac{1}{2!}^2 PD^2P^{-1} + \frac{1}{3!}PD^3P^{-1} + \dots \\ &= PIP^{-1} + PDP^{-1} + \frac{1}{2!}^2 PD^2P^{-1} + \frac{1}{3!}PD^3P^{-1} + \dots \\ &= P\left(I + D + \frac{1}{2!}D^2 + \frac{1}{3!}D^3 + \dots\right)P^{-1} \\ &= Pe^DP^{-1} = P\begin{bmatrix} e^{\lambda_1} & & \\ 0 & \ddots & \\ & & e^{\lambda_n} \end{bmatrix}P^{-1}, \end{aligned}$$

where $\lambda_1, \dots, \lambda_n$ denote the eigenvalues of A.

Exercise. $(**)$ The set of diagonalizable $n \times n$ complex matrices is *dense* in the set of all $n \times n$ complex matrices, that is, given any $A \in \mathbb{C}^{n\times n}$, there exists a $B \in \mathbb{C}^{n\times n}$ arbitrarily close to A (meaning that $|b_{ij} - a_{ij}|$ can be made arbitrarily small for all $i, j \in \{1, \dots, n\}$) such that B has n distinct eigenvalues.

HINT: Use the fact that every complex $n \times n$ matrix A can be 'upper-triangularized': that is, there exists an invertible complex matrix P such that PAP^{-1} is upper triangular. Clearly the diagonal entries of this new upper triangular matrix are the eigenvalues of A.

In order to use the matrix exponential to solve systems of differential equations, we need to extend some of the properties of the ordinary exponential to it. The most fundamental property is $e^{a+b} = e^ae^b$. This property can be expressed as a formal identity between the two infinite series which are obtained by expanding

$$\begin{aligned} &e^{a+b} = 1 + \tfrac{(a+b)}{1!} + \tfrac{(a+b)^2}{2!} + \dots \quad \text{and} \\ &e^ae^b = \left(1 + \tfrac{a}{1!} + \tfrac{a^2}{2!} + \dots\right)\left(1 + \tfrac{b}{1!} + \tfrac{b^2}{2!} + \dots\right). \end{aligned} \tag{3.7}$$

We cannot substitute matrices into this identity because the commutative law is needed to obtain equality of the two series. For instance, the quadratic terms of (3.7), computed without the commutative law, are $\frac{1}{2}(a^2 + ab + ba + b^2)$ and $\frac{1}{2}a^2 + ab + \frac{1}{2}b^2$. They are not equal unless $ab = ba$. So there is no reason to expect e^{A+B} to equal e^Ae^B in general. However, if two matrices A and B happen to commute, the formal identity can be applied.

Theorem 3.3.2 *If $A, B \in \mathbb{R}^{n\times n}$ commute (that is $AB = BA$), then $e^{A+B} = e^Ae^B$.*

The proof is at the end of this section. Note that the above implies that e^A is always invertible and in fact its inverse is e^{-A}: Indeed $I = e^{A-A} = e^Ae^{-A}$.

Exercises.

1. Give an example of 2×2 matrices A and B such that $e^{A+B} \neq e^Ae^B$.
2. Compute e^A, where $A = \begin{bmatrix} 2 & 3 \\ 0 & 2 \end{bmatrix}$.

   HINT: $A = 2I + \begin{bmatrix} 0 & 3 \\ 0 & 0 \end{bmatrix}$.

We now come to the main result relating the matrix exponential to differential equations. Given an $n \times n$ matrix, we consider the exponential e^{tA}, t being a variable scalar, as a matrix-valued function:

$$e^{tA} = I + tA + \frac{t^2}{2!}A^2 + \frac{t^3}{3!}A^3 + \dots.$$

Theorem 3.3.3 *e^{tA} is a differentiable matrix-valued function of t, and its derivative is Ae^{tA}.*

The proof is at the end of the section.

Theorem 3.3.4 (Product rule.) *Let $A(t)$ and $B(t)$ be differentiable matrix-valued functions of t, of suitable sizes so that their product is defined. Then the matrix product $A(t)B(t)$ is differentiable, and its derivative is*

$$\frac{d}{dt}(A(t)B(t)) = \frac{dA(t)}{dt}B(t) + A(t)\frac{dB(t)}{dt}.$$

The proof is left as an exercise.

Theorem 3.3.5 *The first-order linear differential equation*

$$\frac{dx}{dt}(t) = Ax(t), \quad t \geq a, \quad x(0) = x_0 \tag{3.8}$$

has the unique solution $x(t) = e^{tA}x_0$.

Proof We have

$$\frac{d}{dt}(e^{tA}x_0) = Ae^{tA}x_0,$$

and so $t \mapsto e^{tA}x_0$ solves $\frac{dx}{dt}(t) = Ax(t)$. Furthermore, $x(0) = e^{0A}x_0 = Ix_0 = x_0$.

Finally we show that the solution is unique. Let x be a solution to (3.8). Using the product rule, we differentiate the matrix product $e^{-tA}x(t)$:

$$\frac{d}{dt}(e^{-tA}x(t)) = -Ae^{-tA}x(t) + e^{-tA}Ax(t).$$

From the definition of the exponential, it can be seen that A and e^{tA} commute, and so the derivative of $e^{tA}x(t)$ is zero. Therefore, $e^{tA}x(t)$ is a constant column vector, say C, and $x(t) = e^{tA}C$. As $x(0) = x_0$, we obtain that $x_0 = e^{0A}C$, that is, $C = x_0$. Consequently, $x(t) = e^{tA}x_0$. ∎

Thus the matrix exponential enables us to solve the differential equation (3.8). Since direct computation of the exponential can be quite difficult, the above theorem may not be easy to apply in a concrete situation. But if A is a diagonalizable matrix, then the exponential can be computed: $e^A = Pe^DP^{-1}$. To compute the exponential explicitly in all cases requires putting the matrix into Jordan form.

We now go back to prove Theorems 3.3.1, 3.3.2, and 3.3.3.

For want of a more compact notation, we will denote the i,j-entry of a matrix A by A_{ij} here. So $(AB)_{ij}$ will stand for the entry of the matrix product matrix AB, and $(A^k)_{ij}$ for the entry of A^k. With this notation, the i,j-entry of e^A is the sum of the series

$$(e^A)_{ij} = I_{ij} + A_{ij} + \frac{1}{2!}(A^2)_{ij} + \frac{1}{3!}(A^3)_{ij} + \dots.$$

In order to prove that the series for the exponential converges, we need to show that the entries of the powers A^k of a given matrix do not grow too fast, so that the absolute values of the i,j-entries form a bounded (and hence convergent) series. Consider the following norm on $\mathbb{R}^{n\times n}$:

$$\|A\| = \max\{|A_{ij}| \mid 1 \leq i,j \leq n\}.$$

Thus $|A_{ij}| \leq \|A\|$ for all i,j. This is one of several possible norms on $\mathbb{R}^{n\times n}$, and it has the following property.

Lemma 3.3.6 *If $A, B \in \mathbb{R}^{n\times n}$, then $\|AB\| \leq n\|A\|\|B\|$, and for all $k \in \mathbb{N}$, $\|A^k\| \leq n^{k-1}\|A\|^k$.*

Proof We estimate the size of the i,j-entry of AB:

$$|(AB)_{ij}| = \left|\sum_{k=1}^{n} A_{ik}B_{kj}\right| \leq \sum_{k=1}^{n} |A_{ik}||B_{kj}| \leq n\|A\|\|B\|.$$

Thus $\|AB\| \leq n\|A\|\|B\|$. The second inequality follows from the first inequality by induction. ■

Proof (of Theorem 3.3.1:) To prove that the matrix exponential converges, we show that the series

$$I_{ij} + A_{ij} + \frac{1}{2!}(A^2)_{ij} + \frac{1}{3!}(A^3)_{ij} + \dots$$

is absolutely convergent, and hence convergent. Let $a = n\|A\|$. Then

$$\begin{aligned} |I_{ij}| + |A_{ij}| + \frac{1}{2!}|(A^2)_{ij}| + \frac{1}{3!}|(A^3)_{ij}| + \dots &\leq 1 + \|A\| + \frac{1}{2!}n\|A\|^2 + \frac{1}{3!}n^2\|A\|^3 + \dots \\ &= 1 + \frac{1}{n}\left(a + \frac{1}{2!}a^2 + \frac{1}{3!}a^3 + \dots\right) = 1 + \frac{e^a - 1}{n}. \end{aligned}$$

■

Proof (of Theorem 3.3.2:) The terms of degree k in the expansions of (3.7) are

$$\frac{1}{k!}(A+B)^k = \frac{1}{k!}\sum_{r+s=k}\binom{k}{r}A^rB^s \quad \text{and} \quad \sum_{r+s=k}\frac{1}{r!}A^r\frac{1}{s!}B^s.$$

These terms are equal since for all k, and all r,s such that $r+s=k$,

$$\frac{1}{k!}\binom{k}{r} = \frac{1}{r!s!}.$$

Define

$$S_n(A) = 1 + \frac{1}{1!}A + \frac{1}{2!}A^2 + \dots + \frac{1}{n!}A^n.$$

Then

$$\begin{aligned} S_n(A)S_n(B) &= \left(1 + \frac{1}{1!}A + \frac{1}{2!}A^2 + \dots + \frac{1}{n!}A^n\right)\left(1 + \frac{1}{1!}B + \frac{1}{2!}B^2 + \dots + \frac{1}{n!}B^n\right) \\ &= \sum_{r,s=0}^{n}\frac{1}{r!}A^r\frac{1}{s!}B^s, \end{aligned}$$

while

$$\begin{aligned} S_n(A+B) &= 1 + \frac{1}{1!}(A+B) + \frac{1}{2!}(A+B)^2 + \dots + \frac{1}{n!}(A+B)^n \\ &= \sum_{k=0}^{n}\sum_{r+s=k}\frac{1}{k!}\binom{k}{r}A^rB^s = \sum_{k=0}^{n}\sum_{r+s=k}\frac{1}{r!}A^r\frac{1}{s!}B^s. \end{aligned}$$

Comparing terms, we find that the expansion of the partial sum $S_n(A+B)$ consists of the terms in $S_n(A)S_n(B)$ such that $r+s \leq n$. We must show that the sum of the remaining terms tends to zero as k tends to ∞.

Lemma 3.3.7 *The series*

$$\sum_k \sum_{r+s=k} \left| \left(\frac{1}{r!} A^r \frac{1}{s!} B^s \right)_{ij} \right|$$

converges for all i, j.

