The Maximum Principle of Pontryagin in control and in optimal control

Andrew D. Lewis[1]

16/05/2006
Last updated: 23/05/2006

[1]Associate Professor, Department of Mathematics and Statistics, Queen's University, Kingston, ON K7L 3N6, Canada
Email: andrew@mast.queensu.ca, URL: http://penelope.mast.queensu.ca/~andrew/

Preface

These notes provide an introduction to Pontryagin's Maximum Principle. Optimal control, and in particular the Maximum Principle, is one of the real triumphs of mathematical control theory. Certain of the developments stemming from the Maximum Principle are now a part of the standard tool box of users of control theory. While the Maximum Principle has proved to be extremely valuable in applications, here our emphasis is on understanding the Maximum Principle, where it comes from, and what light it sheds on the subject of control theory in general. For this reason, readers looking for many instances of applications of the Maximum Principle will not find what we say here too rewarding. Such applications can be found in many other places in the literature. What is more difficult to obtain, however, is an appreciation for the Maximum Principle. It is true that the original presentation of Pontryagin, Boltyanskii, Gamkrelidze, and Mishchenko [1986] is still an excellent one in many respects. However, it is probably the case that one can benefit by focusing exclusively on understanding the ideas behind the Maximum Principle, and this is what we try to do here.

Let us outline the structure of our presentation.

1. The Maximum Principle can be thought of as a far reaching generalisation of the classical subject of the calculus of variations. For this reason, we begin our development with a discussion of those parts of the calculus of variations that bear upon the Maximum Principle. The usual Euler–Lagrange equations only paint part of the picture, with the necessary conditions of Legendre and Weierstrass filling in the rest of the canvas. We hope that readers familiar with the calculus of variations can, at the end of this development, at least find the Maximum Principle plausible.
2. After our introduction through the calculus of variations, we give a precise statement of the Maximum Principle.
3. With the Maximum Principle stated, we next wind our way towards its proof. While it is not necessary to understand the proof of the Maximum Principle to use it,[1] a number of important ideas in control theory come up in the proof, and these are explored in independent detail.
   (a) The notion of a "control variation" is fundamental in control theory, particularly in optimal control and controllability theory. It is the common connection with control variations that accounts for the links, at first glance unexpected, between controllability and optimal control.
   (b) The notion of the reachable set lies at the heart of the Maximum Principle. As we shall see in Section 6.2, a significant step in understanding the Maximum Principle occurs with the recognition that optimal trajectories lie on the boundary of the reachable set for the so-called extended system.
4. With the buildup to the proof of the Maximum Principle done, it is possible to complete the proof. In doing so, we identify the key steps and what they rely upon.

[1] Like much of mathematics, one can come to some sort of understanding of the Maximum Principle by applying it enough times to enough interesting problems. However, the understanding one obtains in this way may be compromised by looking only at simple examples. As Bertrand Russell wrote, "The man who has fed the chicken every day throughout its life at last wrings its neck instead, showing that more refined views as to the uniformity of nature would have been useful to the chicken."

5. Since one should, at the end of the day, be able to apply the Maximum Principle, we do this in two special cases: (1) linear quadratic optimal control and (2) linear time-optimal control. In both cases one can arrive at a decent characterisation of the solution of the problem by simply applying the Maximum Principle.
6. In three appendices we collect some details that are needed in the proof of the Maximum Principle. Readers not caring too much about technicalities can probably pretend that these appendices do not exist.

An attempt has been made to keep prerequisites to a minimum. For example, the treatment is not differential geometric. Readers with a background in advanced analysis (at the level of, say, "Baby Rudin" [Rudin 1976]) ought to be able to follow the development. We have also tried to make the treatment as self-contained as possible. The only ingredient in the presentation that is not substantially self-contained is our summary of measure theory and differential equations with measurable time-dependence. It is not really possible, and moreover not very useful, to attempt a significant diversion into these topics. We instead refer the reader to the many excellent references.

Acknowledgements. These notes were prepared for a short graduate course at the Universitat Politècnica de Catalunya in May 2006. I extend warm thanks to Miguel Muñoz–Lecanda for the invitation to give the course and for his hospitality during my stay in Barcelona. I would also like to thank the attendees of the lectures for indulging my excessively pedantic style and onerous notation for approximately three gruelling (for the students, not for me) hours of every day of the course.

Contents

Chapter 1

Control systems and optimal control problems

We begin by indicating what we will mean by a control system in these notes. We will be slightly fussy about classes of trajectories and in this chapter we give the notation attendant to this. We follow this by formulating precisely the problem in optimal control on which we will focus for the remainder of the discussion. Then, by means of motivation, we consider a few typical concrete problems in optimal control.

Notation. Here is a list of standard notation we shall use.

1. $\mathbb{R}$ denotes the real numbers and $\overline{\mathbb{R}} = \{-\infty\} \cup \mathbb{R} \cup \{\infty\}$ denotes the extended reals.
2. The set of linear maps from $\mathbb{R}^m$ to $\mathbb{R}^n$ is denoted by $\mathrm{L}(\mathbb{R}^m; \mathbb{R}^n)$.
3. The standard inner product on $\mathbb{R}^n$ is denoted by $\langle \cdot, \cdot \rangle$ and the standard norm by $\|\cdot\|$. We also use $\|\cdot\|$ for the induced norms on linear and multilinear maps from copies of Euclidean space.
4. For $x \in \mathbb{R}^n$ and $r > 0$ we denote by
$$B(x, r) = \{y \in \mathbb{R}^n \mid \|y - x\| < r\},$$
$$\overline{B}(x, r) = \{y \in \mathbb{R}^n \mid \|y - x\| \leq r\}$$
the open and closed balls of radius r centred at x.
5. We denote by
$$\mathbb{S}^n = \{x \in \mathbb{R}^{n+1} \mid \|x\| = 1\}$$
the ***unit sphere*** of n-dimensions and by
$$\mathbb{D}^n = \{x \in \mathbb{R}^n \mid \|x\| \leq 1\}$$
the ***unit disk*** of n-dimensions.
6. The interior, boundary, and closure of a set $A \subset \mathbb{R}^n$ are denoted by $\mathrm{int}(A)$, $\mathrm{bd}(A)$, and $\mathrm{cl}(A)$, respectively. If $A \subset \mathbb{R}^n$ then recall that the ***relative topology*** on A is that topology whose open sets are of the form $\mathcal{U} \cap A$ with $\mathcal{U} \subset \mathbb{R}^n$ open. If $S \subset A \subset \mathbb{R}^n$, then $\mathrm{int}_A(S)$ is the interior of S with respect to the relative topology on A.

7. If $\mathcal{U} \subset \mathbb{R}^n$ is an open set and if $\phi\colon \mathcal{U} \to \mathbb{R}^m$ is differentiable, the derivative of ϕ at $x \in \mathcal{U}$ is denoted by $\boldsymbol{D}\phi(x)$, and we think of this as a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$. The rth derivative at x we denote by $\boldsymbol{D}^r\phi(x)$ and we think of this as a symmetric multilinear map from $(\mathbb{R}^n)^r$ to $\mathbb{R}^m$.
8. If $\mathcal{U}_a \subset \mathbb{R}^{n_a}$, $a \in \{1, \dots, k\}$, are open sets and if

$$\phi\colon \mathcal{U}_1 \times \cdots \times \mathcal{U}_k \to \mathbb{R}^m$$

is differentiable, we denote by $\boldsymbol{D}_a\phi(x_1, \dots, x_k)$ the ath partial derivative for $a \in \{1, \dots, k\}$. By definition, this is the derivative at x_a of the map from $\mathcal{U}_a$ to $\mathbb{R}^m$ defined by

$$x \mapsto \phi(x_1, \dots, x_{a-1}, x, x_{a+1}, \dots, x_k).$$

We denote by $\boldsymbol{D}_a^r\phi$ the rth partial derivative with respect to the ath component.
9. Let $\mathcal{U} \subset \mathbb{R}^n$ be an open set. A map $\phi\colon \mathcal{U} \to \mathbb{R}^m$ is of ***class C^r*** if it is r-times continuously differentiable.
10. The expression $\dot{f}$ will always mean the derivative of the function $f\colon \mathbb{R} \to \mathbb{R}^k$ with respect to the variable which will be "time" in the problem.
11. We denote by $o(\epsilon^k)$ a general continuous function of ϵ satisfying $\lim_{\epsilon \to 0} \frac{o(\epsilon^k)}{\epsilon^k} = 0$. This is the so-called ***Landau symbol***.
12. The $n \times n$ identity matrix will be denoted by I_n and the $m \times n$ matrix of zeros will be denoted by $0_{m \times n}$.

1.1. Control systems

Control systems come in many flavours, and these flavours do not form a totally ordered set under the relation of increasing generality. Thus one needs to make a choice about the sort of system one is talking about. In these notes we will mean the following.

1.1 Definition: (Control system) A ***control system*** is a triple $\Sigma = (\mathcal{X}, f, U)$ where

(i) $\mathcal{X} \subset \mathbb{R}^n$ is an open set,

(ii) $U \subset \mathbb{R}^m$, and

(iii) $f\colon \mathcal{X} \times \mathrm{cl}(U) \to \mathbb{R}^n$ has the following properties:

  (a) f is continuous;

  (b) the map $x \mapsto f(x, u)$ is of class C^1 for each $u \in \mathrm{cl}(U)$. •

The differential equation associated to a control system $\Sigma = (\mathcal{X}, f, U)$ is

$$\dot{\xi}(t) = f(\xi(t), \mu(t)). \tag{1.1}$$

One might additionally consider a system where time-dependence enters explicitly, rather than simply through the control. However, we do not pursue this level of generality (or any of the other possible levels of generality) since it contributes little to the essential ideas we develop.

Let us give important specific classes of control systems.

1.2 Example: (Control-affine system) A control system $\Sigma = (\mathcal{X}, f, U)$ is a ***control-affine system*** if f has the form

$$f(x, u) = f_0(x) + f_1(x) \cdot u,$$

where $f\colon \mathcal{X} \to \mathbb{R}^n$ and $f_1\colon \mathcal{X} \to \mathrm{L}(\mathbb{R}^m; \mathbb{R}^n)$ are of class C^1. Note that f is an affine function of u, hence the name. Many of the systems one encounters in practice are control-affine systems. •

A special class of control-affine system is the following.

1.3 Example: (Linear control system) A ***linear control system*** (more precisely, a ***linear time-invariant control system***) is a triple (A, B, U) where $A\colon \mathbb{R}^n \to \mathbb{R}^n$ and $B\colon \mathbb{R}^m \to \mathbb{R}^n$ are linear maps, and where $U \subset \mathbb{R}^m$. We associate to a linear control system the control system $(\mathcal{X}, f, U)$ with $\mathcal{X} = \mathbb{R}^n$, $f(x, u) = A(x) + B(u)$, and with "$U = U$." Thus the equations governing a linear control system are

$$\dot{\xi}(t) = A(\xi(t)) + B(\mu(t)).$$

The solution to this differential equation with initial condition $\xi(0)$ at time $t = 0$ is given by

$$\xi(t) = \exp(At)\xi(0) + \int_0^t \exp(A(t - \tau))B\mu(\tau)\,\mathrm{d}\tau, \tag{1.2}$$

where $\exp(\cdot)$ is the matrix exponential. •

1.2. Controls and trajectories

It will be worthwhile for us to be quite careful about characterising the sorts of controls we will consider, and the trajectories generated by them. A consequence of this care is a pile of notation.

We should first place some conditions on the character of the control functions $t \mapsto \mu(t)$ that will allow solutions to (1.1). The following definition encodes one of the weakest notion of control that one can allow.

1.4 Definition: (Admissible control, controlled trajectory, controlled arc) Let $\Sigma = (\mathcal{X}, f, U)$ be a control system.

(i) An ***admissible control*** is a measurable map $\mu\colon I \to U$ defined on an interval $I \subset \mathbb{R}$ such that $t \mapsto f(x, \mu(t))$ is locally integrable for each $x \in \mathcal{X}$. The set of admissible controls defined on I is denoted by $\mathscr{U}(I)$.

(ii) A ***controlled trajectory*** is a pair (ξ, μ) where, for some interval $I \subset \mathbb{R}$,

  (a) $\mu \in \mathscr{U}(I)$ and

  (b) $\xi\colon I \to \mathcal{X}$ satisfies (1.1).

  We call I the ***time interval*** for (ξ, μ).

(iii) A ***controlled arc*** is a controlled trajectory defined on a compact time interval.

If (ξ, μ) is a controlled trajectory, we call ξ the ***trajectory*** and μ the ***control***. •

For $x_0 \in \mathcal{X}$ and $t_0 \in I$ we denote by $t \mapsto \xi(\mu, x_0, t_0, t)$ the solution of the differential equation (1.1) satisfying $\xi(\mu, x_0, t_0, t_0) = x_0$. We denote by $\xi(\mu, x_0, t_0, \cdot)$ the map $t \mapsto \xi(\mu, x_0, t_0, t)$.

Corresponding to admissible controls we shall denote

$$\begin{aligned}
\text{Ctraj}(\Sigma) &= \{(\xi, \mu) \mid (\xi, \mu) \text{ is a controlled trajectory}\},\\
\text{Carc}(\Sigma) &= \{(\xi, \mu) \mid (\xi, \mu) \text{ is a controlled arc}\},\\
\text{Ctraj}(\Sigma, I) &= \{(\xi, \mu) \mid (\xi, \mu) \text{ is a controlled trajectory with time interval } I\},\\
\text{Carc}(\Sigma, I) &= \{(\xi, \mu) \mid (\xi, \mu) \text{ is a controlled arc with time interval } I\}.
\end{aligned}$$

Because of the rather general nature of the controls we allow, the existence and uniqueness of controlled trajectories does not quite follow from the basic such theorems for ordinary differential equations. We consider such matters in Appendix A.

It is also essential to sometimes restrict controls to not be merely integrable, but bounded.[1] To encode this in notation, we denote by $\mathscr{U}_{\text{bdd}}(I)$ the set of admissible controls defined on the interval $I \subset \mathbb{R}$ that are also bounded.

Sometimes we shall merely wish to consider trajectories emanating from a specified initial condition and whose existence can be guaranteed for some duration of time. This leads to the following notion.

1.5 Definition: (Controls giving rise to trajectories defined on a specified interval) Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, let $I \subset \mathbb{R}$ be an interval, and let $t_0 \in I$. We denote by $\mathscr{U}(x_0, t_0, I)$ the set of admissible controls such that the solution to the initial value problem

$$\dot{\xi}(t) = f(\xi(t), \mu(t)), \quad \xi(t_0) = x_0,$$

exists for all $t \in I$. •

1.3. Two general problems in optimal control

Now let us introduce into the problem the notion of optimisation. We will focus on a rather specific sort of problem.