Proof Let $a = n\|A\|$ and $b = n\|B\|$. We estimate the terms in the sum:

$$|(A^r B^s)_{ij}| \leq n(n^{r-1}\|A\|^r)(n^{s-1}\|B\|^s) \leq a^r b^s.$$

Therefore

$$\sum_k \sum_{r+s=k} \left| \left(\frac{1}{r!} A^r \frac{1}{s!} B^s \right)_{ij} \right| \leq \sum_k \sum_{r+s=k} \frac{a^r}{r!} \frac{b^s}{s!} = e^{a+b}.$$

The theorem follows from this lemma because, on the one hand, the i,j-entry of $(S_k(A)S_k(B) - S_k(A+B))_{ij}$ is bounded by

$$\sum_{r+s>k} \left| \left(\frac{1}{r!} A^r \frac{1}{s!} B^s \right)_{ij} \right|.$$

According to the lemma, this sum tends to 0 as k tends to ∞. And on the other hand, $S_k(A)S_k(B) - S_k(A+B)$ tends to $e^A e^B - e^{A+B}$. ■

This completes the proof of Theorem 3.3.2. ■

Proof (of Theorem 3.3.3:) By definition,

$$\frac{d}{dt} e^{tA} = \lim_{h \to 0} \frac{1}{h} (e^{(t+h)A} - e^{tA}).$$

Since the matrices tA and hA commute, we have

$$\frac{1}{h}(e^{(t+h)A} - e^{tA}) = \left(\frac{1}{h}(e^{hA} - I) \right) e^{tA}.$$

So our theorem follows from this lemma:

Lemma 3.3.8 $\lim_{h \to 0} \frac{1}{h}(e^{hA} - I) = A.$

Proof The series expansion for the exponential shows that

$$\frac{1}{h}(e^{hA} - I) - A = \frac{h}{2!} A^2 + \frac{h^2}{3!} A^3 + \dots. \tag{3.9}$$

We estimate this series. Let $a = |h|n\|A\|$. Then

$$\begin{aligned} \left| \left(\frac{h}{2!} A^2 + \frac{h^2}{3!} A^3 + \dots \right)_{ij} \right| &\leq \left| \frac{h}{2!} (A^2)_{ij} \right| + \left| \frac{h^2}{3!} (A^3)_{ij} \right| + \dots \\ &\leq \frac{1}{2!} |h| n \|A\|^2 + \frac{1}{3!} |h|^2 n^2 \|A\|^3 + \dots \\ &= \|A\| \left(\frac{1}{2!} a + \frac{1}{3!} a^2 + \dots \right) \\ &= \frac{\|A\|}{a} (e^a - 1 - a) = \|A\| \left(\frac{e^a - 1}{a} - 1 \right). \end{aligned}$$

Note that $a \to 0$ as $h \to 0$. Since the derivative of e^x is e^x,

$$\lim_{a\to 0} \frac{e^a - 1}{a} = \left.\frac{d}{dx} e^x\right|_{x=0} = e^0 = 1.$$

So (3.9) tends to 0 with h. ■

This completes the proof of Theorem 3.3.3. ■

3.4 Solutions to the linear control system

Using the results from the previous section, we give a formula for the solution of a linear control system.

Theorem 3.4.1 *Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times 1}$. If u is a continuous function, then the differential equation*

$$\frac{dx}{dt}(t) = Ax(t) + Bu(t), \quad x(0) = x_0, \quad t \geq 0 \tag{3.10}$$

has the unique solution $x(\cdot)$ in $[0, +\infty)$ given by

$$x(t) = e^{tA}x_0 + \int_0^t e^{(t-\tau)A}Bu(\tau)d\tau. \tag{3.11}$$

Proof We have

$$\begin{aligned}
\frac{d}{dt}\left(e^{tA}x_0 + \int_0^t e^{(t-\tau)A}Bu(\tau)d\tau\right) &= \frac{d}{dt}\left(e^{tA}x_0 + e^{tA}\int_0^t e^{-\tau A}Bu(\tau)d\tau\right) \\
&= Ae^{tA}x_0 + Ae^{tA}\int_0^t e^{-\tau A}Bu(\tau)d\tau + e^{tA}e^{-tA}Bu(t) \\
&= A\left(e^{tA}x_0 + e^{tA}\int_0^t e^{-\tau A}Bu(\tau)d\tau\right) + e^{tA-tA}Bu(t) \\
&= A\left(e^{tA}x_0 + e^{tA}\int_0^t e^{-\tau A}Bu(\tau)d\tau\right) + Bu(t),
\end{aligned}$$

and so it follows that $x(\cdot)$ given by (3.11) satisfies $x'(t) = Ax(t) + Bu(t)$. Furthermore,

$$e^{0A}x_0 + \int_0^0 e^{(0-\tau)A}Bu(\tau)d\tau = Ix_0 + 0 = x_0.$$

Finally we show uniqueness. If x_1, x_2 are both solutions to (3.10), then it follows that $x := x_1 - x_2$ satisfies $x'(t) = Ax(t)$, $x(0) = 0$, and so from Theorem 3.3.5 it follows that $x(t) = 0$ for all $t \geq 0$, that is $x_1 = x_2$. ■

We end this section with the following result concerning *nonlinear* differential equations, which we will use in the next chapter. This result says, roughly speaking, that if we perturb the input function u a little, then the new state differs from the old state by a function which is the solution of a *linear* control system, with the input to this linear system being the perturbation (in the original input). In other words, 'around' some solution (x_0, u_0) of the original nonlinear system, the system behaves as if it is a linear system. This result in Theorem 3.4.2 shows the importance of linear systems: not only are they simple (in the sense that we have an explicit description of the solution in terms of the input and the initial condition), but in fact nonlinear systems behave like linear systems 'locally'.

Theorem 3.4.2 *Consider the equation*

$$\frac{dx}{dt}(t) = f(x(t), u(t)), \quad t \in [0, T], \quad x(0) = x_0. \tag{3.12}$$

Let $\dfrac{\partial f}{\partial \upsilon}$ *be continuous, and let* $v \in C[0, T]$. *For* $\epsilon \in [0, \eta]$, *let* x_ϵ *be the solution to (3.12) corresponding to the input*

$$u_\epsilon(t) = u(t) + \epsilon v(t), \quad t \in [0, T],$$

with the same initial condition $x_\epsilon(0) = x_0$. *Then*

$$x_\epsilon(t) = x(t) + \epsilon y(t) + o(t, \epsilon), \tag{3.13}$$

where y *is the solution to*

$$y'(t) = \frac{\partial f}{\partial \xi}(x(t), u(t))y(t) + \frac{\partial f}{\partial \upsilon}(x(t), u(t))v(t), \quad t \in [0, T], \quad y(0) = 0.$$

In the following exercises, we learn to solve a certain type of nonlinear differential equation, called a *Riccati equation*, which will play an important role in the sequel. We solve the Riccati equation by making a transformation that results in a linear control system.

Exercises.

1. Suppose that $p \in C^1[0, T]$ is such that for all $t \in [0, T]$, $p(t) + \alpha \neq 0$, and it satisfies the scalar Riccati equation $p'(t) = \gamma(p(t) + \alpha)(p(t) + \beta)$. Prove that q given by

   $$q(t) := \frac{1}{p(t) + \alpha}, \quad t \in [0, T],$$

   satisfies $q'(t) = \gamma(\alpha - \beta)q(t) - \gamma$, $t \in [0, T]$.

2. Find $p \in C^1[0, 1]$ such that $p'(t) = (p(t))^2 - 1$, $t \in [0, 1]$, $p(1) = 0$.

3.5 Controllability of linear control systems

A characteristic of underdetermined equations is that one can choose the free variable in a way that some desirable effect is produced on the other dependent variable.

For example, if with our algebraic equation $x + u = 10$ we wish to make $x < 5$, then we can achieve this by choosing the free variable u to be strictly larger than 5.

Control theory is all about doing similar things with differential equations of the type (3.2). The state variables x comprise the 'to-be-controlled' variables, which depend on the free variables u, the inputs. For example, in the case of an aircraft, the speed, altitude and so on are the to-be-controlled variables, while the angle of the wing flaps, the speed of the propeller and so on, which the pilot can specify, are the inputs. So one of the basic questions in control theory is then the following:

How do we choose the control inputs to achieve regulation of the state variables?

For instance, one may wish to drive the state to zero or some other desired value of the state at some time instant T. This brings us naturally to the notion of controllability which, roughly speaking, means that any state can be driven to any other state using an appropriate control.

For the sake of simplicity, we restrict ourselves to linear systems: $x'(t) = Ax(t) + Bu(t)$, $t \geq 0$, where $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times 1}$. We first give the definition of controllability for such a linear control system.

Example. Suppose a lake contains two species of fish, which we simply call 'big fish' and 'small fish', which form a predator-prey pair. Suppose that the evolution of their populations x_b and x_s are reasonably accurately modelled by

$$\begin{bmatrix} x_b'(t) \\ x_s'(t) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_b(t) \\ x_s(t) \end{bmatrix}.$$

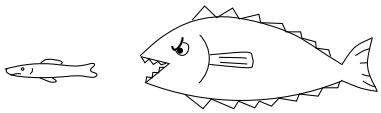


Figure 3.2: Big fish and small fish.

Now suppose that one is harvesting these fish at harvesting rates h_b and h_s (which are inputs, since they can be decided by the fisherman). The model describing the evolution of the populations then becomes:

$$\begin{bmatrix} x_b'(t) \\ x_s'(t) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_b(t) \\ x_s(t) \end{bmatrix} - \begin{bmatrix} h_b(t) \\ h_s(t) \end{bmatrix}.$$

The goal is to harvest the species of fish over some time period $[0, T]$ in such a manner starting from the initial population levels

$$\begin{bmatrix} x_b(0) \\ x_s(0) \end{bmatrix} = \begin{bmatrix} x_{b,i} \\ x_{s,i} \end{bmatrix},$$

we are left with the desired population levels

$$\begin{bmatrix} x_b(T) \\ x_s(T) \end{bmatrix} = \begin{bmatrix} x_{b,f} \\ x_{s,f} \end{bmatrix}.$$

For xemaple, if one of the species of fish is nearing extinction, it might be important to maintain some crtical levels of the populations of the predator versus the prey. Thus we see that controllability problems arise quite naturally from applications. ◊

Definition. The system

$$\frac{dx}{dt}(t) = Ax(t) + Bu(t), \quad t \in [0, T] \tag{3.14}$$

is said to be *controllable at time* T if for every pair of vectors x_0, x_1 in $\mathbb{R}^n$, there exists a control $u \in C[0, T]$ such that the solution x of (3.14) with $x(0) = x_0$ satisfies $x(T) = x_1$.

Examples.