1.6 Definition: (Lagrangian, objective function) Let $\Sigma = (\mathcal{X}, f, U)$ be a control system.

(i) A ***Lagrangian*** for Σ is a map $L\colon \mathcal{X} \times \text{cl}(U) \to \mathbb{R}$ for which
   (a) L is continuous and
   (b) the function $x \mapsto L(x, u)$ is of class C^1 for each $u \in \text{cl}(U)$.

(ii) If L is a Lagrangian, $(\xi, \mu) \in \text{Ctraj}(\Sigma)$ is ***L-acceptable*** if the function $t \mapsto L(\xi(t), \mu(t))$ is integrable, where I is the time interval for (ξ, μ). The set of L-acceptable controlled trajectories (resp. controlled arcs) for Σ is denoted by $\text{Ctraj}(\Sigma, L)$ (resp. $\text{Carc}(\Sigma, L)$).

[1] More generally, one can consider controls that are essentially bounded. However, since trajectories for the class of bounded controls and the class of unbounded controls are identical by virtue of these controls differing only on sets of measure zero, there is nothing gained by carrying around the word "essential."

(iii) If L is a Lagrangian, the corresponding ***objective function*** is the map $J_{\Sigma,L}\colon \mathrm{Ctraj}(\Sigma) \to \overline{\mathbb{R}}$ defined by

$$J_{\Sigma,L}(\xi,\mu) = \int_I L(\xi(t),\mu(t))\,\mathrm{d}t,$$

where we adopt the convention that $J_{\Sigma,L}(\xi,\mu) = \infty$ if (ξ,μ) is not L-acceptable. •

We will seek, essentially, to minimise the objective function over some set of controlled trajectories. It is interesting and standard to consider controlled trajectories that steer the system from some subset S_0 of $\mathcal{X}$ to some other subset S_1 of $\mathcal{X}$. Let us define precisely the problems we will address in these notes. Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let L be a Lagrangian, and let S_0 and S_1 be subsets of $\mathcal{X}$. Denote by $\mathrm{Carc}(\Sigma, L, S_0, S_1) \subset \mathrm{Carc}(\Sigma)$ the set of controlled arcs with the following properties:

1. if $(\xi,\mu) \in \mathrm{Carc}(\Sigma, L, S_0, S_1)$ then (ξ,μ) is defined on a time interval of the form $[t_0, t_1]$ for some $t_0, t_1 \in \mathbb{R}$ satisfying $t_0 < t_1$;
2. if $(\xi,\mu) \in \mathrm{Carc}(\Sigma, L, S_0, S_1)$ then $(\xi,\mu) \in \mathrm{Carc}(\Sigma, L)$;
3. if $(\xi,\mu) \in \mathrm{Carc}(\Sigma, L, S_0, S_1)$ is defined on the time interval $[t_0, t_1]$, then $\xi(t_0) \in S_0$ and $\xi(t_1) \in S_1$.

This leads to the following problem.

1.7 Problem: (Free interval optimal control problem) Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let L be a Lagrangian for Σ, and let $S_0, S_1 \subset \mathcal{X}$ be sets. A controlled trajectory $(\xi_*, \mu_*) \in \mathrm{Carc}(\Sigma, L, S_0, S_1)$ is a solution to the ***free interval optimal control problem*** for Σ, L, S_0, and S_1 if $J_{\Sigma,L}(\xi_*,\mu_*) \leq J_{\Sigma,L}(\xi,\mu)$ for each $(\xi,\mu) \in \mathrm{Carc}(\Sigma, L, S_0, S_1)$. The set of solutions to this problem is denoted by $\mathscr{P}(\Sigma, L, S_0, S_1)$. •

For $t_0, t_1 \in \mathbb{R}$ satisfying $t_0 < t_1$, we denote by $\mathrm{Carc}(\Sigma, L, S_0, S_1, [t_0, t_1])$ the subset of $\mathrm{Carc}(\Sigma, L, S_0, S_1)$ comprised of those controlled arcs defined on $[t_0, t_1]$. This gives the following problem.

1.8 Problem: (Fixed interval optimal control problem) Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let L be a Lagrangian for Σ, let $S_0, S_1 \subset \mathcal{X}$ be sets, and let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$. A controlled trajectory $(\xi_*, \mu_*) \in \mathrm{Carc}(\Sigma, L, S_0, S_1, [t_0, t_1])$ is a solution to the ***fixed interval optimal control problem*** for Σ, L, S_0, and S_1 if $J_{\Sigma,L}(\xi_*,\mu_*) \leq J_{\Sigma,L}(\xi,\mu)$ for each $(\xi,\mu) \in \mathrm{Carc}(\Sigma, L, S_0, S_1, [t_0, t_1])$. The set of solutions to this problem is denoted by $\mathscr{P}(\Sigma, L, S_0, S_1, [t_0, t_1])$. •

Giving necessary conditions for the solution of these optimal control problems is what the Maximum Principle is concerned with. By providing such necessary conditions, the Maximum Principle does not *solve* the problem, but it does often provide enough restrictions on the set of possible solutions that insight can be gained to allow a solution to the problem. We shall see instances of this in Chapters 8 and 9.

1.4. Some examples of optimal control problems

Let us now take the general discussion of the preceding section and consider some special case which themselves have varying levels of generality.

1. In the case where the Lagrangian is defined by $L(x,u) = 1$, the objective function is then the time along the trajectory, and solutions of the problem $\mathscr{P}(\Sigma, L, S_0, S_1)$ are called ***time-optimal trajectories*** for Σ.
2. A commonly encountered optimal control problem arises in the study of linear systems. Let Σ be the control system associated with a linear system (A, B, U), let Q be a symmetric bilinear form on the state space $\mathbb{R}^n$, and let R be a symmetric positive-definite bilinear form on the control space $\mathbb{R}^m$. We then consider the Lagrangian defined by $L_{Q,R}(x,u) = \frac{1}{2}Q(x,x) + \frac{1}{2}R(u,u)$. This class of optimal control problems are called ***linear quadratic optimal control problems***. The utility of this class of problem is not readily apparent at first blush. However, this optimal control problem leads, in a way that we will spell out, to a technique for designing a stabilising linear state feedback for (A, B, U). We shall consider this in Chapter 8.
3. Let $f_1, \dots, f_m$ be vector fields on the open subset $\mathcal{X}$ of $\mathbb{R}^n$ and define

$$f(x,u) = \sum_{a=1}^{m} u^a f_a(x).$$

   This sort of system is called a ***driftless control system***. Let us take as Lagrangian the function defined by $L(x,u) = \frac{1}{2}\|u\|$. This sort of optimal control problem is a sort of clumsy first step towards ***sub-Riemannian geometry***. A more elegant description of sub-Riemannian geometry involves distributions.
4. As a very concrete example we take
   (a) $n = 2$ and $m = 1$,
   (b) $\mathcal{X} = \mathbb{R}^2$,
   (c) $f(x,u) = (x^2, u)$,
   (d) $U = [-1, 1]$,
   (e) $L((x^1, x^2), u) = \frac{1}{2}(x^1)^2$, and
   (f) $S_0 = \{(x_0^1, x_0^2)\}$ and $S_1 = \{(0,0)\}$.

   The fixed interval optimal control problem associated with this data has a surprisingly complicated solution exhibiting what is known as ***Fuller's phenomenon***. What happens is that, as one approaches S_1, the optimal control undergoes an infinite number of switches between the boundary points of the control set U. This indicates that simple problems in optimal control can have complicated, even undesirable solutions. We refer the reader to Exercise E7.3 to see an outline of what one can say in this example after an application of the Maximum Principle.

Exercises

E1.1 Prove (1.2).

E1.2 Do the following.

(a) Find a control system $\Sigma = (\mathcal{X}, f, U)$ and a locally integrable function $\mu\colon [0,1] \to U$ that is not an admissible control.

(b) Find a control system $\Sigma = (\mathcal{X}, f, U)$ and an admissible control $\mu\colon [0,1] \to U$ that is not locally integrable.

Chapter 2

From the calculus of variations to optimal control

The calculus of variations is a subject with a distinguished history in mathematics. The subject as we know it began in earnest with the so-called "brachistochrone problem," the object of which is to determine the path along which a particle must fall in a gravitational field in order to minimise the time taken to reach a desired point (see Figure 2.1). This

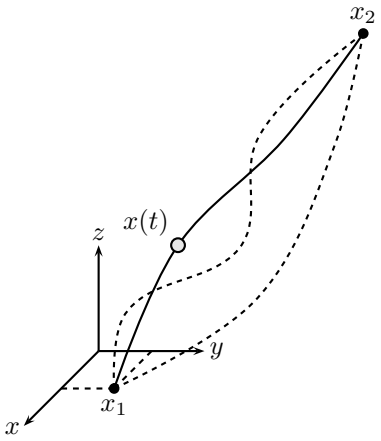

Figure 2.1. The brachistochrone problem

problem was posed in 1696 by Johann Bernoulli to the mathematical community, with solutions being given by some of the luminaries of the time: Johann himself, Johann's brother Jakob, Leibniz, Newton, and Tschirnhaus. This problem is typical of the calculus of variations in that its solution is a curve. We refer to [Goldstine 1980] for an historical account of the calculus of variations.

In this chapter we will review some parts of the classical calculus of variations with an eye towards motivating the Maximum Principle. The Maximum Principle itself can seem to be a bit of a mystery, so we hope that this warmup via the calculus of variations will be helpful in the process of demystification. The presentation here owes a great deal to the very interesting review article of Sussmann and Willems [1997].

There are a huge number of books dealing with the calculus of variations. Introductory books include [Gelfand and Fomin 2000, Lanczos 1949, Yan 1995]. More advanced treatments include [Bliss 1946, Bolza 1961, Carathéodory 1935, Troutman 1996], and the two volumes of Giaquinta and Hildebrandt [1996].

2.1. Three necessary conditions in the calculus of variations

We let $\mathcal{X}$ be an open subset of $\mathbb{R}^n$ and let $L\colon \mathcal{X} \times \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function which we call the ***Lagrangian***.[1] We let $x_0, x_1 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and denote by $C^2(x_0, x_1, [t_0, t_1])$ the collection of twice continuously differentiable curves $\xi\colon [t_0, t_1] \to \mathcal{X}$ which satisfy $\xi(t_0) = x_0$ and $\xi(t_1) = x_1$. For $\xi \in C^2(x_0, x_1, [t_0, t_1])$ we denote

$$J_L(\xi) = \int_{t_0}^{t_1} L(\xi(t), \dot{\xi}(t))\, \mathrm{d}t.$$

The basic problem in the calculus of variations is then the following.

2.1 Problem: (Problem of calculus of variations) Find $\xi_* \in C^2(x_0, x_1, [t_0, t_1])$ such that $J_L(\xi_*) \leq J_L(\xi)$ for every $\xi \in C^2(x_0, x_1, [t_0, t_1])$. The set of solutions to this problem will be denoted by $\mathscr{P}(L, x_0, x_1, [t_0, t_1])$. •

We will state and prove three necessary conditions which must be satisfied by elements of $\mathscr{P}(L, x_0, x_1, [t_0, t_1])$.

2.1.1. The necessary condition of Euler–Lagrange. The first necessary condition is the most important one since it gives a differential equation that must be satisfied by solutions of the basic problem in the calculus of variations. We denote a point in $\mathcal{X} \times \mathbb{R}^n$ by (x, v), and we think of x as representing "position" and v as representing "velocity."

2.2 Theorem: (The necessary condition of Euler–Lagrange) *Let $\mathcal{X}$ be an open subset of $\mathbb{R}^n$, let $x_0, x_1 \in \mathcal{X}$, and let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$. Suppose that $L\colon \mathcal{X} \times \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable Lagrangian and let $\xi \in \mathscr{P}(L, x_0, x_1, [t_0, t_1])$. Then ξ satisfies the* ***Euler–Lagrange equations****:*

$$\frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{D}_2 L(\xi(t), \dot{\xi}(t))) - \boldsymbol{D}_1 L(\xi(t), \dot{\xi}(t)) = 0.$$

Proof: Let $\xi \in C^2(x_0, x_1, [t_0, t_1])$ and let $\zeta\colon [t_0, t_1] \to \mathbb{R}^n$ be twice continuously differentiable and satisfy $\zeta(t_0) = \zeta(t_1) = 0$. Then, for $\epsilon > 0$ sufficiently small, the map $\xi_\zeta\colon (-\epsilon, \epsilon) \times [t_0, t_1] \to \mathcal{X}$ defined by $\xi_\zeta(s, t) = \xi(t) + s\zeta(t)$ has the following properties:

1. it makes sense (i.e., takes values in $\mathcal{X}$);
2. $\xi_\zeta(0, t) = \xi(t)$ for each $t \in [t_0, t_1]$;
3. $\frac{\mathrm{d}}{\mathrm{d}s}\big|_{s=0} \xi_\zeta(s, t) = \zeta(t)$ for each $t \in [t_0, t_1]$.

[1] Note that we consider Lagrangians of class C^2 in this section. This is because we want to be able to always write the Euler–Lagrange equations. In the problems of optimal control we will relax the smoothness on Lagrangians, cf. Definition 1.6.

For $s \in (-\epsilon, \epsilon)$ let us denote by $\xi^s_\zeta \in C^2(x_0, x_1, [t_0, t_1])$ the curve defined by $\xi^s_\zeta(t) = \xi_\zeta(s, t)$. Note that the function $s \mapsto J_L(\xi^s_\zeta)$ is continuously differentiable. We then compute

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}s} J_L(\xi^s_\zeta) &= \int_{t_0}^{t_1} \frac{\mathrm{d}}{\mathrm{d}s} L(\xi^s_\zeta(t), \dot{\xi}^s_\zeta(t)) \, \mathrm{d}t \\
&= \int_{t_0}^{t_1} \Big(\boldsymbol{D}_1 L(\xi^s_\zeta(t), \dot{\xi}^s_\zeta(t)) \cdot \zeta(t) + \boldsymbol{D}_2 L(\xi^s_\zeta(t), \dot{\xi}^s_\zeta(t)) \cdot \dot{\zeta}(t) \Big) \, \mathrm{d}t \qquad (2.1) \\
&= \int_{t_0}^{t_1} \Big(\boldsymbol{D}_1 L(\xi^s_\zeta(t), \dot{\xi}^s_\zeta(t)) - \frac{\mathrm{d}}{\mathrm{d}t} (\boldsymbol{D}_2 L(\xi^s_\zeta(t), \dot{\xi}^s_\zeta(t))) \Big) \cdot \zeta(t) \, \mathrm{d}t \\
&\quad + \boldsymbol{D}_2 L(\xi^s_\zeta(t), \dot{\xi}^s_\zeta(t)) \cdot \zeta(t) \Big|_{t=t_0}^{t=t_1} \\
&= \int_{t_0}^{t_1} \Big(\boldsymbol{D}_1 L(\xi^s_\zeta(t), \dot{\xi}^s_\zeta(t)) - \frac{\mathrm{d}}{\mathrm{d}t} (\boldsymbol{D}_2 L(\xi^s_\zeta(t), \dot{\xi}^s_\zeta(t))) \Big) \cdot \zeta(t) \, \mathrm{d}t,
\end{aligned}$$

where we have used integration by parts in the penultimate step and the fact that $\zeta(t_0) = \zeta(t_1) = 0$ in the last step.