1. *(A controllable system)* Consider the system $x'(t) = u(t)$, $t \in [0, T]$, so that $A = 0$, $B = 1$. Given $x_0, x_1 \in \mathbb{R}$, define $u \in C[0, T]$ to be the constant function

$$u(t) = \frac{x_1 - x_0}{T}, \quad t \in [0, T].$$

By the fundamental theorem of calculus,

$$x(T) = x(0) + \int_0^T x'(\tau)d\tau = x_0 + \int_0^T u(\tau)d\tau = x_0 + \frac{x_1 - x_0}{T}(T-0) = x_1.$$

2. *(An uncontrollable system)* Consider the system

$$x_1'(t) = x_1(t) + u(t), \tag{3.15}$$
$$x_2'(t) = x_2(t), \tag{3.16}$$

so that

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

The equation (3.16) implies that $x_2(t) = e^t x_2(0)$, and so if $x_2(0) > 0$, then $x_2(t) > 0$ for all $t \geq 0$. So a final state with the x_2-component negative is never reachable by any control. $\Diamond$

We would like to characterize the property of controllability in terms of the matrices A and B. For this purpose, we introduce the notion of reachable space at time T:

Definition. The *reachable space* of (3.14) at time T, denoted by $\mathscr{R}_T$, is defined as the set of all $x \in \mathbb{R}^n$ for which there exists a control $u \in C[0,T]$ such that

$$x = \int_0^T e^{(T-\tau)A} B u(\tau) d\tau. \tag{3.17}$$

Note that the above simply says that if we run the differential equation (3.14) with the input u, and with initial condition $x(0) = 0$, then x is the set of all points in the state-space that are 'reachable' at time T starting from 0 by means of some input $u \in C[0,T]$.

We now prove that $\mathscr{R}_T$ is a subspace of $\mathbb{R}^n$.

Lemma 3.5.1 *$\mathscr{R}_T$ is a subspace of $\mathbb{R}^n$.*

Proof We verify that $\mathscr{R}_T$ is nonempty, and closed under addition and scalar multiplication.

S1 If we take $u = 0$, then

$$\int_0^T e^{(T-\tau)A} B u(\tau) d\tau = 0,$$

and so $0 \in \mathscr{R}_T$.

S2 If $x_1, x_2 \in \mathscr{R}_T$, then there exist u_1, u_2 in $C[0,T]$ such that

$$x_1 = \int_0^T e^{(T-\tau)A} B u_1(\tau) d\tau \quad \text{and} \quad x_2 = \int_0^T e^{(T-\tau)A} B u_2(\tau) d\tau.$$

Thus $u := u_1 + u_2 \in C[0,T]$ and

$$\int_0^T e^{(T-\tau)A} B u(\tau) d\tau = \int_0^T e^{(T-\tau)A} B u_1(\tau) d\tau + x_2 = \int_0^T e^{(T-\tau)A} B u_2(\tau) d\tau = x_1 + x_2.$$

Consequently $x_1 + x_2 \in \mathscr{R}_T$.

S3 If $x \in \mathscr{R}_T$, then there exists a $u \in C[0,T]$ such that

$$x = \int_0^T e^{(T-\tau)A} B u(\tau) d\tau.$$

If $\alpha \in \mathbb{R}$, then $\alpha u \in C[0,T]$ and

$$\int_0^T e^{(T-\tau)A} B(\alpha u)(\tau) d\tau = \alpha \int_0^T e^{(T-\tau)A} B u(\tau) d\tau = \alpha x.$$

Consequently $\alpha x \in \mathscr{R}_T$.

Thus $\mathscr{R}_T$ is a subspace of $\mathbb{R}^n$. ■

We now prove Theorem 3.5.3, which will yield Corollary 3.5.4 below on the characterization of the property of controllability. In order to prove Theorem 3.5.3, we will use the Cayley-Hamilton theorem, and for the sake of completeness, we have included a sketch of proof of this result here.

Theorem 3.5.2 (Cayley-Hamilton) *If $A \in \mathbb{C}^{n\times n}$ and $p(t) = t^n + c_{n-1}t^{n-1} + \cdots + c_1 t + c_0$ is its characteristic polynomial, then $p(A) = A^n + c_{n-1}A^{n-1} + \cdots + c_1 A + c_0 I = 0$.*

Proof *(Sketch)* This is easy to see if A is diagonal, since

$$p\left(\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}\right) = \begin{bmatrix} p(\lambda_1) & & \\ & \ddots & \\ & & p(\lambda_n) \end{bmatrix} = 0.$$

It is also easy to see if A is diagonalizable, since if $A = PDP^{-1}$, then

$$p(A) = p(PDP^{-1}) = Pp(D)P^{-1} = P0P^{-1} = 0.$$

As $\det : \mathbb{C}^{n\times n} \to \mathbb{C}$ is a continuous function, it follows that the coefficients of the characteristic polynomial are continuous functions of the matrix entries. Using the fact that the set of diagonalizable matrices is dense in $\mathbb{C}^{n\times n}$, we see that the result extends to all complex matrices by continuity. ■

Theorem 3.5.3 $\mathscr{R}_T = \mathbb{R}^n$ *iff* $\operatorname{rank}\left[\begin{array}{c|c|c|c} B & AB & \ldots & A^{n-1}B \end{array}\right] = n$.

Proof If: If $\mathscr{R}_T \neq \mathbb{R}^n$, then there exists a $x_0 \neq 0$ in $\mathbb{R}^n$ such that for all $x \in \mathscr{R}_T$, $x_0^\top x = 0$. Consequently,

$$x_0^\top \int_0^T e^{(T-\tau)A} B u(\tau) d\tau = 0 \quad \forall u \in C[0,T].$$

In particular, u_0 defined by $u_0(t) = x_0^\top e^{(T-\tau)A} B$, $t \in [0,T]$, belongs to $C[0,T]$, and so

$$\int_0^T \left(x_0^\top e^{(T-\tau)A} B\right)^2 d\tau = 0,$$

and so

$$x_0^\top e^{(T-\tau)A} B = 0, \quad t \in [0,T]. \tag{3.18}$$

With $t = T$, we obtain $x_0^\top B = 0$. Differentiating (3.18), we obtain $x_0^\top e^{(T-t)A}AB = 0$, $t \in [0,T]$, and so with $t = T$, we have $x_0^\top AB = 0$. Proceeding in this manner (that is, successively differentiating (3.18) and setting $t = T$), we see that $x_0^\top A^k B = 0$ for all $k \in \mathbb{N}$, and so in particular,

$$x_0^\top \left[\begin{array}{c|c|c|c} B & AB & \dots & A^{n-1}B \end{array}\right] = 0.$$

As $x_0 \neq 0$, we obtain $\operatorname{rank}\left[\begin{array}{c|c|c|c} B & AB & \dots & A^{n-1}B \end{array}\right] < n$.

Only if: Let $\mathscr{C} := \operatorname{rank}\left[\begin{array}{c|c|c|c} B & AB & \dots & A^{n-1}B \end{array}\right] < n$. Then there exists a nonzero $x_0 \in \mathbb{R}^n$ such that

$$x_0^\top \left[\begin{array}{c|c|c|c} B & AB & \dots & A^{n-1}B \end{array}\right] = 0. \tag{3.19}$$

By the Cayley-Hamilton theorem, it follows that

$$x_0^\top A^n B = x_0^\top \left[\alpha_0 I + \alpha_1 A + \cdots + \alpha_n A^{n-1}\right] B = 0.$$

By induction,

$$x_0^\top A^k B = 0 \quad \forall k \geq n. \tag{3.20}$$

From (3.19) and (3.20), we obtain $x_0^\top A^k B = 0$ for all $k \geq 0$, and so $x_0^\top e^{tA} B = 0$ for all $t \in [0,T]$. But this implies that $x_0 \notin \mathscr{R}_T$, since otherwise if some $u \in C[0,T]$,

$$x_0 = \int_0^T e^{(T-\tau)A} B u(\tau) d\tau,$$

then

$$x_0^\top x_0 = \int_0^T x_0 e^{(T-\tau)A} B u(\tau) d\tau = \int_0^T 0 u(\tau) d\tau = 0,$$

which yields $x_0 = 0$, a contradiction. ∎

The following result gives an important characterization of controllability.

Corollary 3.5.4 *Let $T > 0$. The system (3.14) is controllable at T iff*

$$\operatorname{rank}\left[\begin{array}{c|c|c|c} B & AB & \dots & A^{n-1}B \end{array}\right] = n,$$

where n denotes the dimension of the state space.

Proof Only if: Let $x'(t) = Ax(t) + Bu(t)$ be controllable at time T. Then with $x_0 = 0$, all the states $x_1 \in \mathbb{R}^n$ can be reached at time T. So $\mathscr{R}_T = \mathbb{R}^n$. Hence by Theorem 3.5.3, $\operatorname{rank}\left[\begin{array}{c|c|c|c} B & AB & \dots & A^{n-1}B \end{array}\right] = n$.

If: Let $\operatorname{rank}\left[\begin{array}{c|c|c|c} B & AB & \dots & A^{n-1}B \end{array}\right] = n$. Then by Theorem 3.5.3, $\mathscr{R}_T = \mathbb{R}^n$. Given $x_0, x_1 \in \mathbb{R}^n$, we have $x_1 - e^{TA}x_0 \in \mathbb{R}^n = \mathscr{R}_T$, and so there exists a $u \in C[0,T]$ such that

$$x_1 - e^{TA}x_0 = \int_0^T e^{(T-\tau)A} B u(\tau) d\tau, \text{ that is, } x_1 = e^{TA}x_0 + \int_0^T e^{(T-\tau)A} B u(\tau) d\tau.$$

In other words $x(T) = x_1$, where $x(\cdot)$ denotes the unique solution to $x'(t) = Ax(t) + Bu(t)$, $t \in [0,T]$, $x(0) = x_0$. ∎

We remark that the test:

$$\operatorname{rank}\left[\begin{array}{c|c|c|c} B & AB & \dots & A^{n-1}B \end{array}\right] = n$$

is independent of T, and so it follows that if $T_1, T_2 > 0$, then the system $x'(t) = Ax(t) + Bu(t)$ is controllable at T_1 iff it is controllable at T_2. So for the system $x'(t) = Ax(t) + Bu(t)$, we usually talk about 'controllability' instead of 'controllability at $T > 0$'.

Examples. Consider the two examples on page 54.

1. *(Controllable system)* In the first example,

$$\operatorname{rank}\begin{bmatrix} B & AB & \dots & A^{n-1}B \end{bmatrix} \overset{(n=1)}{=} \operatorname{rank}\begin{bmatrix} B \end{bmatrix} = \operatorname{rank}\begin{bmatrix} 1 \end{bmatrix} = 1 = n,$$

   the dimension of the state space ($\mathbb{R}$).

2. *(Uncontrollable system)* In the second example, note that

$$\operatorname{rank}\begin{bmatrix} B & AB & \dots & A^{n-1}B \end{bmatrix} \overset{(n=2)}{=} \operatorname{rank}\begin{bmatrix} B & AB \end{bmatrix} = \operatorname{rank}\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} = 1 \neq 2 = n,$$

   the dimension of the state space ($\mathbb{R}^2$). ◊

Exercises.

1. For what values of α is the system (3.14) controllable, if

$$A = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ \alpha \end{bmatrix}?$$

2. (∗) Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times 1}$. Prove that if the system (3.14) is controllable, then every matrix commuting with A is a polynomial in A.

3. Let

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

   Show that the reachable subspace $\mathscr{R}_T$ at time T of $x'(t) = Ax(t) + Bu(t)$, $t \geq 0$, is equal to

$$\operatorname{span}\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \text{ that is, the set } \left\{ \begin{bmatrix} \alpha \\ 0 \end{bmatrix} \;\middle|\; \alpha \in \mathbb{R} \right\}.$$

4. A nonzero vector $v \in \mathbb{R}^{1\times n}$ is called a *left eigenvector* of $A \in \mathbb{R}^{n\times n}$ if there exists a $\lambda \in \mathbb{R}$ such that $vA = \lambda v$.

   Show that if the system described by $x'(t) = Ax(t) + Bu(t)$, $t \geq 0$, is controllable, then for every left eigenvector v of A, there holds that $vB \neq 0$.

   HINT: Observe that if $vA = \lambda v$, then $vA^k = \lambda^k v$ for all $k \in \mathbb{N}$.

3.6 How do we control optimally?

The questions of controlling a system optimally arise naturally. For example, in the case of an aircraft, we are not just interested in flying from one place to another, but we would also like to do so in a way so that the total travel time is minimized or the fuel consumption is minimized. With our algebraic equation $x + u = 10$, in which we want $x < 5$, suppose that furthermore we

wish to do so in manner such that u is the least possible integer. Then the only possible choice of the (input) u is 6. Optimal control addresses similar questions with differential equations of the type (3.2), together with a 'performance index functional', which is a function that measures optimality.

Example. Consider the model discussed in the Example on page 45:

$$x'(t) = f(x(t)) - x(t)u(t),$$

where x denotes the population of the fish, and u is the control input (the harvesting effort). Suppose that the profit associated with the fishing harvest over a time interval $[0, T]$ is given by

$$I(u) = \int_0^T e^{-rt} \left(px(t)u(t) - cu(t)\right) dt.$$

(Here p is the profit per unit harvest, so that p multiplied by the actual harvest $x(t)u(t)$ gives the profit at time t, and c is the cost per unit effort, so that c times the harvesting effort $u(t)$ gives the cost incurred at time t. The factor e^{-rt} is the discounting factor.)

The problem of deciding how to harvest so that the above profit is maximized now arises: that is, what is the $u \in C[0, T]$ that maximizes I? So we see that optimal control problems arise naturally in applications. ◊

In the next two chapters, we will learn about the basic principles behind optimal control theory.

Chapter 4

Optimal control

4.1 The simplest optimal control problem

In this section, we wish to find the functions u_0 such that the function I defined below has a local extremum at u_0:

$$I(u) = \int_0^T F(x(t), u(t), t)dt,$$

where x is the unique solution of the differential equation

$$x'(t) = f(x(t), u(t)), \quad t \in [0, T], \quad x(0) = x_i.$$

Such a local extremum u_0 is henceforth referred to as an *optimal control*.

We prove the following result.

Theorem 4.1.1 *Let $F(\xi, \upsilon, \tau)$ and $f(\xi, \upsilon)$ be continuously differentiable functions of each of their arguments. Suppose that $u_0 \in C[0, T]$ is an optimal control for the function $I : C[0, T] \to \mathbb{R}$ defined as follows: If $u \in C[0, T]$, then*

$$I(u) = \int_0^T F(x(t), u(t), t)dt,$$

where x denotes the unique solution to the differential equation

$$x'(t) = f(x(t), u(t)), \quad t \in [0, T], \quad x(0) = x_i. \tag{4.1}$$

If x_0 denotes the state corresponding to the input u_0, then there exists a $p_0 \in C^1[0, T]$ such that

$$\frac{\partial F}{\partial \xi}(x_0(t), u_0(t), t) + p_0(t)\frac{\partial f}{\partial \xi}(x_0(t), u_0(t)) = -p_0'(t), \quad t \in [0, T], \quad p_0(T) = 0, \tag{4.2}$$

$$\frac{\partial F}{\partial \upsilon}(x_0(t), u_0(t), t) + p_0(t)\frac{\partial f}{\partial \upsilon}(x_0(t), u_0(t)) = 0, \qquad t \in [0, T]. \tag{4.3}$$

Proof The proof can be divided into three main steps.

Step 1. In this step we consider an associated function $I_v : \mathbb{R} \to \mathbb{R}$ (that is defined below in terms of the functional I). Using the optimality of u_0 for I, we then conclude that the function

I_v must have a local extremum at 0 ($\in \mathbb{R}$). Thus applying the necessity of the condition that the derivative must vanish at extremal points (now simply for a function from $\mathbb{R}$ to $\mathbb{R}$!), we obtain a certain condition, given by equation (4.6).

Let $v \in C[0,T]$ be such that $v(0) = v(T) = 0$. Define $u_\epsilon(t) = u_0(t) + \epsilon v(t)$, $\epsilon \in \mathbb{R}$. Then from Theorem 3.2.2, for all ϵ such that $|\epsilon| < \delta$, with δ small enough, there exists a unique x_ϵ satisfying

$$x_\epsilon'(t) = f(x_\epsilon(t), u_\epsilon(t)), \quad t \in [0,T], \quad x_\epsilon(0) = x_i. \tag{4.4}$$

Let $I_v : (-\delta, \delta) \to \mathbb{R}$ be defined by

$$I_v(\epsilon) = \int_0^T F(x_\epsilon(t), u_\epsilon(t), t)dt.$$

Since I has a local has a local extremum at u_0, it follows that I_v has a local extremum at 0, and so $\frac{dI_v}{d\epsilon}(0) = 0$.

We have

$$\frac{dI_v}{d\epsilon}(0) = \int_0^T \left[\frac{\partial F}{\partial \xi}(x_0(t), u_0(t), t)\frac{d}{d\epsilon}(x_\epsilon(t))(0) + \frac{\partial F}{\partial \upsilon}(x_0(t), u_0(t), t)\frac{d}{d\epsilon}(u_\epsilon(t))(0)\right] dt.$$

(Differentiation under the integral sign can be justified!) Clearly $\frac{d}{d\epsilon}(u_\epsilon(t))(0) = v(t)$, and from Theorem 3.4.2, we also have that

$$x_\epsilon(t) = x_0(t) + \epsilon y(t) + o(t, \epsilon), \tag{4.5}$$

where y is the solution to

$$y'(t) = \frac{\partial f}{\partial \xi}(x_0(t), u_0(t))y(t) + \frac{\partial f}{\partial \upsilon}(x_0(t), u_0(t))v(t), \quad t \in [0,T], \quad y(0) = 0, \tag{4.6}$$

so that

$$\frac{d}{d\epsilon}(x_\epsilon(t))(0) = \lim_{\epsilon \to 0} \frac{x_\epsilon(t) - x_0(t)}{\epsilon} = \lim_{\epsilon \to 0} \frac{\epsilon y(t) + o(t,\epsilon)}{\epsilon} = y(t).$$

Hence

$$\frac{dI_v}{d\epsilon}(0) = \int_0^T \left[\frac{\partial F}{\partial \xi}(x_0(t), u_0(t), t)y(t) + \frac{\partial F}{\partial \upsilon}(x_0(t), u_0(t), t)v(t)\right] dt,$$

and so we obtain

$$\int_0^T \left[\frac{\partial F}{\partial \xi}(x_0(t), u_0(t), t)y(t) + \frac{\partial F}{\partial \upsilon}(x_0(t), u_0(t), t)v(t)\right] dt = 0. \tag{4.7}$$

Step 2. We now introduce an function p in order to rewrite (4.7) in a different manner, which will eventually help us to obtain (4.2) and (4.3).

Let $p \in C^1[0,T]$ be an unspecified function right now. Multiplying (4.6) by p, we have that

$$p(t)\left[\frac{\partial f}{\partial \xi}(x_0(t), u_0(t))y(t) + \frac{\partial f}{\partial \upsilon}(x_0(t), u_0(t))v(t) - y'(t)\right] = 0, \quad t \in [0,T]. \tag{4.8}$$

Thus adding the left hand side of (4.8) to the integrand in (4.7) does not change the integral. Consequently,

$$\int_0^T \left[\left(\frac{\partial F}{\partial \xi}(x_0(t), u_0(t), t) + p(t)\frac{\partial f}{\partial \xi}(x_0(t), u_0(t))\right) y(t) + \right.$$
$$\left.\left(\frac{\partial F}{\partial \upsilon}(x_0(t), u_0(t), t) + p(t)\frac{\partial f}{\partial \upsilon}(x_0(t), u_0(t))\right) v(t) - p(t)y'(t)\right] dt = 0.$$

Hence

$$\int_0^T \left[\left(\frac{\partial F}{\partial \xi}(x_0(t), u_0(t), t) + p(t) \frac{\partial f}{\partial \xi}(x_0(t), u_0(t)) + \dot{p}(t) \right) y(t) + \left(\frac{\partial F}{\partial v}(x_0(t), u_0(t), t) + p(t) \frac{\partial f}{\partial v}(x_0(t), u_0(t)) \right) v(t) \right] dt + p(t) y(t) \Big|_{t=0}^{t=T} = 0. \quad (4.9)$$

STEP 3. In this final step, we choose the 'right p': one which makes the first summand in the integrand appearing in (4.9) vanish (in other other words a solution of the differential equation in (4.7)) and impose a boundary condition for this special (denoted by p_0) in such a manner that the boundary term in (4.9) also disappears. With this choice of p, (4.9) allows one to conclude that (4.3) holds too!

Now choose $p = p_0$, where p_0 is such that

$$\frac{\partial F}{\partial \xi}(x_0(t), u_0(t), t) + p_0(t) \frac{\partial f}{\partial \xi}(x_0(t), u_0(t)) + p_0'(t) = 0, \quad t \in [0, T], \quad p_0(T) = 0. \quad (4.10)$$

(It is easy to verify that

$$p_0(t) = \int_t^T \frac{\partial F}{\partial \xi}(x_0(s), u_0(s), s) e^{\int_t^s \frac{\partial f}{\partial \xi}(x_0(\tau), u_0(\tau)) d\tau} ds$$

satisfies (4.10).) Thus (4.9) becomes

$$\int_0^T \left(\frac{\partial F}{\partial v}(x_0(t), u_0(t), t) + p(t) \frac{\partial f}{\partial v}(x_0(t), u_0(t)) \right) v(t) dt = 0.$$

Since the choice of $v \in C[0, T]$ satisfying $v(0) = v(T) = 0$ was arbitrary, it follows that (4.3) holds: Indeed, if not, then the left hand side of (4.3) is nonzero (say positive) at some point in $[0, T]$, and by continuity, it is also positive in some interval $[t_1, t_2]$ contained in $[0, T]$. Set

$$v(t) = \begin{cases} (t - t_1)(t_2 - t) & \text{if } t \in [t_1, t_2], \\ 0 & \text{if } t \notin [t_1, t_2]. \end{cases}$$

Then $v \in C[0, T]$ and $v(0) = v(T) = 0$. However,

$$\begin{aligned} & \int_0^T \left(\frac{\partial F}{\partial v}(x_0(t), u_0(t), t) + p(t) \frac{\partial f}{\partial v}(x_0(t), u_0(t)) \right) v(t) dt \\ = & \int_{t_1}^{t_2} \left(\frac{\partial F}{\partial v}(x_0(t), u_0(t), t) + p(t) \frac{\partial f}{\partial v}(x_0(t), u_0(t)) \right) (t - t_1)(t_2 - t) dt \\ > & \ 0, \end{aligned}$$

a contradiction. This completes the proof of the theorem. ■

Remarks.