Now suppose that for some $\bar{t} \in [t_0, t_1]$ we have

$$\boldsymbol{D}_1 L(\xi(\bar{t}), \dot{\xi}(\bar{t})) - \frac{\mathrm{d}}{\mathrm{d}t} (\boldsymbol{D}_2 L(\xi(\bar{t}), \dot{\xi}(\bar{t}))) \neq 0.$$

Then there exists $\zeta_0 \in \mathbb{R}^n$ such that

$$\Big(\boldsymbol{D}_1 L(\xi(\bar{t}), \dot{\xi}(\bar{t})) - \frac{\mathrm{d}}{\mathrm{d}t} (\boldsymbol{D}_2 L(\xi(\bar{t}), \dot{\xi}(\bar{t}))) \Big) \cdot \zeta_0 > 0.$$

Then, since L is twice continuously differentiable, there exists $\delta > 0$ such that

$$\Big(\boldsymbol{D}_1 L(\xi(t), \dot{\xi}(t)) - \frac{\mathrm{d}}{\mathrm{d}t} (\boldsymbol{D}_2 L(\xi(t), \dot{\xi}(t))) \Big) \cdot \zeta_0 > 0$$

for $t \in [\bar{t} - \delta, \bar{t} + \delta] \cap [t_0, t_1]$. We may, therefore, suppose without loss of generality that $\bar{t}$ and δ satisfy $\bar{t} \in (t_0, t_1)$ and $[\bar{t} - \delta, \bar{t} + \delta] \subset [t_0, t_1]$.

Now let $\phi \colon [t_0, t_1] \to \mathbb{R}$ be of class C^2 and be such that $\phi(t) > 0$ for $|t - \bar{t}| < \delta$ and $\phi(t) = 0$ for $|t - \bar{t}| \geq \delta$ (can you think of such a function?). Then define $\zeta(t) = \phi(t)\zeta_0$. This then gives

$$\frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} J_L(\xi^s_\zeta) > 0.$$

Therefore, it cannot hold that the function $s \mapsto J_L(\xi^s_\zeta)$ has a minimum at $s = 0$. This shows that if the Euler–Lagrange equations do not hold, then $\xi \notin \mathscr{P}(L, x_0, x_1, [t_0, t_1])$. ■

The way to think of the Euler–Lagrange equations is as analogous to the condition that the derivative of a function at a minimum must be zero. Thus these equations are a "first-order" necessary condition. Note that solutions to the Euler–Lagrange equations are *not* generally elements of $\mathscr{P}(L, x_0, x_1, [t_0, t_1])$. However, such solutions are important and so are given a name.

2.3 Definition: (Extremal in the calculus of variations) Let $\mathcal{X}$ be an open subset of $\mathbb{R}^n$, let $x_0, x_1 \in \mathcal{X}$, and let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$. Suppose that $L \colon \mathcal{X} \times \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable Lagrangian. A solution $\xi \colon [t_0, t_1] \to \mathcal{X}$ to the Euler–Lagrange equations is an ***extremal*** for $\mathscr{P}(L, x_0, x_1, [t_0, t_1])$. •

2.1.2. The necessary condition of Legendre. Next we look at second-order conditions which are analogous to the condition that, at a minimum, the Hessian of a function must be positive-semidefinite.

2.4 Theorem: (The necessary condition of Legendre) *Let $\mathcal{X}$ be an open subset of $\mathbb{R}^n$, let $x_0, x_1 \in \mathcal{X}$, and let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$. Suppose that $L\colon \mathcal{X} \times \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable Lagrangian and let $\xi \in \mathscr{P}(L, x_0, x_1, [t_0, t_1])$. Then the symmetric bilinear map*

$$\boldsymbol{D}_2^2 L(\xi(t), \dot{\xi}(t))$$

is positive-semidefinite for each $t \in [t_0, t_1]$.

Proof: Let $\xi \in C^2(x_0, x_1, [t_0, t_1])$ and let $\zeta\colon [t_0, t_1] \to \mathbb{R}^n$ be as in the proof of Theorem 2.2. Let ξ_ζ^s be the curve as defined in the proof of Theorem 2.2. Since L is of class C^2 the function $s \mapsto J_L(\xi_\zeta^s)$ is also of class C^2. Moreover, using (2.1) as a starting point,

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}s^2} J_L(\xi_\zeta^s) = \int_{t_0}^{t_1} \Big(&\boldsymbol{D}_1^2 L(\xi_\zeta^s(t), \dot{\xi}_\zeta^s(t)) \cdot (\zeta(t), \zeta(t)) \\ &+ 2\boldsymbol{D}_1 \boldsymbol{D}_2 L(\xi_\zeta^s(t), \dot{\xi}_\zeta^s(t)) \cdot (\zeta(t), \dot{\zeta}(t)) + \boldsymbol{D}_2^2 L(\xi_\zeta^s(t), \dot{\xi}_\zeta^s(t)) \cdot (\dot{\zeta}(t), \dot{\zeta}(t)) \Big) \, \mathrm{d}t.\end{aligned}$$

Let $C_{1,\xi}$ and $C_{2,\xi}$ be defined by

$$\begin{aligned} C_{1,\xi} &= \sup \{ \|\boldsymbol{D}_1^2 L(\xi(t), \dot{\xi}(t))\| \mid t \in [t_0, t_1] \}, \\ C_{2,\xi} &= \sup \{ 2\|\boldsymbol{D}_1 \boldsymbol{D}_2 L(\xi(t), \dot{\xi}(t))\| \mid t \in [t_0, t_1] \}. \end{aligned}$$

Suppose that there exists $\bar{t} \in [t_0, t_1]$ such that $\boldsymbol{D}_2^2 L(\xi(\bar{t}), \dot{\xi}(\bar{t}))$ is not positive-semidefinite. There then exists a vector $\zeta_0 \in \mathbb{R}^n$ such that

$$\boldsymbol{D}_2^2 L(\xi(\bar{t}), \dot{\xi}(\bar{t})) \cdot (\zeta_0, \zeta_0) < 0.$$

Then there exists $\delta_0 > 0$ and $C_{3,\xi} > 0$ such that

$$\boldsymbol{D}_2^2 L(\xi(t), \dot{\xi}(t)) \cdot (\zeta_0, \zeta_0) \leq -C_{3,\xi} < 0$$

for all $t \in [\bar{t} - \delta_0, \bar{t} + \delta_0] \cap [t_0, t_1]$. We may thus suppose without loss of generality that $\bar{t}$ and δ_0 satisfy $\bar{t} \in (t_0, t_1)$ and $[\bar{t} - \delta_0, \bar{t} + \delta_0] \subset [t_0, t_1]$. We also suppose, again without loss of generality, that $\|\zeta_0\| = 1$.

Let us define $\ddot{\phi}\colon \mathbb{R} \to \mathbb{R}$ be defined by

$$\ddot{\phi}(t) = \begin{cases} 0, & |t| \geq 1, \\ 32(t+1), & t \in (-1, -\frac{3}{4}), \\ -32(t+\frac{1}{2}), & t \in [-\frac{3}{4}, -\frac{1}{4}), \\ 32t, & t \in [-\frac{1}{4}, 0), \\ -32t, & t \in [0, \frac{1}{4}), \\ 32(t-\frac{1}{2}), & t \in [\frac{1}{4}, \frac{3}{4}), \\ 32(-t+1), & t \in (\frac{3}{4}, 1). \end{cases}$$

Then define

$$\dot\phi(t) = \int_{-\infty}^{t} \ddot\phi(\tau)\,\mathrm{d}\tau, \quad \phi(t) = \int_{-\infty}^{t} \dot\phi(\tau)\,\mathrm{d}\tau.$$

The function ϕ is designed to have the following properties:

1. ϕ is of class C^2;
2. $\phi(-t) = \phi(t)$;
3. $\phi(t) = 0$ for $|t| \geq 1$;
4. $0 \leq \phi(t) \leq 1$ for $t \in [-1, 1]$;
5. $\dot\phi(t) > 0$ for $t \in (-1, 0)$;
6. $\dot\phi$ is monotonically increasing on $(-1, -\frac{1}{2})$;
7. $\dot\phi$ is monotonically decreasing on $(-\frac{1}{2}, 0)$.

Define

$$d = \inf\{\dot\phi(t) \mid t \in [-\tfrac{9}{10}, -\tfrac{1}{10}]\}, \quad D = \sup\{\dot\phi(t) \mid t \in [-\tfrac{9}{10}, -\tfrac{1}{10}]\}.$$

The properties of ϕ ensure that $|\dot\phi(t)| < d$ for $t \in (-1, -\frac{9}{10}) \cup (-\frac{1}{10}, \frac{1}{10}) \cup (\frac{9}{10}, 1)$. These features of ϕ ensure that the following estimates hold:

$$\int_{-1}^{1} |\phi(t)|^2\,\mathrm{d}t \leq 1, \quad \int_{-1}^{1} |\phi(t)\dot\phi(t)|\,\mathrm{d}t \leq D, \quad \int_{-1}^{1} |\dot\phi(t)|^2\,\mathrm{d}t \geq \frac{8d}{5}.$$

Now, for $\delta \in (0, \delta_0)$ define $\phi_\delta\colon [t_0, t_1] \to \mathbb{R}$ by $\phi_\delta(t) = \phi(\delta^{-1}(t - \bar t))$. Our estimates for ϕ above then translate to

$$\int_{t_0}^{t_1} |\phi_\delta(t)|^2\,\mathrm{d}t \leq \delta, \quad \int_{t_0}^{t_1} |\phi_\delta(t)\dot\phi_\delta(t)|\,\mathrm{d}t \leq D, \quad \int_{t_0}^{t_1} |\dot\phi_\delta(t)|^2\,\mathrm{d}t \geq \frac{8d}{5\delta}.$$

Now define $\zeta_\delta\colon [t_0, t_1] \to \mathbb{R}^n$ by $\zeta_\delta(t) = \phi_\delta(t)\zeta_0$. With ζ_δ so defined we have

$$\Big|\int_{t_0}^{t_1} \boldsymbol{D}_1^2 L(\xi(t), \dot\xi(t)) \cdot (\zeta_\delta(t), \zeta_\delta(t))\,\mathrm{d}t\Big| \leq C_{1,\xi} \int_{t_0}^{t_1} |\phi_\delta(t)|^2\,\mathrm{d}t \leq \delta C_{1,\xi}$$

and

$$\Big|\int_{t_0}^{t_1} 2\boldsymbol{D}_1\boldsymbol{D}_2 L(\xi(t), \dot\xi(t)) \cdot (\zeta_\delta(t), \dot\zeta_\delta(t))\,\mathrm{d}t\Big| \leq C_{2,\xi} \int_{t_0}^{t_1} |\phi_\delta(t)\dot\phi_\delta(t)|\,\mathrm{d}t \leq DC_{2,\xi}.$$

We also have

$$\int_{t_0}^{t_1} \boldsymbol{D}_2^2 L(\xi(t), \dot\xi(t)) \cdot (\dot\zeta_\delta(t), \dot\zeta_\delta(t))\,\mathrm{d}t \leq -C_{3,\xi} \int_{t_0}^{t_1} |\dot\phi_\delta(t)|^2\,\mathrm{d}t \leq -\frac{8dC_{3,\xi}}{5\delta}.$$

From this we ascertain that for δ sufficiently small we have

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2} J_L(\xi_{\zeta_\delta^s})\Big|_{s=0} < 0.$$

If $\xi \in \mathscr{P}(L, x_0, x_1, [t_0, t_1])$ then we must have

$$\frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} J_L(\xi_{\zeta_\delta^s}) = 0, \quad \frac{\mathrm{d}^2}{\mathrm{d}s^2}\Big|_{s=0} J_L(\xi_{\zeta_\delta^s}) \geq 0.$$

Thus we have shown that, if $\boldsymbol{D}_2^2 L(\xi(t), \dot{\xi}(t))$ is not positive-semidefinite for every $t \in [t_0, t_1]$, then $\xi \notin \mathscr{P}(L, x_0, x_1, [t_0, t_1])$, as desired. ■

2.5 Remark: (Relationship with finding minima in finite dimensions) Theorems 2.2 and 2.4 are exactly analogous to well-known conditions from multivariable calculus. This is fairly easily seen by following the proofs of the theorems. But let us make this connection explicit in any case.

Let $\mathcal{U} \subset \mathbb{R}^n$ be an open set and let $f \colon \mathcal{U} \to \mathbb{R}$ be a function of class C^2. In calculus one shows that if $x_0 \in \mathcal{U}$ is a local minimum of f then

1. $\boldsymbol{D}f(x_0) = 0$ and
2. $\boldsymbol{D}^2 f(x_0)$ is positive-semidefinite.

Theorem 2.2 is exactly analogous to the first of these conditions and Theorem 2.4 is exactly analogous to the second of these conditions. •

2.1.3. The necessary condition of Weierstrass. The last necessary condition we give requires a new object.

2.6 Definition: (Weierstrass excess function) Let $\mathcal{X} \subset \mathbb{R}^n$ be open and let $L \colon \mathcal{X} \times \mathbb{R}^n \to \mathbb{R}$ be a Lagrangian. The ***Weierstrass excess function*** is the function $E_L \colon \mathcal{X} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by

$$E_L(x, v, u) = L(x, u) - L(x, v) - \boldsymbol{D}_2 L(x, v) \cdot (u - v). \qquad \bullet$$

The excess function appears in the following necessary condition for a solution of the problem in the calculus of variations.

2.7 Theorem: (The necessary condition of Weierstrass) *Let $\mathcal{X}$ be an open subset of $\mathbb{R}^n$, let $x_0, x_1 \in \mathcal{X}$, and let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$. Suppose that $L \colon \mathcal{X} \times \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable Lagrangian and let $\xi \in \mathscr{P}(L, x_0, x_1, [t_0, t_1])$. Then*

$$E_L(\xi(t), \dot{\xi}(t), u) \geq 0$$

for all $t \in [t_0, t_1]$ and for all $u \in \mathbb{R}^n$.