1. The p_0 is analogous to the Lagrange multiplier encountered in constrained optimization problems in finite dimensions: a necessary condition for $x_0 \in \mathbb{R}^n$ to be an extremum of $F : \mathbb{R}^n \to \mathbb{R}$ subject to $G : \mathbb{R}^n \to \mathbb{R}^k$ is that $D\tilde{F}(x_0) = 0$, where $\tilde{F} = F + p_0^\top G$ for some $p_0 \in \mathbb{R}^k$. The role of the Lagrange multiplier $p_0 \in \mathbb{R}^k$, which is vector in the *finite*-dimensional vector space $\mathbb{R}^k$, is now played by the *function* p_0, which is a vector in an *infinite*-dimensional vector space.

2. It should be emphasized that Theorem 4.1.1 provides a *necessary* condition for optimality. Thus not every u_0 that satisfies (4.2) and (4.3) (for some p_0, and with x_0 denoting the unique solution to (4.1)), needs to be optimal. Such a u_0 satisfying is called a *critical control*. However, if we already know that an optimal solution exists and that there is a unique critical control, then this critical control is obviously optimal.

4.2 The Hamiltonian and Pontryagin minimum principle

With the notation from Theorem 4.1.1, define

$$H(\pi, \xi, \upsilon, \tau) = F(\xi, \upsilon, \tau) + \pi f(\xi, \upsilon). \tag{4.11}$$

H is called the *Hamiltonian* and Theorem 4.1.1 can be equivalently be expressed in the following form.

Theorem 4.2.1 *Let $F(\xi, \upsilon, \tau)$ and $f(\xi, \upsilon)$ be continuously differentiable functions of each of their arguments. If $u_0 \in C[0,T]$ is an optimal control for the function I given by*

$$I(u) = \int_0^T F(x(t), u(t), t)dt, \quad u \in C[0,T],$$

where x is the solution to

$$x'(t) = f(x(t), u(t)), \quad t \in [0,T], \quad x(0) = x_i,$$

and if x_0 denotes the state corresponding to u_0, then there exists a $p_0 \in C^1[0,T]$ such that

$$\frac{\partial H}{\partial \xi}(p_0(t), x_0(t), u_0(t), t) = -p_0'(t), \quad t \in [0,T], \quad p_0(T) = 0, \text{ and} \tag{4.12}$$

$$\frac{\partial H}{\partial \upsilon}(p_0(t), x_0(t), u_0(t), t) = 0, \qquad t \in [0,T]. \tag{4.13}$$

Remarks.

1. Note that the differential equation $x_0' = f(x_0, u_0)$ with $x_0(0) = x_i$ can be expressed in terms of the Hamiltonian as follows:

$$\frac{\partial H}{\partial \pi}(p_0(t), x_0(t), u_0(t), t) = x_0'(t), \quad t \in [0,T], \quad x_0(0) = x_i. \tag{4.14}$$

   The equations (4.12) and (4.14) resemble the equations arising in Hamiltonian mechanics, and these equations together are said to comprise a *Hamiltonian differential system*.

   The function p_0 is called the *co-state*, and (4.12) is called the *adjoint differential equation*. This analogy with Hamiltonian mechanics was responsible for the original terminology of calling H the Hamiltonian.

2. In Theorem 4.2.1, it can in fact be shown that for all $t \in [0,T]$,

$$H(p_0(t), x_0(t), u(t), t) \geq H(p_0(t), x_0(t), u_0(t), t) \tag{4.15}$$

   holds for all $u \in C[0,T]$, that is the optimal input u_0 *minimizes* the Hamiltonian (inequality (4.15)). This is known as *Pontryagin minimum principle*. Equation (4.13) is then a corollary of this result.

Exercises.

1. Find a critical control of the function

$$I(u) = \int_0^1 \left[(x(t))^2 + (u(t))^2\right] dt$$

where x is the solution to $x'(t) = u(t)$, $t \in [0,1]$, $x(0) = x_0$.

2. Find a critical control u_T and the corresponding state x_T of the function

$$I(u) = \int_0^T \frac{1}{2}\left(3(x(t))^2 + (u(t))^2\right) dt$$

where x is the solution to $x'(t) = x(t) + u(t)$, $t \in [0,T]$, $x(0) = x_0$. Show that there exists a constant k such that[1]

$$\lim_{T\to\infty} u_T(t) = F \lim_{T\to\infty} x_T(t)$$

for all t. What is the value of F?

4.3 Generalization to vector inputs and states

In the general case when $x(t) \in \mathbb{R}^n$ and $u(t) \in \mathbb{R}^m$, Theorem 4.2.1 holds with p_0 now being a function taking its values in $\mathbb{R}^n$:

Theorem 4.3.1 *Let $F(\xi, \upsilon, \tau)$ and $f(\xi, \upsilon)$ be continuously differentiable functions of each of their arguments. If $u_0 \in (C[0,T])^m$ is an optimal control for the function*

$$I(u) = \int_0^T F(x(t), u(t), t)dt,$$

where x is the solution to $x'(t) = f(x(t), u(t))$, $t \in [0,T]$, $x(0) = x_i$, and if x_0 denotes the state corresponding to u_0, then there exists a $p_0 \in (C^1[0,T])^n$ such that

$$\begin{aligned}\left[\frac{\partial H}{\partial \xi}(p_0(t), x_0(t), u_0(t), t)\right]^\top &= -p_0'(t), \qquad t \in [0,T], \quad p_0(T) = 0, \text{ and}\\ \frac{\partial H}{\partial \upsilon}(p_0(t), x_0(t), u_0(t), t) &= 0, \qquad t \in [0,T],\end{aligned}$$

where $H(\pi, \xi, \upsilon, \tau) = F(\xi, \upsilon, \tau) + \pi^\top f(\xi, \upsilon)$.

Example. *(Linear systems and the Riccati equation)* Let $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times 1}$, $Q \in \mathbb{R}^{n\times n}$ such that $Q = Q^\top \geq 0$ and $R \in \mathbb{R}$ such that $R > 0$. We wish to find[2] optimal controls for the functional

$$I(u) = \int_0^T \frac{1}{2}\left[x(t)^\top Q x(t) + R(u(t))^2\right] dt$$

subject to the differential equation

$$x'(t) = Ax(t) + Bu(t), \quad t \in [0,T], \quad x(0) = x_i.$$

[1] A control of the type $u(t) = Fx(t)$ is said to be a *static state-feedback*.
[2] This is called the *linear quadratic control problem* or *LQ problem*.

The Hamiltonian is given by

$$H(\pi,\xi,\upsilon,\tau) = \frac{1}{2}\left[\xi^\top Q\xi + R(\upsilon)^2\right] + \pi^\top\left[A\xi + B\upsilon\right].$$

From Theorem 4.3.1, it follows that any optimal input u_0 and the corresponding state x_0 satisfies

$$\frac{\partial H}{\partial \upsilon}(p_0(t), x_0(t), u_0(t), t) = 0,$$

that is, $Ru_0(t) + p_0(t)^\top B = 0$. Thus $u_0(t) = -R^{-1}B^\top p_0(t)$. The adjoint equation is

$$\left[\frac{\partial H}{\partial \xi}(p_0(t), x_0(t), u_0(t), t)\right]^\top = -p_0'(t), \quad t \in [0,T], \quad p_0(T) = 0,$$

that is, $(x_0(t)^\top Q + p_0(t)^\top A)^\top = -p_0'(t)$, $t \in [0,T]$, $p_0(T) = 0$. So we have

$$p_0'(t) = -A^\top p_0(t) - Qx_0(t), \quad t \in [0,T], \quad p_0(T) = 0.$$

Consequently,

$$\frac{d}{dt}\begin{bmatrix} x_0(t) \\ p_0(t) \end{bmatrix} = \begin{bmatrix} A & -BR^{-1}B^\top \\ -Q & -A^\top \end{bmatrix}\begin{bmatrix} x_0(t) \\ p_0(t) \end{bmatrix}, \quad t \in [0,T], \quad x_0(0) = x_i, \quad p_0(T) = 0. \quad (4.16)$$

This is a linear, time-invariant differential equation in (x_0, p_0). If we would only have to deal with *initial* boundary conditions exclusively or *final* boundary conditions exclusively, then we could easily solve (4.16). However, here we have *combined* initial and final conditions, and so it is not clear how we could solve (4.16). It is unclear if (4.16) has a solution at all! We now prove the following.

Theorem 4.3.2 *Let P be a solution of the following* Riccati equation

$$P'(t) = -P(t)A - A^\top P(t) + P(t)BR^{-1}B^\top P(t) - Q, \quad t \in [0,T], \quad P(T) = 0.$$

Let x_0 be the solution of

$$x_0'(t) = \left[A - BR^{-1}B^\top P(t)\right]x_0(t), \quad t \in [0,T], \quad x_0(0) = x_i,$$

and let $p_0(t) = P(t)x_0(t)$. Then (x_0, p_0) above is the unique solution of (4.16).

Proof We have

$$\begin{aligned}
\frac{d}{dt}\begin{bmatrix} x_0(t) \\ p_0(t) \end{bmatrix} &= \begin{bmatrix} (A - BR^{-1}B^\top P(t))x_0(t) \\ P'(t)x_0(t) + P(t)x_0'(t) \end{bmatrix} \\
&= \begin{bmatrix} Ax_0(t) - BR^{-1}B^\top p_0(t) \\ \left(-P(t)Ax_0(t) - A^\top P(t)x_0(t) + P(t)BR^{-1}B^\top P(t)x_0(t) - Qx_0(t) + \right. \\ \qquad\qquad \left. P(t)Ax_0(t) - P(t)BR^{-1}B^\top P(t)x_0(t)\right) \end{bmatrix} \\
&= \begin{bmatrix} Ax_0(t) - BR^{-1}B^\top p_0(t) \\ -Qx_0(t) - A^\top p_0(t) \end{bmatrix} \\
&= \begin{bmatrix} A & -BR^{-1}B^\top \\ -Q & -A^\top \end{bmatrix}\begin{bmatrix} x_0(t) \\ p_0(t) \end{bmatrix}.
\end{aligned}$$

Furthermore, x_0 and p_0 satisfy $x(0) = x_i$ and $p_0(T) = P(T)x_0(T) = 0x_0(T) = 0$. So the pair (x_0, p_0) satisfies (4.16).