Proof: Suppose that $\xi \in \mathscr{P}(L, x_0, x_1, [t_0, t_1])$ and that

$$E_L(\xi(\bar{t}), \dot{\xi}(\bar{t}), u) < 0$$

for some $\bar{t} \in [t_0, t_1]$ and for some $u \in \mathbb{R}^n$. Then, since E_L is continuous,

$$E_L(\xi(t), \dot{\xi}(t), u) < 0$$

for all t sufficiently close to $\bar{t}$. We therefore suppose, without loss of generality, that $\bar{t} \in (t_0, t_1)$.

Choose $\epsilon_0 > 0$ sufficiently small that $[\bar{t} - \epsilon_0, \bar{t} + \epsilon_0] \subset [t_0, t_1]$. Now, for $\epsilon \in (0, \epsilon_0)$, define

$$\xi_\epsilon(t) = \begin{cases} \xi(t) + (t - t_0)u_\epsilon, & t \in [t_0, \bar{t} - \epsilon), \\ (t - \bar{t})u + \xi(\bar{t}), & t \in [\bar{t} - \epsilon, \bar{t}), \\ \xi(t), & t \in [\bar{t}, t_1], \end{cases}$$

where

$$u_\epsilon = \frac{\xi(\bar{t}) - \xi(\bar{t} - \epsilon) - \epsilon u}{\bar{t} - t_0 - \epsilon}.$$

Note that u_ϵ has been defined so that $t \mapsto \xi_\epsilon(t)$ is continuous. Note that

$$\dot{\xi}_\epsilon(t) = \begin{cases} \dot{\xi}(t) + u_\epsilon, & t \in (t_0, \bar{t} - \epsilon), \\ u, & t \in (\bar{t} - \epsilon, \bar{t}), \\ \dot{\xi}(t), & t \in (\bar{t}, t_1). \end{cases}$$

Now define

$$\Delta_L(\epsilon) = J_L(\xi_\epsilon) - J_L(\xi)$$

Since $\lim_{\epsilon \to 0} u_\epsilon = 0$,

$$\begin{aligned} \lim_{\epsilon \to 0} \Delta_L(\epsilon) = & \lim_{\epsilon \to 0} \int_{t_0}^{\bar{t} - \epsilon} \left(L(\xi(t) + (t - t_0)u_\epsilon, \dot{\xi}(t) + u_\epsilon) - L(\xi(t), \dot{\xi}(t)) \right) \mathrm{d}t \\ & + \lim_{\epsilon \to 0} \int_{\bar{t} - \epsilon}^{\bar{t}} \left(L((t - \bar{t})u + \xi(\bar{t}), u) - L(\xi(t), \dot{\xi}(t)) \right) \mathrm{d}t \\ = & \, 0 + 0 = 0, \end{aligned}$$

using the fact that both integrands are bounded uniformly in ϵ. We also compute

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\epsilon} J_L(\xi_\epsilon) = & \frac{\mathrm{d}}{\mathrm{d}\epsilon} \int_{t_0}^{\bar{t} - \epsilon} L(\xi_\epsilon(t), \dot{\xi}_\epsilon(t)) \, \mathrm{d}t + \frac{\mathrm{d}}{\mathrm{d}\epsilon} \int_{\bar{t} - \epsilon}^{\bar{t}} L(\xi_\epsilon(t), \dot{\xi}_\epsilon(t)) \, \mathrm{d}t \\ = & - L(\xi(\bar{t} - \epsilon) + (\bar{t} - \epsilon - t_0)u_\epsilon, \dot{\xi}(\bar{t} - \epsilon) + u_\epsilon) + L(-\epsilon u + \xi(\bar{t}), u) \\ & + \int_{t_0}^{\bar{t} - \epsilon} \Big(\boldsymbol{D}_1 L(\xi(t) + (t - t_0)u_\epsilon, \dot{\xi}(t) + u_\epsilon) \cdot (t - t_0) \frac{\mathrm{d}u_\epsilon}{\mathrm{d}\epsilon} \\ & + \boldsymbol{D}_2 L(\xi(t) + (t - t_0)u_\epsilon, \dot{\xi}(t) + u_\epsilon) \cdot \frac{\mathrm{d}u_\epsilon}{\mathrm{d}\epsilon} \Big) \, \mathrm{d}t. \end{aligned}$$

An integration by parts gives

$$\begin{aligned} & \int_{t_0}^{\bar{t} - \epsilon} \boldsymbol{D}_2 L(\xi(t) + (t - t_0)u_\epsilon, \dot{\xi}(t) + u_\epsilon) \cdot \frac{\mathrm{d}u_\epsilon}{\mathrm{d}\epsilon} \, \mathrm{d}t \\ & \qquad = \boldsymbol{D}_2 L(\xi(t) + (t - t_0)u_\epsilon, \dot{\xi}(t) + u_\epsilon)(t - t_0) \frac{\mathrm{d}u_\epsilon}{\mathrm{d}\epsilon} \Big|_{t = t_0}^{t = \bar{t} - \epsilon} \\ & \qquad \quad - \int_{t_0}^{\bar{t} - \epsilon} \Big(\frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{D}_2 L(\xi(t) + (t - t_0)u_\epsilon, \dot{\xi}(t) + u_\epsilon) \Big) \cdot (t - t_0) \frac{\mathrm{d}u_\epsilon}{\mathrm{d}\epsilon} \, \mathrm{d}t. \quad (2.2) \end{aligned}$$

We easily see that

$$\lim_{\epsilon\to 0}\big(-L(\xi(\bar{t}-\epsilon)+(\bar{t}-\epsilon-t_0)u_\epsilon,\dot{\xi}(\bar{t}-\epsilon)+u_\epsilon)+L(-\epsilon u+\xi(\bar{t}),u)\big)$$
$$=-L(\xi(\bar{t}),\dot{\xi}(\bar{t}))+L(\xi(\bar{t}),u). \quad (2.3)$$

We have

$$\frac{\mathrm{d}u_\epsilon}{\mathrm{d}\epsilon}=\frac{(\dot{\xi}(\bar{t}-\epsilon)-u)(\bar{t}-t_0-\epsilon)+(\xi(\bar{t})-\xi(\bar{t}-\epsilon)-\epsilon u)}{(\bar{t}-t_0-\epsilon)^2},$$

whence

$$\frac{\mathrm{d}u_\epsilon}{\mathrm{d}\epsilon}\Big|_{\epsilon=0}=\frac{\dot{\xi}(\bar{t})-u}{\bar{t}-t_0}.$$

Using this, along with the integration by parts formula (2.2) and the fact that ξ satisfies the Euler–Lagrange equations by Theorem 2.2, we ascertain that

$$\lim_{\epsilon\to 0}\int_{t_0}^{\bar{t}-\epsilon}\Big(\boldsymbol{D}_1L(\xi(t)+(t-t_0)u_\epsilon,\dot{\xi}(t)+u_\epsilon)\cdot(t-t_0)\frac{\mathrm{d}u_\epsilon}{\mathrm{d}\epsilon}$$
$$+\boldsymbol{D}_2L(\xi(t)+(t-t_0)u_\epsilon,\dot{\xi}(t)+u_\epsilon)\cdot\frac{\mathrm{d}u_\epsilon}{\mathrm{d}\epsilon}\Big)\,\mathrm{d}t=\boldsymbol{D}_2L(\xi(\bar{t}),\dot{\xi}(\bar{t}))\cdot(\dot{\xi}(\bar{t})-u). \quad (2.4)$$

Combining (2.3) and (2.4) we have shown that

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon}\Big|_{\epsilon=0}\Delta_L(\epsilon)=\frac{\mathrm{d}}{\mathrm{d}\epsilon}\Big|_{\epsilon=0}J_L(\xi_\epsilon)=E_L(\xi(\bar{t}),\dot{\xi}(\bar{t}),u)<0,$$

implying that for some sufficiently small ϵ there exists $\delta>0$ such that $J_L(\xi_\epsilon)-J_L(\xi)<-\delta$. One can now approximate ξ_ϵ with a curve $\tilde{\xi}$ of class C^2 and for which $|J_L(\tilde{\xi})-J_L(\xi_\epsilon)|<\frac{\delta}{2}$; see [Hansen 2005]. In this case we have $J_L(\tilde{\xi})-J_L(\xi)<-\frac{\delta}{2}$ which contradicts $\xi\in\mathscr{P}(L,x_0,x_1,[t_0,t_1])$. ■

The reader is asked to give an interpretation of Weierstrass's necessary condition in Exercise E2.2. This interpretation is actually an interesting one since it reveals that the Weierstrass condition is really one about the existence of what will appear to us as a "separating hyperplane" in our proof of the Maximum Principle.

2.2. Discussion of calculus of variations

Before we proceed with a reformulation of the necessary conditions of the preceding section, let us say some things about the theorems we proved, and the means by which we proved them.

In each of the above theorems, the proof relied on constructing a family of curves in which our curve of interest was embedded. Let us formalise this idea.

2.8 Definition: (Variation of a curve) Let $\mathcal{X}\subset\mathbb{R}^n$ be an open set, let $x_0,x_1\in\mathcal{X}$, let $t_0,t_1\in\mathbb{R}$ satisfy $t_0<t_1$, and let $\xi\in C^2(x_0,x_1,[t_0,t_1])$. A ***variation*** of ξ is a map $\sigma\colon J\times[t_0,t_1]\to\mathcal{X}$ with the following properties:

(i) $J\subset\mathbb{R}$ is an interval for which $0\in\mathrm{int}(J)$;

(ii) σ is of class C^2;

(iii) $\sigma(s, t_0) = x_0$ and $\sigma(s, t_1) = x_1$ for all $s \in J$;

(iv) $\sigma(0, t) = \xi(t)$ for all $t \in [t_0, t_1]$.

For a variation σ of ξ, the corresponding ***infinitesimal variation*** is the map $\delta\sigma\colon [t_0, t_1] \to \mathbb{R}^n$ defined by

$$\delta\sigma(t) = \frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} \sigma(s, t). \qquad \bullet$$

In Figure 2.2 we depict how one should think of variations and infinitesimal variations.

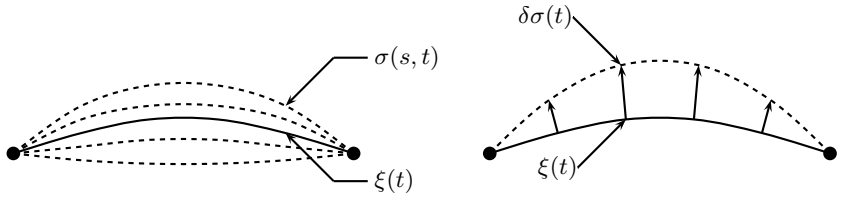

Figure 2.2. A variation (left) and an infinitesimal variation (right)

A variation is a family of curves containing ξ and the corresponding infinitesimal variation measures how the family varies for small values of the parameter indexing the family. In Figure 2.3 we depict the infinitesimal variations used in the proofs of Theorems 2.2, 2.4,

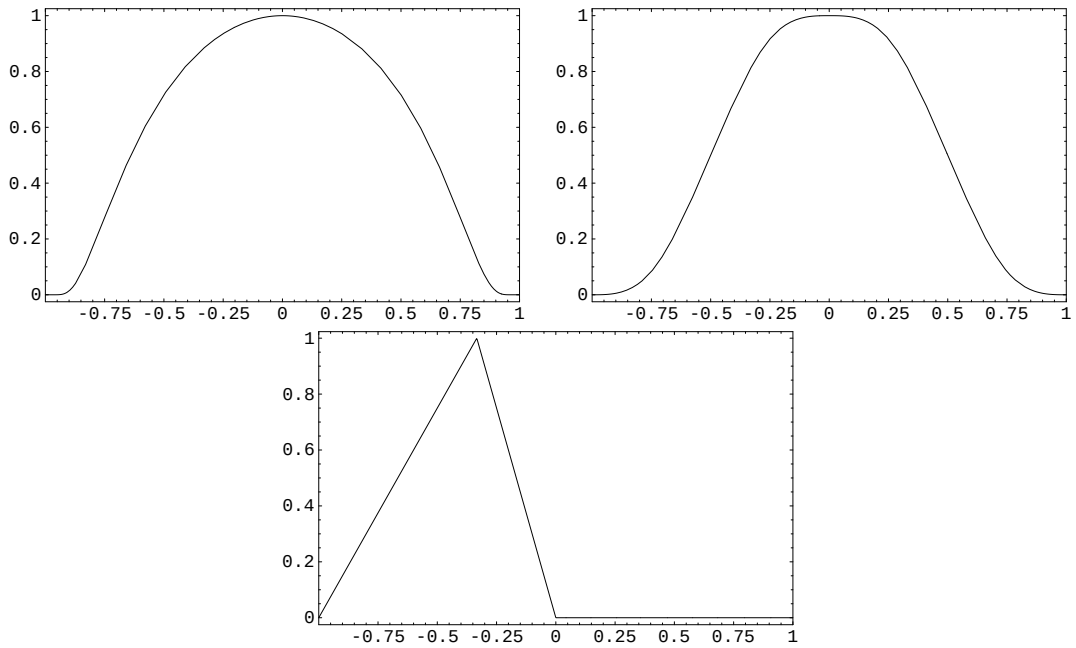

Figure 2.3. Infinitesimal variations used in the necessary conditions of Euler–Lagrange (top left), Legendre (top right), and Weierstrass (bottom). In each case the point $\bar{t}$ in the proofs of Theorems 2.2, 2.4, and 2.7 is to be thought of as being at zero.

and 2.7. Note that the variation used in Theorem 2.7 is of a different character than the

other two. The family of curves associated with the first two variations are obtained by adding the variations scaled by s. Thus, for example, the varied curves and their time derivatives approach the original curve and its time derivative uniformly in time as $s \to 0$. However, the variation for Theorem 2.7 is constructed in a different manner. While the varied curves approach the original one uniformly in time as $s \to 0$, the same does not hold for velocities. Moreover, Theorem 2.7 is a result of a rather different flavour that Theorems 2.2 and 2.4. The latter two theorems have more or less direct analogues to ideas from finite-dimensional optimisation theory, while the former does not. It bears mentioning here that it is the sort of variation appearing in Theorem 2.7 that we will mimic in our proof of the Maximum Principle. These variations will be called needle variations. This name also makes sense for the variations use in proving the Weierstrass necessary condition, if one keeps in mind that it is the variation of the velocity that one is interested in in this case.