The uniqueness can be shown as follows. If (x_1, p_1) and (x_2, p_2) satisfy (4.16), then $\tilde{x} = x_1 - x_2$, $\tilde{p} = p_1 - p_2$ satisfy

$$\frac{d}{dt}\begin{bmatrix} \tilde{x}(t) \\ \tilde{p}(t) \end{bmatrix} = \begin{bmatrix} A & -BR^{-1}B^\top \\ -Q & -A^\top \end{bmatrix}\begin{bmatrix} \tilde{x}(t) \\ \tilde{p}(t) \end{bmatrix}, \quad t \in [0,T], \quad \tilde{x}(0) = 0, \quad \tilde{p}(T) = 0. \tag{4.17}$$

This implies that

$$\begin{aligned} 0 &= \tilde{p}(T)^\top \tilde{x}(T) - \tilde{p}(0)^\top \tilde{x}(0) \\ &= \int_0^T \frac{d}{dt}\left(\tilde{p}(t)^\top \tilde{x}(t)\right) dt \\ &= \int_0^T \left[\tilde{p}'(t)^\top \tilde{x}(t) + \tilde{p}(t)^\top \tilde{x}'(t)\right] dt \\ &= \int_0^T \left[(-Q\tilde{x}(t) - A^\top \tilde{p}(t))\tilde{x}(t) + \tilde{p}(t)^\top (A\tilde{x}(t) - BR^{-1}B^\top \tilde{p}(t))\right] dt \\ &= \int_0^T \left[\tilde{x}(t)^\top Q\tilde{x}(t) + \tilde{p}(t)^\top BR^{-1}B^\top \tilde{p}(t)\right] dt. \end{aligned}$$

Consequently $Q\tilde{x}(t) = 0$ and $R^{-1}B^\top \tilde{p}(t) = 0$ for all $t \in [0,T]$. From (4.17), we obtain

$$\begin{aligned} \tilde{x}'(t) &= A\tilde{x}(t), \qquad t \in [0,R], \quad \tilde{x}(0) = 0, \text{ and} \\ \tilde{p}'(t) &= -A^\top \tilde{p}(t), \quad t \in [0,T], \quad \tilde{p}(T) = 0. \end{aligned}$$

Thus $\tilde{x}(t) = 0$ and $\tilde{p}(t) = 0$ for all $t \in [0,T]$. ■

So we see that the optimal trajectories (x_0, u_0) are governed by

$$\begin{aligned} x_0'(t) &= \left[A - BR^{-1}B^\top P(t)\right] x_0(t), \qquad t \in [0,T], \quad x_0(0) = x_i, \\ u_0(t) &= -R^{-1}B^\top P(t) x_0(t), \qquad t \in [0,T], \end{aligned}$$

where P is the solution of the Riccati equation

$$P'(t) = -P(t)A - A^\top P(t) + P(t)BR^{-1}B^\top P(t) - Q, \quad t \in [0,T], \quad P(T) = 0.$$

Note that the optimal control has the form of a (time-varying) state-feedback law; see Figure 4.1. ◊

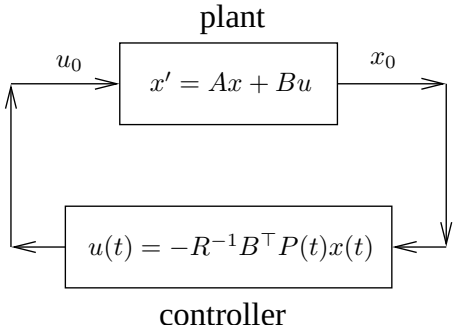


Figure 4.1: The closed loop system.

Exercise. Let $Q \in \mathbb{R}^{n\times n}$ be such that $Q = Q^\top \geq 0$. Show that if $x \in \mathbb{R}^n$ is such that $x^\top Qx = 0$, then $Qx = 0$.

4.4 Constraint on the state at final time.

In many optimization problems, in addition to minimizing the cost, one may also have to satisfy a condition for the final state $x(T)$; for instance, one may wish to drive the state to zero. This brings us naturally to the notion of controllability. For the sake of simplicity, we restrict ourselves to linear systems:

$$x'(t) = Ax(t) + Bu(t), \quad t \geq 0. \tag{4.18}$$

The following theorem tells us how we can calculate the optimal control when $x(T)$ is specified, in the case of controllable linear systems.

Theorem 4.4.1 *Suppose that the system*

$$x'(t) = Ax(t) + Bu(t), \quad t \geq 0$$

is controllable. Let $F(\xi, \upsilon, \tau)$ be a continuously differentiable function of each of their arguments. If $u_0 \in C[0,T]$ is an optimal control for the function I given by

$$I(u) = \int_0^T F(x(t), u(t), t)dt,$$

where x is the solution to

$$x'(t) = Ax(t) + Bu(t), \quad t \in [0,T], \quad x(0) = x_i, \quad x(T)_k = x_{f,k}, \quad k \in \{1, \ldots, r\},$$

and if x_0 denotes the state corresponding to u_0, then there exists a $p_0 \in (C^1[0,T])^n$ such that

$$\begin{aligned} \left[\frac{\partial H}{\partial \xi}(p_0(t), x_0(t), u_0(t), t)\right]^\top &= -p_0'(t), \quad t \in [0,T], \quad p_0(T)_k = 0, \quad k \in \{r+1, \ldots, n\}, \text{ and} \\ \frac{\partial H}{\partial \upsilon}(p_0(t), x_0(t), u_0(t), t) &= 0, \qquad t \in [0,T], \end{aligned}$$

where $H(\pi, \xi, \upsilon, \tau) = F(\xi, \upsilon, \tau) + \pi^\top(A\xi + B\upsilon)$.

We will not prove this theorem. Note that for a differential equation to have a unique solution, there should not be too few or too many initial and final conditions to be satisfied by that solution. Intuitively, one expects as many conditions as there are differential equations. In Theorem 4.4.1, we have, in total, $2n$ differential equations (for x_0 and p_0). We also have the right number of conditions: $n + r$ for x_0, and $n - r$ for p_0.

Exercises.

1. Find a critical control for the function

   $$I(u) = \int_0^1 (u(t))^2 dt$$

   subject to $x'(t) = -2x(t) + u(t)$, $t \in [0,1]$, $x(0) = 1$ and $x(1) = 0$. Is this control unique?

2. Find a critical control for the function

   $$I(u) = \int_0^T (u(t))^2 dt$$

   subject to $x'(t) = -ax(t) + u(t)$, $t \in [0,T]$, $x(0) = x_0$ and $x(T) = 0$. Find an expression for the corresponding state. Prove that the critical control can be expressed as a state-feedback law: $u(t) = F(t,T,a)x(t)$. Find an expression for $F(t,T,a)$.

3. Find a critical control for the function

$$I(u) = \int_0^T \left[(x_T - x(t))^2 + (u(t))^2\right] dt$$

subject to $x'(t) = -ax(t) + u(t)$, $t \in [0, T]$, $x(0) = x_0$ and $x(T) = x_T$.

4. Find a critical control for the function

$$I(u) = \int_0^1 \frac{1}{2}[(x_1(t))^2 + (x_2(t))^2 + (u(t))^2]dt$$

where x_1 and x_2 are solutions to the system

$$\begin{aligned} x_1'(t) &= x_2(t), \\ x_2'(t) &= -x_2(t) + u(t) \end{aligned}$$

$t \in [0, 1]$ and

$$\begin{aligned} x_1(0) &= 1, \qquad x_2(0) = 1, \\ x_1(1) &= 0, \qquad x_2(1) = 0. \end{aligned}$$

5. *(Higher order differential equation constraint.)* Find a critical control for the function

$$I(u) = \int_0^T \frac{1}{2}\left[(y(t))^2 + (u(t))^2\right] dt$$

where y is the solution to the second order differential equation

$$y''(t) + y(t) = u(t), \quad t \in [0, T], \quad y(0) = y_0, \quad y'(0) = v_0, \quad y(T) = y'(T) = 0.$$

HINT: Introduce the state variables $x_1(t) = y(t)$, $x_2(t) = y'(t)$.

Chapter 5

Optimal control II

Bellman and his co-workers pioneered a different approach for solving optimal control problems. So far we have considered *necessary* conditions for the existence of an optimal control. In Theorem 5.2.1 we give *sufficient* conditions for the existence of an optimal control.

5.1 The optimality principle

The underlying idea of the optimality principle is extremely simple. Roughly speaking, the *optimality principle* simply says that any part of an optimal trajectory is optimal.

Theorem 5.1.1 (Optimality principle.) *Let $F(\xi, \upsilon, \tau)$ and $f(\xi, \upsilon)$ be continuously differentiable functions of each of their arguments. Let $u_0 \in C[0,T]$ be an optimal control for the function I given by*

$$I(u) = \int_0^T F(x(t), u(t), t)dt, \quad u \in C[0,T],$$

where x is the solution to

$$x'(t) = f(x(t), u(t)), \quad t \in [0,T], \quad x(0) = x_i. \tag{5.1}$$

Let x_0 be the state corresponding to u_0. If $t_* \in [0,T)$, then the restriction of u_0 to $[t_*, T]$ is an optimal control for the function $\tilde{I}$ given by

$$\tilde{I}(u) = \int_{t_0}^T F(x(t), u(t), t)dt, \quad u \in C[t_*, T]$$

where x is the solution to

$$x'(t) = f(x(t), u(t)), \quad t \in [t_*, T], \quad x(t_*) = x_0(t_*). \tag{5.2}$$

Furthermore,

$$\min_{u \in C[0,T]} I(u) = \int_0^{t_*} F(x_0(t), u_0(t), t)dt + \min_{u \in C[t_*,T]} \tilde{I}(u). \tag{5.3}$$

Proof We have

$$\begin{aligned} I(u_0) &= \int_0^T F(x_0(t), u_0(t), t)dt \\ &= \int_0^{t_*} F(x_0(t), u_0(t), t)dt + \int_{t_*}^T F(x_0(t), u_0(t), t)dt. \end{aligned} \tag{5.4}$$

From Theorem 3.2.1, it follows that the solution to

$$x'(t) = f\left(x(t), u_0|_{[t_*,T]}(t)\right), \quad t \in [t_*, T], \quad x(t_*) = x_0(t_*),$$

is simply the restriction of x_0 to $[t_*, T]$. Thus the second term in (5.4) is the cost $\tilde{I}\left(u_0|_{[t_*,T]}\right)$ corresponding to input $u_0|_{[t_*,T]} \in C[t_*, T]$.

Suppose that there exists a $\tilde{u} \in C[t_*, T]$ such that

$$\int_{t_*}^T F(\tilde{x}(t), \tilde{u}(t), t)dt = \tilde{I}(\tilde{u}) < \tilde{I}\left(u_0|_{[t_*,T]}\right) = \int_{t_*}^T F(x_0(t), u_0(t), t)dt, \tag{5.5}$$

where $\tilde{x}$ is the solution to (5.2) corresponding to $\tilde{u}$.

By Theorem 3.4.2, it follows that [1] we can then choose a $u_* \in C[t_*, T]$ such that $u_*(t_*) = u_0(t_*)$ and such that the inequality in (5.5) is still holds with u_*, that is, $\tilde{I}(u_*) < \tilde{I}\left(u_0|_{[t_*,T]}\right)$.

Define $u \in C[0, T]$ by

$$u(t) = \begin{cases} u_0(t) & \text{for } t \in [0, t_*), \\ u_*(t) & \text{for } t \in [t_*, T], \end{cases}$$

and let x be the corresponding solution to (5.1). From Theorem 3.2.1, it follows that

$$x|_{[0,t_*]} = x_0|_{[0,t_*]}.$$

Hence we have

$$\begin{aligned} I(u) &= \int_0^T F(x(t), u(t), t)dt \\ &= \int_0^{t_*} F(x(t), u(t), t)dt + \int_{t_*}^T F(x(t), u(t), t)dt \\ &= \int_0^{t_*} F(x_0(t), u_0(t), t)dt + \int_{t_*}^T F(x(t), u_*(t), t)dt \\ &= \int_0^{t_*} F(x_0(t), u_0(t), t)dt + \tilde{I}(u_*) \\ &< \int_0^{t_*} F(x_0(t), u_0(t), t)dt + \tilde{I}\left(u_0|_{[t_*,T]}(t)\right) \\ &= \int_0^{t_*} F(x_0(t), u_0(t), t)dt + \int_{t_*}^T F(x_0(t), u_0(t), t)dt \\ &= I(u_0), \end{aligned}$$

which contradicts the optimality of u_0. This proves that an optimal control for $\tilde{I}$ exists and it is given by the restriction of u_0 to $[t_*, T]$. From (5.4), it follows that (5.3) holds. ■

[1] For instance, let $u_*(t) := \tilde{u}(t) + (u_0(t_*) - \tilde{u}(t_*))\Theta(t)$, $t \in [t_*, T]$, where $\Theta(t) = 1 - \frac{1}{\epsilon}(t - t_*)$ if $t < t_* + \epsilon$ and 0 otherwise, with ϵ small enough.

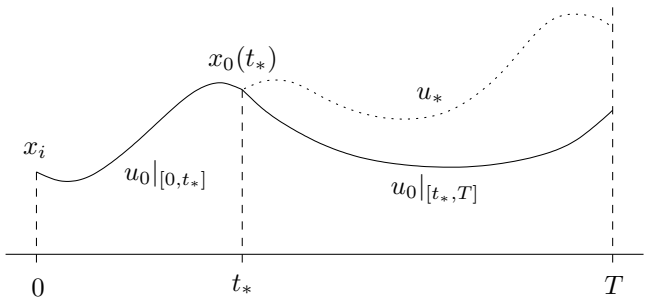


Figure 5.1: Optimality principle.

Note that we have shown that

$$\min_{u\in C[t_*,T]} \tilde{I}(u) = \tilde{I}\left(u_0|_{[t_*,T]}\right).$$

So the theorem above says that if you are on an optimal trajectory, then the best thing you can do is to stay on that trajectory. See Figure 5.1.

Example. Consider the function $I : C[0,T] \to \mathbb{R}$ defined by

$$I(u) = \int_0^T \frac{1}{2}[(x(t))^2 + (u(t))^2]dt,$$

where x is the solution to $x'(t) = u(t)$, $t \in [0,T]$, $x(0) = x_i$. The Hamiltonian is given by

$$H(\pi, \xi, \upsilon, \tau) = \frac{1}{2}(\xi^2 + \upsilon^2) + \pi\upsilon,$$

and so

$$\frac{\partial H}{\partial \xi} = \xi, \quad \frac{\partial H}{\partial \upsilon} = \upsilon.$$

Consequently, the equations governing an optimal control u_0 (with corresponding state denoted by x_0) are given by

$$\begin{aligned} x_0'(t) &= -p_0(t), \text{ and} \\ p_0'(t) &= -x_0(t), \end{aligned}$$

for some $p_0 \in C[0,T]$ such that $p(T) = 0$. Thus we obtain that

$$\begin{bmatrix} x_0(t) \\ p_0(t) \end{bmatrix} = e^{t\begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}} \begin{bmatrix} x_0(0) \\ p_0(0) \end{bmatrix}.$$

Hence

$$\begin{bmatrix} x_0(t) \\ p_0(t) \end{bmatrix} = \begin{bmatrix} \alpha e^t + \beta e^{-t} \\ -\alpha e^t + \beta e^{-t} \end{bmatrix},$$

for some constants α and β. Using $x_0(0) = x_i$ and $p_0(T) = 0$, we find that

$$\alpha = \frac{x_i e^{-2T}}{e^{-2T}+1} \quad \text{and } \beta = \frac{x_i}{e^{-2T}+1}.$$

Consequently

$$u_0(t) = -p_0(t) = x_i \frac{e^{-2T}e^t - e^{-t}}{e^{-2T}+1}, \quad t \in [0,T].$$

Now suppose we evolve along this optimal control starting from x_i till the time is $t_* \in (0,T)$. Then the state $x_0(t_*)$ can be found out as follows:

$$\begin{aligned} x_0(t_*) - x_i &= x_0(t_*) - x_0(0) = \int_0^{t_*} x_0'(t)dt = \int_0^{t_*} u_0(t)dt \\ &= \frac{x_i}{e^{-2T}+1}[e^{-2T}(e^{t_*}-1) + e^{-t_*} - 1] = \frac{x_i}{e^{-2T}+1}[e^{-2T+t_*} + e^{-t_*}] - x_i. \end{aligned}$$

Hence

$$x_0(t_*) = \frac{x_i}{e^{-2T}+1}[e^{-2T+t_*} + e^{-t_*}].$$

Now suppose we consider the optimization problem for the function $\tilde{I} : C[t_*, T] \to \mathbb{R}$, given by

$$\tilde{I}(u) = \int_{t_*}^{T} \frac{1}{2}[(x(t))^2 + (u(t))^2]dt,$$

where x is the solution to $x'(t) = u(t)$, $t \in [t_*, T]$, $x(t_*) = x_0(t_*)$. If we define $I_* : C[0, T-t_*] \to \mathbb{R}$ by $I_*(u) = \tilde{I}(u(\cdot - t_*))$, $u \in C[0, T-t_*]$, then it is easy to see that I_* has an optimal control u_* iff $\tilde{I}$ has the optimal control $\tilde{u}(\cdot) = u_*(\cdot - t_*)$, and moreover

$$I_*(u) = \int_0^{T-t_*} \frac{1}{2}[(x(t))^2 + (u(t))^2]dt,$$

where x is the solution to $x'(t) = u(t)$, $t \in [0, T-t_*]$, $x(0) = x_0(t_*)$. But from the calculation above, we see that if u_* is an optimal control for I_*, then

$$u_*(t) = x_0(t_*)\frac{e^{-2(T-t_*)}e^t - e^{-t}}{e^{-2(T-t_*)}+1}, \quad t \in [0, T-t_*].$$

Hence the optimal control for $\tilde{I}$ is given by

$$\begin{aligned} \tilde{u}(t) &= u_*(t-t_*) = x_0(t_*)\frac{e^{-2(T-t_*)}e^{t-t_*} - e^{-t+t_*}}{e^{-2(T-t_*)}+1} \\ &= \frac{x_i}{e^{-2T}+1}[e^{-2T+t_*} + e^{-t_*}]\frac{e^{-2T}e^{t+t_*} - e^{t_*-t}}{e^{-2(T-t_*)}+1} \\ &= x_i\frac{e^{-2T}e^t - e^{-t}}{e^{-2T}+1}, \end{aligned}$$

for $t \in [t_*, T]$. Hence we see that $\tilde{u} = u_0|_{[t_*,T]}$. ◊

5.2 Bellman's equation

In this section we will prove Theorem 5.2.1 below, which gives a *sufficient* condition for the existence of an optimal control in terms of the existence of an appropriate solution to Bellman's equation (5.7). However, we first provide a heuristic argument that leads one to Bellman's equation: we do not start by asking when the optimal control problem has a solution, but rather we begin by assuming that the optimal control problem is solvable and study the so-called value function, which will lead us to Bellman's equation.

Define the *value function* $V : \mathbb{R}^n \times [0,T] \to \mathbb{R}$ by

$$V(x_*, t_*) = \min_{u \in C[t_*,T]} \int_{t_*}^{T} F(x(t), u(t), t)dt, \tag{5.6}$$

where x is the unique solution to

$$x'(t) = f(x(t), u(t)), \quad t \in [t_*, T], \quad x(t_*) = x_*.$$

With this notation, in Theorem 5.1.1, we have shown that

$$V(x_0(t_*), t_*) = \min_{u \in C[t_*, T]} \int_{t_*}^{T} F(x(t), u(t), t)dt = \int_{t_*}^{T} F(x_0(t), u_0(t), t)dt.$$

Consequently

$$V(x_0(t_* + \epsilon), t_* + \epsilon) - V(x_0(t_*), t_*) = -\int_{t_*}^{t_*+\epsilon} F(x_0(t), u_0(t), t)dt.$$

It is tempting to divide by ϵ on both sides and let ϵ tend to 0. Formally, the left hand side would become

$$\frac{\partial V}{\partial \tau}(x_0(t_*), t_*) + \frac{\partial V}{\partial \xi}(x_0(t_*), t_*) f(x_0(t_*), u_0(t_*)),$$

while the right hand side would become

$$-F(x_0(t_*), u_0(t_*), t_*).$$

Thus we would obtain the equation

$$\frac{\partial V}{\partial \tau}(x_0(t_*), t_*) + \frac{\partial V}{\partial \xi}(x_0(t_*), t_*) f(x_0(t_*), u_0(t_*)) + F(x_0(t_*), u_0(t_*), t_*) = 0.$$

This motivates the following result.

Theorem 5.2.1 *Let $F(\xi, \upsilon, \tau)$ and $f(\xi, \upsilon)$ be continuously differentiable functions of each of their arguments. Suppose that there exists a function $W : \mathbb{R}^n \times [0, T] \to \mathbb{R}$ such that:*

1. *W is continuous on $\mathbb{R}^n \times [0, T]$.*
2. *W is continuously differentiable in $\mathbb{R}^n \times (0, T)$.*
3. *W satisfies* Bellman's equation

$$\frac{\partial W}{\partial \tau}(x, t) + \min_{u \in \mathbb{R}} \left[\frac{\partial W}{\partial \xi}(x, t) f(x, u) + F(x, u, t) \right] = 0, \quad (x, t) \in \mathbb{R}^n \times (0, T). \tag{5.7}$$

4. *$W(x, T) = 0$ for all $x \in \mathbb{R}^n$.*

Then the following implications hold:

1. *If $t_* \in [0, T)$ and $u \in C[t_*, T]$, then*

$$\int_{t_*}^{T} F(x(t), u(t), t)dt \geq W(x_*, t_*),$$

   where x is the unique solution to $x'(t) = f(x(t), u(t))$, $x(t_*) = x_*$, $t \in [t_*, T]$.