The idea of a variation will be of essential importance to us in the proof of the Maximum Principle. Indeed, the idea of a variation is of importance in control theory in general. In our three necessary conditions from the preceding section we were able to be quite explicit about the variations we constructed, and the effect they had on the cost function. In control theory, it is not the usual case that one can so easily measure the effects of a variation. This accounts for some of the technical fussiness we will encounter in proving the Maximum Principle.

Now let us consider a few specific problems that will exhibit some of the sorts of things that can happen in the calculus of variations, and therefore in optimal control theory.

The first example is one where there are no solutions to the problem of the calculus of variations.

2.9 Example: (A problem with no solutions) We take $\mathcal{X} = \mathbb{R}$, $x_0 = 0$, $x_1 = 1$, $t_0 = 0$, and $t_1 = 1$. As a Lagrangian we consider $L(x, v) = x$. Then the first-order necessary condition, namely the Euler–Lagrange equation, is the equation

$$1 = 0.$$

This equation has no solutions, of course.

This can also be seen from the problem itself. The problem is to minimise the integral

$$\int_0^1 \xi(t)\,\mathrm{d}t$$

over all C^2-functions $\xi\colon [0,1] \to \mathbb{R}$ satisfying $\xi(0) = 0$ and $\xi(1) = 1$. Included in this set of functions are the functions $x \mapsto nt^2 + (1-n)t$, $n \in \mathbb{Z}_{>0}$. For these functions we have

$$\int_0^1 (nt^2 + (1-n)t)\,\mathrm{d}t = \frac{3-n}{6}.$$

Therefore, for any $M > 0$ there exists $\xi \in C^2(x_0, x_1, [0,1])$ such that $J_L(\xi) < -M$. Thus $\mathscr{P}(L, 0, 1, [0,1]) = \emptyset$. •

Our next example is one where there are an infinite number of solutions.

2.10 Example: (A problem with many solutions) We again take $\mathcal{X} = \mathbb{R}$, $x_0 = 0$, $x_1 = 1$, $t_0 = 0$, and $t_1 = 1$. The Lagrangian we use is $L(x, v) = v$, and this gives the Euler–Lagrange equations

$$0 = 0.$$

Thus *every* curve in $C^2(0, 1, [0, 1])$ satisfies the first-order necessary conditions of Theorem 2.2. Since $\boldsymbol{D}_2^2 L(x, v) = 0$, the necessary condition of Theorem 2.4 is also satisfied. The Weierstrass excess function is also zero, and so the necessary condition of Theorem 2.7 obtains as well.

However, the satisfaction of the three necessary conditions does not ensure that a curve is a solution of the problem. Nonetheless, in this case it can be seen directly that $\mathscr{P}(L, 0, 1, [0, 1]) = C^2(0, 1, [0, 1])$. Indeed, for $\xi \in C^2(0, 1, [0, 1])$ we have

$$\int_0^1 L(\xi(t), \dot{\xi}(t)) \, \mathrm{d}t = \int_0^1 \dot{\xi}(t) \, \mathrm{d}t = \xi(1) - \xi(0) = 1.$$

Thus $J_L(\xi)$ is actually independent of ξ, and so every curve minimises J_L. •

Next we consider a problem where there are no solutions of class C^2, but solutions exist with weaker differentiability hypotheses. This suggests that the class $C^2(x_0, x_1, [t_0, t_1])$ is not always the correct one to deal with. In optimal control theory, we will deal with very general classes of curves, and this accounts for some of the technicalities we will encounter in the proof of the Maximum Principle.

2.11 Example: (A problem with no solutions of class C^2) We let $\mathcal{X} = \mathbb{R}$, $t_0 = -1$, and $t_1 = 1$ and define $L(t, x, v) = x^2(2t - v)^2$.[2] As boundary conditions we choose $x_0 = 0$ and $x_1 = 1$. Clearly since L is always positive J_L is bounded below by zero. Thus any curve ξ for which $J_L(\xi) = 0$ will be a global minimum. In particular

$$t \mapsto \begin{cases} 0, & -1 \leq t < 0, \\ t^2, & 0 \leq t \leq 1, \end{cases}$$

is a global minimiser of

$$J_L(\xi) = \int_{t_0}^{t_1} \xi(t)^2 (2t - \dot{\xi}(t))^2 \, \mathrm{d}t.$$

Note that ξ as defined is of class C^1 but is not of class C^2. •

There are many phenomenon in the calculus of variations that are not covered by the above examples. However, since we will be considering a far more general theory in the form of the Maximum Principle, we do not dedicate any time to this here. The reader can refer to one of books on the topic of calculus of variations cited at the beginning of this section.

[2]The Lagrangian we consider in this problem is time-dependent. We have not formally talked about time-dependent Lagrangians, although the development for these differs from ours merely by adding a "t" in the appropriate places.

2.3. The Skinner–Rusk formulation of the calculus of variations

In a cluster of papers Skinner [1983] and Skinner and Rusk [1983a, 1983b] developed a framework for mechanics on the Whitney sum of the tangent and cotangent bundles (although surely this idea predates 1983). In this section we review this idea, but in the absence of the geometric machinery (although with this machinery, the picture is somewhat more compelling). Our presentation closely follows Sussmann and Willems [1997].

We let $\mathcal{X} \subset \mathbb{R}^n$ be an open set and we consider the product $\mathcal{X} \times \mathbb{R}^n \times \mathbb{R}^n$ in which typical points are denoted by (x, v, p). Given a Lagrangian $L\colon \mathcal{X} \times \mathbb{R}^n \to \mathbb{R}$ of class C^2 we define a C^2 function H_L by

$$H_L(x, v, p) = \langle p, v \rangle - L(x, v),$$

which we call the ***Hamiltonian*** for the Lagrangian L.

We now state a condition that is equivalent to Theorem 2.2.

2.12 Theorem: (A Hamiltonian formulation of the Euler–Lagrange equations) *The following statements are equivalent for a C^2-curve $t \mapsto (\xi(t), \chi(t), \lambda(t))$ in $\mathcal{X} \times \mathbb{R}^n \times \mathbb{R}^n$:*

(i) $\chi(t) = \dot{\xi}(t)$, and the Euler–Lagrange equations for ξ are satisfied, along with the equation

$$\lambda(t) = \boldsymbol{D}_2 L(\xi(t), \chi(t));$$

(ii) the following equations hold:

$$\begin{aligned} \dot{\xi}(t) &= \boldsymbol{D}_3 H_L(\xi(t), \chi(t), \lambda(t)), \\ \dot{\lambda}(t) &= -\boldsymbol{D}_1 H_L(\xi(t), \chi(t), \lambda(t)), \\ \boldsymbol{D}_2 H_L(\xi(t), \chi(t), \lambda(t)) &= 0. \end{aligned}$$

Proof: (i) $\Longrightarrow$ (ii) The equation additional to the Euler–Lagrange equation immediately implies the third of the equations of part (ii). Since $\boldsymbol{D}_3 H_L(x, v, p) = v$ and since $\chi = \dot{\xi}$, it follows that

$$\dot{\xi}(t) = \boldsymbol{D}_3 H_L(\xi(t), \chi(t), \lambda(t)),$$

which is the first of the equations of part (ii). Finally, taking the derivative of the equation additional to the Euler–Lagrange equation and using the relation

$$\boldsymbol{D}_1 H_L(x, v, p) = -\boldsymbol{D}_1 L(x, v)$$

gives the second of the equations of part (ii).

(ii) $\Longrightarrow$ (i) The third of the equations of part (ii) implies that

$$\lambda(t) = \boldsymbol{D}_2 L(\xi(t), \chi(t)),$$

which is the equation additional to the Euler–Lagrange equations in part (i). The first equation of part (ii) implies that $\dot{\xi}(t) = \chi(t)$. Also, since

$$\boldsymbol{D}_1 H_L(x, v, p) = -\boldsymbol{D}_1 L(x, v),$$

this shows that the equations of part (ii) imply

$$\frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{D}_2 L(\xi(t), \dot{\xi}(t))) = \boldsymbol{D}_1 L(\xi(t), \dot{\xi}(t)),$$

which are the Euler–Lagrange equations. ■

The first two of the three equations of part (ii) of Theorem 2.12 are the classical Hamilton's equations, although they now involve a "parameter" v. The importance of the third of these equations becomes fully realised when one throws Theorems 2.4 and 2.7 into the mix.

2.13 Theorem: (A Hamiltonian maximisation formulation of the necessary conditions) *The following statements are equivalent for a C^2-curve ξ on $\mathcal{X}$:*

(i) ξ satisfies the necessary conditions of Theorems 2.2, 2.4, and 2.7;

(ii) there exists a differentiable map $t \mapsto \lambda(t)$ so that, for the curve $t \mapsto (\xi(t), \dot\xi(t), \lambda(t))$, the following relations hold:

$$\begin{aligned}
&\dot\xi(t) = \boldsymbol{D}_3 H_L(\xi(t), \dot\xi(t), \lambda(t)),\\
&\dot\lambda(t) = -\boldsymbol{D}_1 H_L(\xi(t), \dot\xi(t), \lambda(t)),\\
&H_L(\xi(t), \dot\xi(t), \lambda(t)) = \max\{H_L(\xi(t), u, \lambda(t)) \mid\ u \in \mathbb{R}^n\}.
\end{aligned}$$

Proof: (i) $\Longrightarrow$ (ii) Define $\lambda(t) = \boldsymbol{D}_2 L(\xi(t), \dot\xi(t))$. By Theorem 2.12, the necessary condition of Theorem 2.2 implies the first two of the equations of part (ii). Since

$$\boldsymbol{D}_2^2 H_L(x, v, p) = -\boldsymbol{D}_2^2 L(x, v),$$

the necessary condition of Theorem 2.4 and the third equation of part (ii) of Theorem 2.12 tell us that

$$\begin{aligned}
&\boldsymbol{D}_2 H_L(\xi(t), \dot\xi(t), \lambda(t)) = 0,\\
&\boldsymbol{D}_2^2 H_L(\xi(t), \dot\xi(t), \lambda(t)) \leq 0.
\end{aligned}$$

Now we consider the necessary condition of Theorem 2.7. We note that the excess function satisfies

$$E_L(p, v, u) = H_L(x, v, \boldsymbol{D}_2 L(x, v)) - H_L(x, u, \boldsymbol{D}_2 L(x, v)).$$

Therefore the necessary condition of Theorem 2.7 translates to asserting that, for each $t \in [t_0, t_1]$,

$$H_L(\xi(t), \dot\xi(t), \lambda(t)) - H^L(\xi(t), u, \lambda(t)) \geq 0, \qquad u \in \mathbb{R}^n. \tag{2.5}$$

This is exactly the third of the equations of part (ii).

(ii) $\Longrightarrow$ (i) The third of the equations of part (ii) implies that $\boldsymbol{D}_2 H_L(\xi(t), \dot\xi(t), \lambda(t)) = 0$. The definition of H_L then gives $\lambda(t) = \boldsymbol{D}_2 L(\xi(t), \dot\xi(t))$. This then defines λ. By Theorem 2.12 this also implies that the necessary condition of Theorem 2.2 holds. The third of the equations of part (ii) also implies that (2.5) holds, this then implying that the necessary condition of Theorem 2.7, and particularly those of Theorem 2.4, hold. ■

2.4. From the calculus of variations to the Maximum Principle, almost

We now make a transition to optimal control from the calculus of variations setting of the preceding section. We now suppose that we have a control system $\Sigma = (\mathcal{X}, f, U)$, and we restrict velocities to satisfy the constraint imposed by the control equations:

$$\dot\xi(t) = f(\xi(t), u(t)).$$

Note then that the admissible velocities are now parameterised by the control set U. Therefore, a Lagrangian will now be a function not on $\mathcal{X} \times \mathbb{R}^n$, but on $\mathcal{X} \times U$. Let us try to be a little more explicit about this, since this is an important point to understand in relating the calculus of variations to optimal control. Let us define a subset D_Σ of $\mathcal{X} \times \mathbb{R}^n$ by

$$D_\Sigma = \{(x, f(x, u)) \in \mathcal{X} \times \mathbb{R}^n \mid u \in U\}.$$

Thus D_Σ is the subset of the states x and velocities v that satisfy the equations specifying the control system. That is to say, if $(\xi, u) \in \mathrm{Ctraj}(\Sigma)$ then $\dot{\xi}(t) \in D_\Sigma$. Therefore, to specify a Lagrangian for the system it suffices to define the Lagrangian only on D_Σ. However, this can be done simply by using the Lagrangian $L\colon \mathcal{X} \times U \to \mathbb{R}$. Indeed, we define a function $L_\Sigma\colon D_\Sigma \to \mathbb{R}$ by

$$L_\Sigma(x, f(x, u)) = L(x, u).$$

Thus the optimal control problem can be turned into a problem resembling the calculus of variations problem if it is understood that there are constraints on the subset of $\mathcal{X} \times \mathbb{R}^n$ where the admissible velocities lie.

We now argue from Theorem 2.13 to a plausible (but incorrect) solution to Problem 1.8. The ***control Hamiltonian*** for Σ and L is the function on $\mathcal{X} \times \mathbb{R}^n \times U$ given by

$$H_{\Sigma,L}(x, p, u) = \langle p, f(x, u) \rangle - L(x, u).$$

Where in the calculus of variations we chose the velocity v to maximise the Hamiltonian with $(x, p) \in \mathcal{X} \times \mathbb{R}^n$ fixed, we now fix $(x, p) \in \mathcal{X} \times \mathbb{R}^n$ and maximise the control Hamiltonian with respect to the control:

$$H^{\max}_{\Sigma,L}(x, p) = \max_{u \in U} H_{\Sigma,L}(x, p, u).$$

With Theorem 2.13 in mind, we state the following non-result.

2.14 This is not a theorem: *If* $(\xi, u) \in \mathrm{Carc}(\Sigma)$ *solves Problem 1.8 then there exists a* C^1*-map* λ *such that*

(i) $H_{\Sigma,L}(\xi(t), \lambda(t), u(t)) = H^{\max}_{\Sigma,L}(\xi(t), \lambda(t))$ *and*

(ii) the differential equations

$$\begin{aligned} \dot{\xi}(t) &= \boldsymbol{D}_2 H_{\Sigma,L}(\xi(t), \lambda(t), u(t)), \\ \dot{\lambda}(t) &= -\boldsymbol{D}_1 H_{\Sigma,L}(\xi(t), \lambda(t), u(t)) \end{aligned}$$

are satisfied.

While the preceding attempt at a solution to the problems of optimal control from the necessary conditions of the calculus of variations is not quite correct, it nevertheless captures some part of the essence of the Maximum Principle. A somewhat uninteresting (although still important) way in which our attempt differs from the actual Maximum Principle is through the fact that its direct adaptation from the calculus of variations demands more smoothness of the objects involved than is necessary. A more substantial way that the attempt deviates from the correct version involves the absence in the attempt of the "abnormal multiplier." This is contained in the correct version, to whose precise statement we now turn.

Exercises

E2.1 Recall that a function $f\colon \mathbb{R}^n \to \mathbb{R}$ is ***convex*** if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$$

for all $\lambda \in [0,1]$ and all $x, y \in \mathbb{R}^n$.

Show that if, for fixed $x \in \mathcal{X}$, the function $v \mapsto L(x,v)$ is convex then the excess function satisfies $E_L(x,v,u) \geq 0$ for all $u \in \mathbb{R}^n$.

In the following exercise you will give a geometric interpretation of the necessary condition of Weierstrass. We recommend that the reader refer back to this exercise during the course of the proof of Lemma 6.3 and "compare notes."

E2.2 For a Lagrangian $L\colon \mathcal{X} \times \mathbb{R}^n \to \mathbb{R}$ consider fixing $x \in \mathcal{X}$ and thinking of the resulting function $L_x\colon v \mapsto L(x,v)$. This is sometimes called the ***figurative***.

(a) In Figure E2.1 we depict the function L_x in the case where $\mathcal{X} \subset \mathbb{R}$. Explain

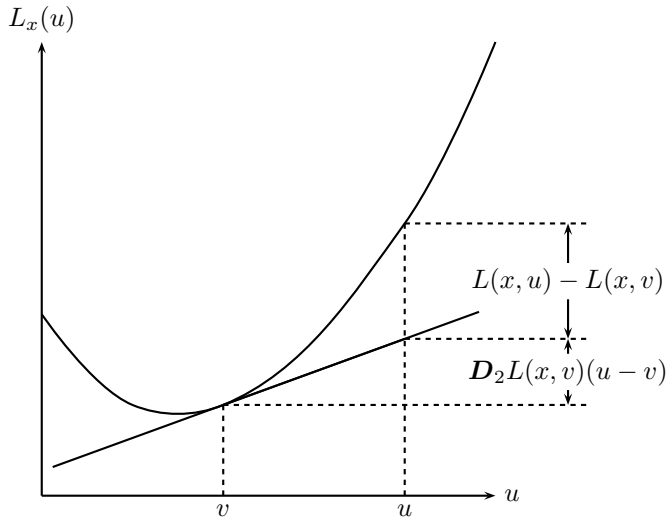

Figure E2.1. The figurative when $n = 1$

all the labels in the figure and use the figure to give an interpretation of the necessary condition of Theorem 2.7 in this case.

(b) Extend your explanation from part (a) to the case of general n.

(c) After understanding the proof of the Maximum Principle, restate Theorem 2.7 using separating hyperplanes.

The following exercise gives the initial steps in the important connection between the calculus of variations and classical mechanics.

E2.3 Let $m \in \mathbb{R}_{>0}$, let $V : \mathbb{R}^3 \to \mathbb{R}$ be of class C^2, and define a Lagrangian on $\mathbb{R}^3 \times \mathbb{R}^3$ by

$$L(x, v) = \tfrac{1}{2} m \|v\|^2 - V(x).$$

Show that the Euler–Lagrange equations are Newton's equations for a particle of mass m moving in the potential field V.

E2.4 Take $\mathcal{X} = \mathbb{R}$ and define a Lagrangian by $L(x, v) = (v^2 - 1)^2$. Take $t_0 = 0$, $t_1 = 1$, $x_0 = 0$, and $x_1 = 1$.

(a) What is the solution to Problem 2.1 in this case?

(b) Does the solution you found in part (a) satisfy the Euler–Lagrange equations?

(c) Can you find piecewise differentiable solutions to Problem 2.1 in this case?

Chapter 3

The Maximum Principle

In this chapter we give a precise statement of one version of the Maximum Principle, one that is essentially that given by [Pontryagin, Boltyanskii, Gamkrelidze, and Mishchenko 1961] as this still holds up pretty well after all these years. We shall only prove the Maximum Principle in Chapter 6 after we have taken the time to properly prepare ourselves to understand the proof. After the proof, in Chapter 7 we will discuss some of the important issues that arise from the statement of the Maximum Principle.

Of course, since the Maximum Principle has been around for nearly a half century at this point, one can find many textbook treatments. Some of these include books by Lee and Markus [1967], Hestenes [1966], Berkovitz [1974], and Jurdjevic [1997]. Our treatment bears some resemblance to that of Lee and Markus [1967].

For convenience, let us reproduce here the statement of the two problems that our version of the Maximum Principle deals with. We refer the reader to Section 1.3 for notation.

Problem 1.7: (Free interval optimal control problem) Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let L be a Lagrangian for Σ, and let $S_0, S_1 \subset \mathcal{X}$ be sets. A controlled trajectory $(\xi_*, \mu_*) \in \text{Carc}(\Sigma, L, S_0, S_1)$ is a solution to the ***free interval optimal control problem*** for Σ, L, S_0, and S_1 if $J_{\Sigma,L}(\xi_*, \mu_*) \leq J_{\Sigma,L}(\xi, \mu)$ for each $(\xi, \mu) \in \text{Carc}(\Sigma, L, S_0, S_1)$. The set of solutions to this problem is denoted by $\mathscr{P}(\Sigma, L, S_0, S_1)$. •

Problem 1.8: (Fixed interval optimal control problem) Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let L be a Lagrangian for Σ, let $S_0, S_1 \subset \mathcal{X}$ be sets, and let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$. A controlled trajectory $(\xi_*, \mu_*) \in \text{Carc}(\Sigma, L, S_0, S_1, [t_0, t_1])$ is a solution to the ***fixed interval optimal control problem*** for Σ, L, S_0, and S_1 if $J_{\Sigma,L}(\xi_*, \mu_*) \leq J_{\Sigma,L}(\xi, \mu)$ for each $(\xi, \mu) \in \text{Carc}(\Sigma, L, S_0, S_1, [t_0, t_1])$. The set of solutions to this problem is denoted by $\mathscr{P}(\Sigma, L, S_0, S_1, [t_0, t_1])$. •

3.1. Preliminary definitions

As is suggested by our development of Section 2.4, an important ingredient in the Maximum Principle is the Hamiltonian associated to the optimal control problem for the system $\Sigma = (\mathcal{X}, f, U)$ with Lagrangian L. We shall have occasion to use various Hamiltonians, so let us formally define these now.

3.1 Definition: (Hamiltonian, extended Hamiltonian, maximum Hamiltonian, maximum extended Hamiltonian) Let $\Sigma = (\mathcal{X}, f, U)$ be a control system and let L be a Lagrangian.

(i) The ***Hamiltonian*** is the function $H_\Sigma\colon \mathcal{X} \times \mathbb{R}^n \times U \to \mathbb{R}$ defined by

$$H_\Sigma(x, p, u) = \langle p, f(x, u)\rangle.$$

(ii) The ***extended Hamiltonian*** is the function $H_{\Sigma,L}\colon \mathcal{X} \times \mathbb{R}^n \times U \to \mathbb{R}$ defined by

$$H_{\Sigma,L}(x, p, u) = \langle p, f(x, u)\rangle + L(x, u).$$

(iii) The ***maximum Hamiltonian*** is the function $H_\Sigma^{\max}\colon \mathcal{X} \times \mathbb{R}^n \to \overline{\mathbb{R}}$ defined by

$$H_\Sigma^{\max}(x, p) = \sup\{H_\Sigma(x, p, u) \mid u \in U\}.$$

(iv) The ***maximum extended Hamiltonian*** is the function $H_{\Sigma,L}^{\max}\colon \mathcal{X} \times \mathbb{R}^n \to \overline{\mathbb{R}}$ defined by

$$H_{\Sigma,L}^{\max}(x, p) = \sup\{H_{\Sigma,L}(x, p, u) \mid u \in U\}.$$ •

The variable "p" on which the various Hamiltonians depend is sometimes called the ***costate***.

In Section 6.1 we shall see that the extended Hamiltonian is, in fact, the Hamiltonian for what we shall call the extended system.

Another key idea in the Maximum Principle is the notion of an adjoint response.

3.2 Definition: (Adjoint response) Let $\Sigma = (\mathcal{X}, f, U)$ be a control system and let $(\xi, \mu) \in \text{Ctraj}(\Sigma)$ have time interval I.

(i) An ***adjoint response*** for Σ along (ξ, μ) is a locally absolutely continuous map $\lambda\colon I \to \mathbb{R}^n$ such that the differential equation

$$\begin{aligned}\dot\xi(t) &= \boldsymbol{D}_2 H_\Sigma(\xi(t), \lambda(t), \mu(t)),\\ \dot\lambda(t) &= -\boldsymbol{D}_1 H_\Sigma(\xi(t), \lambda(t), \mu(t))\end{aligned}$$

is satisfied.

Now additionally let L be a Lagrangian, and let $(\xi, \mu) \in \text{Ctraj}(\Sigma, L)$ have time interval I.

(ii) An ***adjoint response*** for (Σ, L) along (ξ, μ) is a locally absolutely continuous map $\lambda\colon I \to \mathbb{R}^n$ such that the differential equation

$$\begin{aligned}\dot\xi(t) &= \boldsymbol{D}_2 H_{\Sigma,L}(\xi(t), \lambda(t), \mu(t)),\\ \dot\lambda(t) &= -\boldsymbol{D}_1 H_{\Sigma,L}(\xi(t), \lambda(t), \mu(t))\end{aligned}$$

is satisfied. •

Part of the Maximum Principle deals with the case when the initial set S_0 and the final set S_1 have a nice property. In a properly geometric treatment of the subject one might ask that these subsets be submanifolds. In more general treatments, one might relax smoothness conditions on these subsets. However, since we are striving for neither geometric elegance nor generality, the subsets we consider are of the following sort.

3.3 Definition: (Constraint set) A ***smooth constraint set*** for a control system $\Sigma = (\mathcal{X}, f, U)$ is a subset $S \subset \mathcal{X}$ of the form $S = \Phi^{-1}(0)$ where $\Phi\colon \mathcal{X} \to \mathbb{R}^k$ is of class C^1 and has the property that $\boldsymbol{D}\Phi(x)$ is surjective for each $x \in \Phi^{-1}(0)$. •

Of course, a smooth constraint set is a submanifold if you are familiar with the notion of a submanifold.

3.2. The statement of the Maximum Principle

We now have the data needed to give necessary conditions which can be applied to the Problems 1.7 and 1.8. First we consider the free interval problem.

3.4 Theorem: (Maximum Principle for free interval problems) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let L be a Lagrangian, and let S_0 and S_1 be subsets of $\mathcal{X}$. If $(\xi_*, \mu_*) \in \mathscr{P}(\Sigma, L, S_0, S_1)$ is defined on $[t_0, t_1]$ then there exists an absolutely continuous map $\lambda_* \colon [t_0, t_1] \to \mathbb{R}^n$ and $\lambda_*^0 \in \{0, -1\}$ with the following properties:*

(i) either $\lambda_*^0 = -1$ or $\lambda_*(t_0) \neq 0$;

(ii) λ_* is an adjoint response for $(\Sigma, \lambda_*^0 L)$ along (ξ_*, μ_*);

(iii) $H_{\Sigma,\lambda_*^0 L}(\xi_*(t), \lambda_*(t), \mu_*(t)) = H^{\max}_{\Sigma,\lambda_*^0 L}(\xi_*(t), \lambda_*(t))$ for almost every $t \in [t_0, t_1]$.

If, additionally, $\mu_* \in \mathscr{U}_{\text{bdd}}([t_0, t_1])$, then

(iv) $H^{\max}_{\Sigma,\lambda_*^0 L}(\xi_*(t), \lambda_*(t)) = 0$ for every $t \in [t_0, t_1]$.

Moreover, if S_0 and S_1 are smooth constraint sets defined by

$$S_0 = \Phi_0^{-1}(0), \quad S_1 = \Phi_1^{-1}(0)$$

for maps $\Phi_a \colon \mathbb{R}^n \to \mathbb{R}^{k_a}$, $a \in \{1, 2\}$, then λ_* can be chosen such that

(v) $\lambda_*(t_0)$ is orthogonal to $\ker(\boldsymbol{D}\Phi_0(\xi(t_0)))$ and $\lambda_*(t_1)$ is orthogonal to $\ker(\boldsymbol{D}\Phi_1(\xi(t_1)))$.

For the fixed interval problem we merely lose the fact that the maximum Hamiltonian is zero.

3.5 Theorem: (Maximum Principle for fixed interval problems) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let L be a Lagrangian, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let S_0 and S_1 be subsets of $\mathcal{X}$. If $(\xi_*, \mu_*) \in \mathscr{P}(\Sigma, L, S_0, S_1, [t_0, t_1])$ then there exists an absolutely continuous map $\lambda_* \colon [t_0, t_1] \to \mathbb{R}^n$ and $\lambda_*^0 \in \{0, -1\}$ with the following properties:*

(i) either $\lambda_*^0 = -1$ or $\lambda_*(t_0) \neq 0$;

(ii) λ_* is an adjoint response for $(\Sigma, \lambda_*^0 L)$ along (ξ_*, μ_*);

(iii) $H_{\Sigma,\lambda_*^0 L}(\xi_*(t), \lambda_*(t), \mu_*(t)) = H^{\max}_{\Sigma,\lambda_*^0 L}(\xi_*(t), \lambda_*(t))$ for almost every $t \in [t_0, t_1]$.

If, additionally, $\mu_* \in \mathscr{U}_{\text{bdd}}([t_0, t_1])$, then

(iv) $H^{\max}_{\Sigma,\lambda_*^0 L}(\xi_*(t), \lambda_*(t)) = H^{\max}_{\Sigma,L}(\xi_*(t_0), \lambda_*(t_0))$ for every $t \in [t_0, t_1]$.

Moreover, if S_0 and S_1 are smooth constraint sets defined by

$$S_0 = \Phi_0^{-1}(0), \quad S_1 = \Phi_1^{-1}(0)$$

for maps $\Phi_a \colon \mathbb{R}^n \to \mathbb{R}^{k_a}$, $a \in \{1, 2\}$, then λ_* can be chosen such that

(v) $\lambda_*(t_0)$ is orthogonal to $\ker(\boldsymbol{D}\Phi_0(\xi(t_0)))$ and $\lambda_*(t_1)$ is orthogonal to $\ker(\boldsymbol{D}\Phi_1(\xi(t_1)))$.