2. *If there exists a function $\varphi : \mathbb{R}^n \times [0, T] \to \mathbb{R}$ such that:*

(a) For all $(x,t) \in \mathbb{R}^n \times (0,T)$,

$$\frac{\partial W}{\partial \xi}(x,t) f(x,\varphi(x,t)) + F(x,\varphi(x,t),t) = \min_{u\in\mathbb{R}} \left[\frac{\partial W}{\partial \xi}(x,t) f(x,u) + F(x,u,t) \right].$$

(b) The equation

$$x'(t) = f(x(t),\varphi(x(t),t)), \quad t \in [0,T], \quad x(0) = x_i,$$

has a solution x_0.

(c) u_0 *defined by* $u_0(t) = \varphi(x_0(t),t)$, $t \in [0,T]$ *is an element in* $C[0,T]$.

Then u_0 *is an optimal control for the function* I *defined by*

$$I(u) = \int_0^T F(x(t),u(t),t)dt,$$

where x *is the solution to* $x'(t) = f(x(t),u(t))$, $t \in [0,T]$, $x(0) = x_i$, *and furthermore,*

$$I(u_0) = \int_0^T F(x_0(t),u_0(t),t)dt = W(x_i,0). \tag{5.8}$$

3. Let φ *be the function from part 2. If for every* $t_* \in [0,T)$ *and every* $x_* \in \mathbb{R}^n$, *the equation*

$$x'(t) = f(x(t),\varphi(x(t),t)), \quad t \in [t_*,T], \quad x(t_*) = x_*,$$

has a solution, then W *is the value function* V *defined in (5.6).*

Proof *1.* We have

$$\begin{aligned}
\int_{t_*}^T F(x(t),u(t),t)dt &= \int_{t_*}^T \left[\frac{\partial W}{\partial \xi}(x(t),t) f(x(t),u(t)) + F(x(t),u(t),t) \right] dt - \\
&\quad \int_{t_*}^T \frac{\partial W}{\partial \xi}(x(t),t) f(x(t),u(t))dt \\
&\geq \int_{t_*}^T \min_{u\in\mathbb{R}} \left[\frac{\partial W}{\partial \xi}(x(t),t) f(x(t),u) + F(x(t),u,t) \right] dt - \\
&\quad \int_{t_*}^T \frac{\partial W}{\partial \xi}(x(t),t) f(x(t),u(t))dt \\
&= \int_{t_*}^T \left[-\frac{\partial W}{\partial \tau}(x(t),t) - \frac{\partial W}{\partial \xi}(x(t),t) f(x(t),u(t)) \right] dt \\
&= -\int_{t_*}^T \left(\frac{d}{dt} W(x(\cdot),\cdot) \right)(t)dt \\
&= -W(x(T),T) + W(x(t_*),t_*) \\
&= W(x_*,t_*).
\end{aligned}$$

2. Let x_0 be a solution of $x'(t) = f(x(t),\varphi(x(t),t))$, $t \in [0,T]$, $x(0) = x_i$. Then we proceed as in

part 1:

$$\begin{aligned}
\int_0^T F(x_0(t),u_0(t),t)dt &= \int_0^T \left[\frac{\partial W}{\partial \xi}(x_0(t),t)f(x_0(t),\varphi(x_0(t),t)) + F(x_0(t),\varphi(x_0(t),t),t)\right]dt - \\
&\quad \int_0^T \frac{\partial W}{\partial \xi}(x_0(t),t)f(x_0(t),\varphi(x_0(t),t))dt \\
&= \int_0^T \min_{u\in\mathbb{R}}\left[\frac{\partial W}{\partial \xi}(x_0(t),t)f(x_0(t),u) + F(x_0(t),u,t)\right]dt - \\
&\quad \int_0^T \frac{\partial W}{\partial \xi}(x_0(t),t)f(x_0(t),\varphi(x_0(t),t))dt \\
&= \int_0^T \left[-\frac{\partial W}{\partial \tau}(x_0(t),t) - \frac{\partial W}{\partial \xi}(x_0(t),t)f(x_0(t),\varphi(x_0(t),t))\right]dt \\
&= -\int_0^T \left(\frac{d}{dt}W(x_0(\cdot),\cdot)\right)(t)dt \\
&= W(x_i,0).
\end{aligned}$$

But from part 1 (with $t_* = 0$), we know that if $u \in C[0,T]$, then

$$I(u) = \int_0^T F(x(t),u(t),t)dt \geq W(x_i,0).$$

This shows that $u_0(\cdot) = \varphi(x_0(\cdot),\cdot)$ is an optimal control and (5.8) holds.

3. We simply repeat the argument from part 2 for the time interval $[t_*,T]$. This yields

$$V(x,t_*) = \min_{u\in C[t_*,T]} \int_{t_*}^T F(x(t),u(t),t)dt = W(x,t_*).$$

■

In the following example, we show how Theorem 5.2.1 can be used to calculate an optimal control.

Example. Consider the function I given by

$$I(u) = \int_0^1 \left[(x(t))^2 + (u(t))^2\right]dt,$$

where x is the solution to

$$x'(t) = u(t), \quad t \in [0,1], \quad x(0) = x_i.$$

Bellman's equation is given by

$$\frac{\partial W}{\partial \tau}(x,t) + \min_{u\in\mathbb{R}}\left[\frac{\partial W}{\partial \xi}(x,t)u + x^2 + u^2\right] = 0, \quad (x,t) \in \mathbb{R}\times(0,1), \quad W(x,1) = 0.$$

It is easy to see that the minimum in the above is assumed for

$$u = \varphi(x,t) = -\frac{1}{2}\frac{\partial W}{\partial \xi}(x,t).$$

Thus we obtain

$$\frac{\partial W}{\partial \tau}(x,t) + x^2 - \frac{1}{4}\left(\frac{\partial W}{\partial \xi}(x,t)\right)^2 = 0, \quad (x,t) \in \mathbb{R}\times(0,1), \quad W(x,1) = 0.$$

This is a nonlinear partial differential equation. It is easy to see that if (x_0, u_0) is an optimal trajectory with initial state x_i, then for every $\lambda \in \mathbb{R}$, $(\lambda x_0, \lambda u_0)$ is an optimal trajectory with respect to the initial state λx_i. Therefore the value function is quadratic in x: $W(\lambda x, t) = \lambda^2 W(x, t)$. In particular,

$$W(x, t) = x^2 W(1, t) = x^2 P(t),$$

where $P(t) := W(1, t)$. Consequently,

$$x^2 P'(t) + x^2 - \frac{1}{4}(2x)^2 (P(t))^2 = 0, \quad (x, t) \in \mathbb{R} \times (0, 1), \quad P(1) = 0.$$

Dividing by x^2, we obtain the Riccati equation

$$P'(t) = (P(t))^2 - 1, \quad t \in (0, 1), \quad P(1) = 0.$$

This has the solution

$$P(t) = \frac{e^{-t+1} - e^{t-1}}{e^{-t+1} + e^{t-1}}.$$

(See the exercise on page 53.) Thus

$$W(x, t) = x^2 \frac{e^{-t+1} - e^{t-1}}{e^{-t+1} + e^{t-1}}.$$

Also the linear time varying differential equation

$$x'(t) = \varphi(x(t), t) = -x(t)P(t), \quad t \in [0, 1], \quad x(0) = x_i$$

has the solution

$$x_0(t) = x_i e^{-\int_0^t P(\tau) d\tau}. \tag{5.9}$$

We note that all the conditions from Theorem 5.2.1 are satisfied, so the optimization problem is solvable. The optimal input is given by

$$u_0(t) = \varphi(x_0(t), t) = -x_0(t)P(t),$$

where x_0 is the optimal state given by (5.9). Note that the optimal control is given in the form of a (time-varying) state feedback. The value function is given by $V(x, t) = x^2 P(t)$. ♢

Exercises.

1. Consider the problem of minimizing the function $I : C[0, 1] \to \mathbb{R}$ defined by

$$I(u) = \int_0^1 [(x(t))^4 + (x(t))^2 (u(t))^2] dt, \tag{5.10}$$

where x denotes the unique solution to

$$x'(t) = u(t), \quad t \in [0, 1], \quad x(0) = 1. \tag{5.11}$$

   (a) Write Bellman's equation associated with the minimization of the function (5.10).

   (b) Let P denote the unique solution to the Riccati equation

$$P'(t) = 4(P(t))^2 - 1, \quad t \in [0, 1], \quad P(1) = 0. \tag{5.12}$$

   Verify that $W : \mathbb{R} \times [0, 1]$ given by $W(x, t) = x^4 P(t)$ for $(x, t) \in \mathbb{R} \times (0, 1)$ satisfies Bellman's equation found in part 1a.

(c) Using Bellman's theorem, conclude that the optimal control u_0 that minimizes I is given by

$$u_0(t) = -2x_0(t)P(t), \quad t \in [0,1],$$

where P is the unique solution to (5.12), and x_0 is the unique solution to

$$x_0'(t) = -2P(t)x_0(t), \quad t \in [0,1].$$

(d) It can be shown that the Riccati equation (5.12) has the solution given by

$$P(t) = \frac{e^{-4t} - e^{-4}}{2(e^{-4t} + e^{-4})}, \quad t \in [0,1].$$

Without calculating the optimal control u_0, determine the value of the corresponding cost $I(u_0)$.
HINT: What is $W(x_0(0), 0)$?

2. Its good to keep in mind that not every optimal control problem is solvable. Prove that for the following problem, there is no minimizing input $u \in C[0,T]$ for the function I given by

$$I(u) = \int_0^1 (x(t))^2 dt,$$

where x is the solution to

$$x'(t) = u(t), \quad t \in [0,1], \quad x(0) = 1.$$

HINT: First show that there exists a sequence of inputs $u_n \in C[0,1]$ such that $I(u_n) \to 0$ as $n \to \infty$. Conclude that if there exists a minimizing control u_0, then $I(u_0) = 0$. But then $x(0)$ cannot be 1.

Bibliography

[1] D.N. Burghes and A. Graham. *Introduction to Control Theory, Including Optimal Control*. John Wiley, 1980.

[2] I.M. Gelfand and S.V. Fomin. *Calculus of Variations*. Dover, 1963.

[3] D.G. Luenberger. *Optimization by Vector Space methods*. Wiley, 1969.

[4] H.J. Sussmann and J.C. Willems. 300 years of optimal control: from the brachystochrone problem to the maximum principle. *IEEE Control Systems*, 17:32-44, 1997.

[5] J.L. Troutman. *Variational Calculus and Optimal Control: Optimization with Elementary Convexity*, 2nd Edition. Springer, 1996.

[6] R. Weinstock. *Calculus of Variations with applications to physics and engineering*. Dover, 1974.

Index