3.6 Remark: (Nonvanishing of the "total" adjoint vector) We shall see in the course of the proof of the Maximum Principle (specifically in the proof of Lemma 6.3) that the condition that either $\lambda_*^0 = -1$ or $\lambda_*(t_0) \neq 0$ amounts to the condition that $(\lambda_*^0)^2 + \|\lambda_*(t)\|^2 \neq 0$ for all $t \in [t_0, t_1]$. Thus the "total" adjoint vector is nonzero. •

3.7 Remark: (Transversality conditions) The final of the conclusions of Theorems 3.4 and 3.5 are called ***transversality conditions*** since they are simply stating the transversality of the initial and final adjoint vectors with the tangent spaces to the smooth constraint sets. •

Note that if (ξ, μ) satisfies the hypotheses of the suitable Theorem 3.4 or Theorem 3.5, then this does not imply that (ξ, μ) is a solution of Problem 1.7 or 1.8, respectively. However, controlled trajectories satisfying the necessary conditions are sufficiently interesting that they merit their own name and some classification based on their properties.

First we give the notation we attach to these controlled trajectories. We shall discriminate between various forms of such controlled trajectories in Sections 7.1 and 7.2.

3.8 Definition: (Controlled extremal, extremal, extremal control) Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let L be a Lagrangian, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $S_0, S_1 \subset \mathcal{X}$.

(i) A controlled trajectory (ξ, μ) is a ***controlled extremal*** for $\mathscr{P}(\Sigma, L, S_0, S_1)$ (resp. $\mathscr{P}(\Sigma, L, S_0, S_1, [t_0, t_1])$) if it satisfies the necessary conditions of Theorem 3.4 (resp. Theorem 3.5).

(ii) An absolutely continuous curve ξ is an ***extremal*** for $\mathscr{P}(\Sigma, L, S_0, S_1)$ (resp. $\mathscr{P}(\Sigma, L, S_0, S_1, [t_0, t_1])$) if there exists an admissible control μ such that (ξ, μ) satisfies the necessary conditions of Theorem 3.4 (resp. Theorem 3.5).

(iii) An admissible control μ is an ***extremal control*** for $\mathscr{P}(\Sigma, L, S_0, S_1)$ (resp. $\mathscr{P}(\Sigma, L, S_0, S_1, [t_0, t_1])$) if there exists an absolutely continuous curve ξ such that (ξ, μ) satisfies the necessary conditions of Theorem 3.4 (resp. Theorem 3.5). •

One way to think of the Maximum Principle is as identifying a subset of controlled trajectories as being candidates for optimality. Sometimes this restricted candidacy is enough to actually characterise the optimal trajectories. In other instances it can turn out that much additional work needs to be done beyond the Maximum Principle to distinguish the optimal trajectories from the extremals.

3.3. The mysterious parts and the useful parts of the Maximum Principle

The statement of the Maximum Principle, despite our attempts to motivate it using the calculus of variations, probably seems a bit mysterious. Some of the more mysterious points include the following.

1. *Where does the adjoint response λ_* come from?* A complete understanding of this question is really what will occupy us for the next three chapters. We shall see in the proofs of Theorems 5.16 and 5.18 that the adjoint response has a very concrete geometric meaning in terms of the boundary of the reachable set.
2. *What is the meaning of the pointwise (in time) maximisation of the Hamiltonian?* As with the adjoint response λ_*, the meaning of the Hamiltonian maximisation condition is only understood by digging deeply into the proof of the Maximum Principle. When we do, we shall see that the Hamiltonian maximisation condition is essentially a condition that ensures that one is doing as best as one can, at almost every instant of time, to reduce the cost.

3. *What is the rôle of the constant λ_*^0?* In Section 7.1 we shall visit this question in some detail, but it is only possible to really appreciate this *after* one has understood the proof of the Maximum Principle. There is a finite-dimensional analogue of this which we consider in Exercise E3.2.

Despite these mysterious points, it is possible to *use* the Maximum Principle without actually understanding it. Let us indicate how this is often done. We will employ the strategy we outline in the developments of Chapters 8 and 9.

1. *Use the Hamiltonian maximisation condition to determine the control.* It is very often the case that the third condition in Theorems 3.4 and 3.5 allow one to solve for the control as a function of the state x and the costate p. Cases where this is not possible are called "singular." We shall say something about these in Section 7.2.
2. *Study the Hamiltonian equations.* If one does indeed have the control as a function of the state and costate from the previous step, then one studies the equations

$$\begin{aligned}\dot{\xi}(t) &= \boldsymbol{D}_2 H_{\Sigma,\lambda_*^0 L}(\xi(t), \lambda(t), \mu(\xi(t), \lambda(t))),\\ \dot{\lambda}(t) &= -\boldsymbol{D}_1 H_{\Sigma,\lambda_*^0 L}(\xi(t), \lambda(t), \mu(\xi(t), \lambda(t)))\end{aligned}$$

   to understand the behaviour of the extremals. Note that the control may not be a continuous function of x and p, so care needs to be exercised in understanding how solutions to these equations might be defined.
3. *Work, pray, or learn some other things to determine which extremals are optimal.* The preceding two steps allow a determination of the differential equations governing extremals. One now has to ascertain which extremals, if any, correspond to optimal solutions. Sometimes this can be done "by hand." However, very often this is not feasible. However, much work has been done on determining sufficient conditions, so one can look into the literature on such things.

Exercises

E3.1 Consider a control system $\Sigma = (\mathcal{X}, f, U)$ that is control-affine:

$$f(x, u) = f_0(x) + f_1(x) \cdot u.$$

Make the following assumptions:

(i) $U = \mathbb{R}^m$;

(ii) for each $x \in \mathcal{X}$ the linear map $f_1(x)$ is an isomorphism.

Answer the following questions.

(a) Show that if $L\colon \mathcal{X} \times U \to$ is a Lagrangian then there is a naturally induced Lagrangian $\tilde{L}\colon \mathcal{X} \times \mathbb{R}^n \to \mathbb{R}$.

(b) Show that the solutions of class C^2 of Problem 1.8 for the system Σ and the Lagrangian L coincide with solutions of Problem 2.1 for the Lagrangian $\tilde{L}$.

(c) Argue that the above conclusions almost show how Theorems 2.2, 2.4, and 2.7 in the calculus of variations are special cases of the Maximum Principle. Why does it not completely show this?

The following exercise considers a sort of finite-dimensional version of the Maximum Principle, indicating how, in the finite-dimensional setting, the constant λ_*^0 can arise. This rests on the following so-called Lagrange Multiplier Theorem.

3.9 Theorem: (Lagrange Multiplier Theorem) *Let $\mathcal{U} \subset \mathbb{R}^n$ be an open subset and let $f\colon \mathcal{U} \to \mathbb{R}$ and $g\colon \mathcal{U} \to \mathbb{R}^m$ be of class C^1 with $m < n$. If x_0 is a local minimum of $f|g^{-1}(0)$ then there exist $\lambda^0 \in \mathbb{R}$ and $\lambda \in \mathbb{R}^m$, not simultaneously zero, so that x_0 is a critical point of*

$$f_{\lambda^0,\lambda}\colon x \mapsto \lambda^0 f(x) + \langle \lambda, g(x) \rangle.$$

Furthermore, if $\boldsymbol{D}g(x_0)$ is surjective then $\lambda^0 \neq 0$. Conversely, if $x_0 \in g^{-1}(0)$ is a critical point of $f_{\lambda^0,\lambda}$ with $\lambda^0 = 0$ then $\boldsymbol{D}g(x_0)$ is not surjective.

E3.2 We consider the Lagrange Multiplier Theorem in a couple of special cases. In each case, use the theorem to find the minimum of $f|g^{-1}(0)$, and comment on the rôle of λ^0.

(a) Let $f\colon \mathbb{R}^2 \to \mathbb{R}$ and $g\colon \mathbb{R}^2 \to \mathbb{R}$ be defined by

$$f(x^1, x^2) = x^1 + x^2, \quad g(x^1, x^2) = (x^1)^2 + (x^2)^2 - 1.$$

(b) Let $f\colon \mathbb{R}^2 \to \mathbb{R}$ and $g\colon \mathbb{R}^2 \to \mathbb{R}$ be defined by

$$f(x^1, x^2) = x^1 + x^2, \quad g(x^1, x^2) = (x^1)^2 + (x^2)^2.$$

Chapter 4

Control variations

In this chapter we provide the technical details surrounding the construction of certain sorts of variations of trajectories of a control system. The idea of a control variation is of fundamental importance in control theory in general, and here we only skim the surface of what can be said and done. The basic idea of a control variation is that it allows one to explore the reachable set in the neighbourhood of a given trajectory by "wiggling" the trajectory. This is somewhat like what we explained in Section 2.2 in the calculus of variations. However, in control theory one can consider variations arising from two mechanisms. Analogously to Definition 2.8, one can fix a control and vary a trajectory, essentially by varying the initial condition. However, one can also vary the control in some way and measure the effects of this on the trajectory. The variations we discuss in this section are obtained by combining these mechanisms. Also, in control theory one often looks for families of variations that are closed under taking convex combinations. We see this idea in our development as well.

An interesting paper where the variational approach to controllability is developed is [Bianchini and Stefani 1993]. Here one can find a fairly general definition of a control variation. Let us remark that through the idea of a control variation one may arrive at the importance of the Lie bracket in control theory. This idea is explored in great depth in the geometric control literature; we refer to the books [Agrachev and Sachkov 2004, Jurdjevic 1997] for an introduction to the geometric point of view and for references to the literature. The book on mechanical control systems by Bullo and Lewis [2004] also contains an extensive bibliography concerning geometric control.

4.1. The variational and adjoint equations

As we saw in Section 2.1 in our proofs of the necessary conditions in the calculus of variations, it is useful to be able to "wiggle" a trajectory and measure the effects of this in some precise way. For control systems, a useful tool in performing such measurements is the variational equation. Associated to this, essentially through orthogonal complementarity, is the adjoint equation.

4.1 Definition: (Variational equation, adjoint equation) Let $\Sigma = (\mathcal{X}, f, U)$ be a control system and let $\mu\colon I \to U$ be an admissible control.

(i) The ***variational equation*** for Σ with control μ is the differential equation

$$\begin{aligned}\dot{\xi}(t) &= f(\xi(t), \mu(t)),\\ \dot{v}(t) &= \boldsymbol{D}_1 f(\xi(t), \mu(t)) \cdot v(t)\end{aligned}$$

on $\mathcal{X} \times \mathbb{R}^n$.

(ii) The ***adjoint equation*** for Σ with control μ is the differential equation

$$\begin{aligned}\dot{\xi}(t) &= f(\xi(t), \mu(t)),\\ \dot{\lambda}(t) &= -\boldsymbol{D}_1 f^T(\xi(t), \mu(t)) \cdot \lambda(t)\end{aligned}$$

on $\mathcal{X} \times \mathbb{R}^n$. •

Let us indicate how one should think about these differential equations. For the variational equation the interpretation is quite direct, while the interpretation of the adjoint equation perhaps is most easily explained by relating it to the variational equation. Thus let us first consider the variational equation.

Note that, if one fixes the control $\mu\colon I \to \mathbb{R}^n$, then one can think of the differential equation

$$\dot{\xi}(t) = f(\xi(t), \mu(t)) \tag{4.1}$$

as simply being a nonautonomous differential equation. The nicest interpretation of the variational equation involves variations of solutions of (4.1).

4.2 Definition: (Variation of trajectory) Let $\Sigma = (\mathcal{X}, f, U)$, let $x_0 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$. A ***variation*** of the trajectory $\xi(\mu, x_0, t_0, \cdot)$ is a map $\sigma\colon J \times [t_0, t_1] \to \mathcal{X}$ such that

(i) $J \subset \mathbb{R}$ is an interval for which $0 \in \operatorname{int}(J)$,

(ii) $\sigma(0, t) = \xi(\mu, x_0, t_0, t)$ for each $t \in [t_0, t_1]$,

(iii) $s \mapsto \sigma(s, t)$ is of class C^1 for each $t \in [t_0, t_1]$, and

(iv) $t \mapsto \sigma(s, t)$ is a solution of (4.1).

For a variation σ of $\xi(\mu, x_0, t_0, \cdot)$, the corresponding ***infinitesimal variation*** is the map $\delta\sigma\colon [t_0, t_1] \to \mathbb{R}^n$ defined by

$$\delta\sigma(t) = \frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} \sigma(s, t). \qquad \bullet$$

The intuition of the notion of a variation and an infinitesimal variation can be gleaned from Figure 4.1.

The following result explains the connection between variations and the variational equation.

4.3 Proposition: (Infinitesimal variations are solutions to the variational equation and vice versa) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$. For a map $v\colon [t_0, t_1] \to \mathbb{R}^n$ the following statements are equivalent:*

(i) there exists a variation σ of $\xi(\mu, x_0, t_0, \cdot)$ such that $v = \delta\sigma$;

(ii) $t \mapsto (\xi(\mu, x_0, t_0, t), v(t))$ satisfies the variational equation.

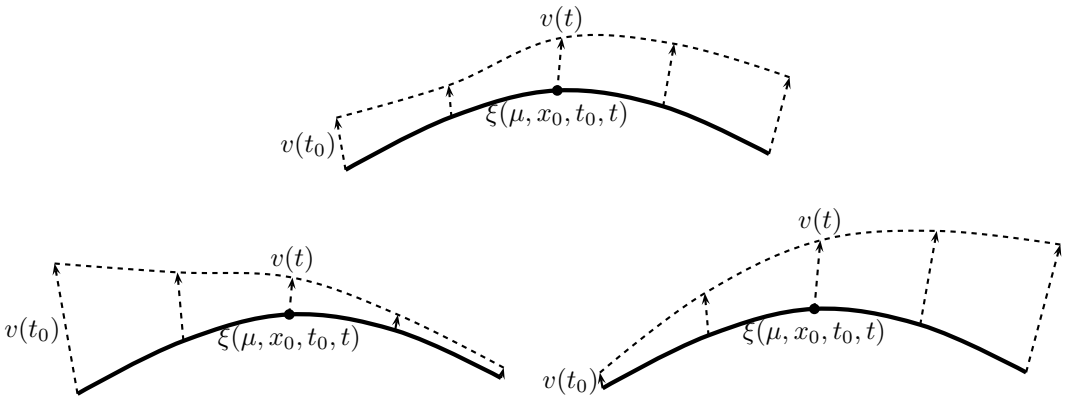

Figure 4.1. The idea of a variation of a trajectory is shown in the top figure. To further assist in the intuition of the variational equation, one can refer to the bottom two pictures; the one on the left shows a variation for a "stable" trajectory and the one of the right shows a variation for an "unstable" trajectory.

Proof: (i) $\Longrightarrow$ (ii) Suppose that σ has domain $J \times [t_0, t_1]$ and define $\gamma\colon J \to \mathcal{X}$ by $\gamma(s) = \sigma(s, t_0)$. Then we have

$$\sigma(s,t) = \xi(\mu, \gamma(s), t_0, t)$$

since each of the curves $t \mapsto \sigma(s,t)$ and $t \mapsto \xi(\mu, \gamma(s), t_0, t)$ is a solution to (4.1) with initial condition $\gamma(s)$ at time t_0. Therefore,

$$v(t) = \delta\sigma(t) = \frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} \sigma(s,t) = \boldsymbol{D}_2\xi(\mu, x_0, t_0, t) \cdot \delta\sigma(t_0).$$

Now let us define a linear map $\Phi(t)\colon \mathbb{R}^n \to \mathbb{R}^n$ by

$$\Phi(t) \cdot w = \boldsymbol{D}_2\xi(\mu, x_0, t_0, t) \cdot w.$$

Note that

$$\boldsymbol{D}_4\xi(\mu, x_0, t_0, t) = f(\xi(\mu, x_0, t_0, t), \mu(t))$$

so that, by the Chain Rule,

$$\begin{aligned}\boldsymbol{D}_4\boldsymbol{D}_2\xi(\mu, x_0, t_0, t) &= \boldsymbol{D}_1 f(\xi(\mu, x_0, t_0, t), \mu(t)) \circ \boldsymbol{D}_2\xi(\mu, x_0, t_0, t) \\ &= \boldsymbol{D}_1 f(\xi(\mu, x_0, t_0, t), \mu(t)) \circ \Phi(t).\end{aligned}$$

Thus Φ satisfies the matrix differential initial value problem

$$\dot{\Phi}(t) = \boldsymbol{D}_1 f(\xi(\mu, x_0, t_0, t), \mu(t)) \circ \Phi(t), \quad \Phi(t_0) = I_n.$$

Therefore, $v(t) = \Phi(t) \cdot \delta\sigma(t_0)$ is the solution of the vector initial value problem

$$\dot{v}(t) = \boldsymbol{D}_1 f(\xi(\mu, x_0, t_0, t), \mu(t)) \cdot v(t), \quad v(t_0) = \delta\sigma(t_0),$$

which gives this part of the proof.

(ii) $\implies$ (i) Let $s_0 > 0$, take $J = [-s_0, s_0]$, and let $\gamma\colon J \to \mathcal{X}$ be a curve such that $\frac{\mathrm{d}}{\mathrm{d}s}\big|_{s=0}\gamma(s) = v(t_0)$. By a compactness argument using continuity of solutions of (4.1), the details of which we leave to the reader, one can choose s_0 sufficiently small that $\xi(\mu, \gamma(s), t_0, \cdot)$ is defined on $[t_0, t_1]$ for each $s \in J$. We then define $\sigma(s,t) = \xi(\mu, \gamma(s), t_0, t)$. It is easy to see that σ is a variation. Moreover, using our computations from the first part of the proof we have $\delta\sigma(t) = v(t)$. ■

Motivated by the proof of the preceding proposition, for $\tau, t \in [t_0, t_1]$ we denote by $\Phi(\mu, x_0, t_0, \tau, t)$ the solution to the matrix initial value problem

$$\dot{\Phi}(t) = \boldsymbol{D}_1 f(\xi(\mu, x_0, t_0, t), \mu(t)) \circ \Phi(t), \quad \Phi(\tau) = I_n.$$

Geometrically we should think of $\Phi(\mu, x_0, t_0, \tau, t)$ as an isomorphism from the tangent space at $\xi(\mu, x_0, t_0, \tau)$ to the tangent space at $\xi(\mu, x_0, t_0, t)$; see Figure 4.2. In the proof we showed

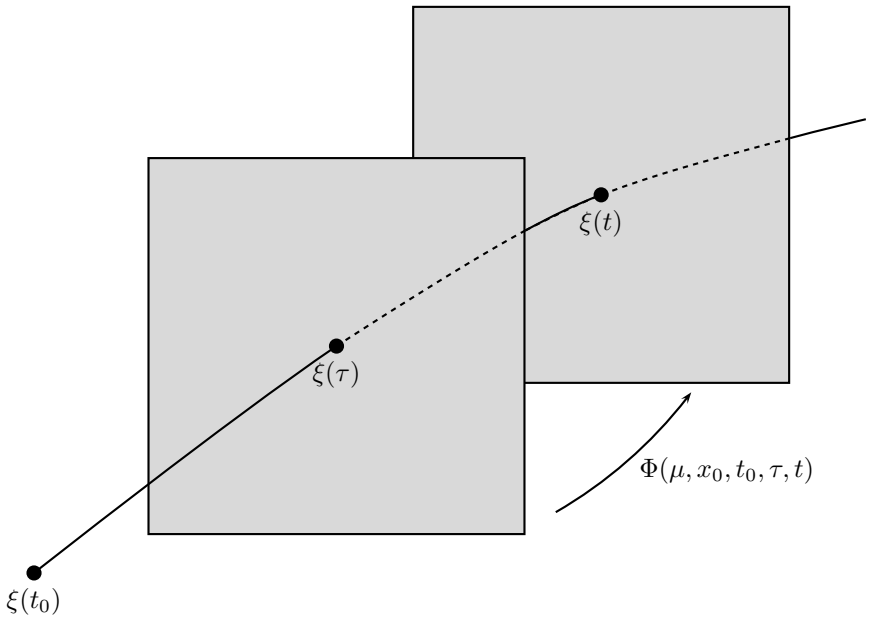

Figure 4.2. The interpretation of $\Phi(\mu, x_0, t_0, \tau, t)$; $\xi(\mu, x_0, t_0, \cdot)$ is abbreviated by ξ

that $t \mapsto v(t) = \Phi(\mu, x_0, t_0, t_0, t) \cdot v(t_0)$ is then the solution of the linear differential equation

$$\dot{v}(t) = \boldsymbol{D}_1 f(\xi(\mu, x_0, t_0, t), \mu(t)) \cdot v(t)$$

with initial condition $v(t_0)$ at time t_0. The intuition is that $t \mapsto \Phi(\mu, x_0, t_0, t_0, t) \cdot v$ indicates how the linearised flow of the differential equation (4.1) translates the vector v along the trajectory $\xi(\mu, x_0, t_0, \cdot)$. This is a sort of parallel transport, and is indeed the "Lie drag" for those familiar with some differential geometry. Using standard methods for linear differential equations one can verify the following:

1. $\Phi(\mu, x_0, t_0, \tau_1, \tau_1) = I_n$;

2. $\Phi(\mu, x_0, t_0, \tau_1, \tau_2) = \Phi(\mu, x_0, t_0, \tau_2, \tau_1)^{-1}$;

3. $\Phi(\mu, x_0, t_0, \tau_1, \tau_2) \circ \Phi(\mu, x_0, t_0, \tau_0, \tau_1) = \Phi(\mu, x_0, t_0, \tau_0, \tau_2)$.

The connection between the adjoint equation and things Hamiltonian is also of interest. We recall that for a system $\Sigma = (\mathcal{X}, f, U)$ we define the Hamiltonian

$$H_\Sigma(x, p, u) = \langle p, f(x, u)\rangle.$$

The following result, which is proved merely by a direct computation, relates this Hamiltonian with the adjoint equation.

4.4 Proposition: (A Hamiltonian interpretation of the adjoint equation) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system and let $\mu\colon I \to U$ be an admissible control. For maps $\xi\colon I \to \mathcal{X}$ and $\lambda\colon I \to \mathbb{R}^n$, the following statements are equivalent:*

(i) the curve $t \mapsto (\xi(t), \lambda(t))$ satisfies the adjoint equation;

(ii) the curve $t \mapsto (\xi(t), \lambda(t))$ satisfies the differential equation

$$\begin{aligned}\dot\xi(t) &= \boldsymbol{D}_2 H_\Sigma(\xi(t), \lambda(t), \mu(t)),\\ \dot\lambda(t) &= -\boldsymbol{D}_1 H_\Sigma(\xi(t), \lambda(t), \mu(t)).\end{aligned}$$

The equations in the second part of the proposition are ***Hamilton's equations*** for the time-dependent Hamiltonian $(t, (x, p)) \mapsto H_\Sigma(x, p, \mu(t))$. We discuss this a little bit more in Section 7.4.

For us, the most useful feature of the adjoint equation is its following relationship with the variational equation.

4.5 Proposition: (Property of the adjoint equation) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$. If $v\colon [t_0, t_1] \to \mathbb{R}^n$ and $\lambda\colon [t_0, t_1] \to \mathbb{R}^n$ are such that $t \mapsto (\xi(\mu, x_0, t_0, t), v(t))$ and $t \mapsto (\xi(\mu, x_0, t_0, t), \lambda(t))$ satisfy the variational and adjoint equations, respectively, then*

$$\langle \lambda(t), v(t)\rangle = \langle \lambda(t_0), v(t_0)\rangle$$

for all $t \in [t_0, t_1]$.

Proof: Abbreviate $\xi(t) = \xi(\mu, x_0, t_0, t)$. We compute

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\langle \lambda(t), v(t)\rangle &= \langle \dot\lambda(t), v(t)\rangle + \langle \lambda(t), \dot v(t)\rangle\\ &= -\langle \boldsymbol{D}_1 f^T(\xi(t), \mu(t)) \cdot \lambda(t), v(t)\rangle + \langle \lambda(t), \boldsymbol{D}_1 f(\xi(t), \mu(t)) \cdot v(t)\rangle = 0.\end{aligned}$$

The result now follows since λ and v are absolutely continuous, and so $t \mapsto \langle \lambda(t), \xi(t)\rangle$ is also absolutely continuous. ■

The following immediate consequence of the preceding result will be useful.

4.6 Corollary: (Hyperplanes and the variational and adjoint equations) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$.*

Let $P_0 \subset \mathbb{R}^n$ be a subspace of codimension one and let $\lambda_0 \in \mathbb{R}^n \setminus \{0\}$ be orthogonal to P_0. For $t \in [t_0, t_1]$, define $P_t \subset \mathbb{R}^n$ and $\lambda_t \in \mathbb{R}^n$ by asking that

$$P_t = \{v(t) \mid v\colon [t_0, t_1] \to \mathbb{R}^n \text{ satisfies the variational equation with } v(t_0) \in P_0\}$$

and that $t \mapsto \lambda_t$ is the solution of the adjoint equation with initial condition λ_0.

Then λ_t is orthogonal to P_t for every $t \in [t_0, t_1]$.

Thus one can think of the adjoint equation as describing the evolution of a hyperplane along the trajectory $\xi(\mu, x_0, t_0, \cdot)$.

The solutions of the adjoint equation can be characterised in terms of the solutions of the variational equation.

4.7 Proposition: (Solutions of the adjoint equation) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$, and let $\tau \in [t_0, t_0]$. Then the solution of the initial value problem*

$$\dot{\lambda}(t) = -\boldsymbol{D}_1 f^T(\xi(\mu, x_0, t_0, t), \mu(t)) \cdot \lambda(t), \quad \lambda(\tau) = \lambda_\tau$$

is

$$t \mapsto \lambda(t) = \Phi(\mu, x_0, t_0, t, \tau)^T \cdot \lambda_\tau.$$

Proof: Let $t \mapsto \Psi(\mu, x_0, t_0, \tau, t)$ be the solution to the matrix initial value problem

$$\boldsymbol{D}_5\Psi(\mu, x_0, t_0, \tau, t) = -\boldsymbol{D}_1 f^T(\xi(\mu, x_0, t_0, t), \mu(t)) \cdot \Psi(\mu, x_0, t_0, \tau, t), \quad \Psi(\mu, x_0, t_0, \tau, \tau) = I_n,$$

so that the solution to the initial value problem

$$\dot{\lambda}(t) = -\boldsymbol{D}_1 f(\xi(\mu, x_0, t_0, t), \mu(t)) \cdot \lambda(t), \quad \lambda(\tau) = \lambda_\tau,$$

is $t \mapsto \Psi(\mu, x_0, t_0, \tau, t) \cdot \lambda_\tau$. Now let $v_\tau \in \mathbb{R}^n$. Then, by Proposition 4.5, we have

$$\langle \Psi(\mu, x_0, t_0, \tau, t) \cdot \lambda_\tau, \Phi(\mu, x_0, t_0, \tau, t) \cdot v_\tau \rangle = \langle \lambda_\tau, v_\tau \rangle.$$

Since this must hold for every $\lambda_\tau, v_\tau \in \mathbb{R}^n$ we deduce that

$$\Psi(\mu, x_0, t_0, \tau, t) = \Phi(\mu, x_0, t_0, \tau, t)^{-T} = \Phi(\mu, x_0, t_0, t, \tau)^T,$$

giving the result. ■

4.2. Needle variations

As we shall see, the proof of the Maximum Principle involves approximating the reachable set with convex cones. In order to generate these approximations we will use specific variations of controls, called needle variations, that are conjured to exactly give the sort of approximation we need. There are two sorts of needle variations we will use, depending on whether we are considering the free interval or the fixed interval optimal control problem. We shall first examine in detail the control variations used for the fixed interval problem, and then after we understand that will "tack on" the additional data needed for the free interval problem.

So let us first consider needle variations for the fixed interval problem.

4.8 Definition: (Fixed interval needle variation) Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, let $x_0 \in \mathcal{X}$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$.

(i) ***Fixed interval needle variation data*** is a triple $\theta = (\tau_\theta, l_\theta, \omega_\theta)$ where
  (a) $\tau_\theta \in (t_0, t_1]$,
  (b) $l_\theta \in \mathbb{R}_{\geq 0}$, and
  (c) $\omega_\theta \in U$.

(ii) The ***control variation*** of μ associated to the fixed interval needle variation data $\theta = (\tau_\theta, l_\theta, \omega_\theta)$ is the map $\mu_\theta \colon J \times [t_0, t_1] \to U$ defined by

$$\mu_\theta(s,t) = \begin{cases} \omega_\theta, & t \in [\tau_\theta - s l_\theta, \tau_\theta], \\ \mu(t), & \text{otherwise}, \end{cases}$$

and where $J = [0, s_0]$ is an interval sufficiently small that $\mu_\theta(s, \cdot) \colon t \mapsto \mu_\theta(s,t)$ is an admissible control for each $s \in J$.

(iii) The ***fixed interval needle variation*** associated with the control μ, the trajectory $\xi(\mu, x_0, t_0, \cdot)$, and the fixed interval needle variation data $\theta = (\tau_\theta, l_\theta, \omega_\theta)$ is the element $v_\theta \in \mathbb{R}^n$ defined by

$$v_\theta = \frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} \xi(\mu_\theta(s, \cdot), x_0, t_0, \tau_\theta),$$

when the limit exists. •

In Figure 4.3 we depict how one might think of μ_θ. The precise reason why this is a

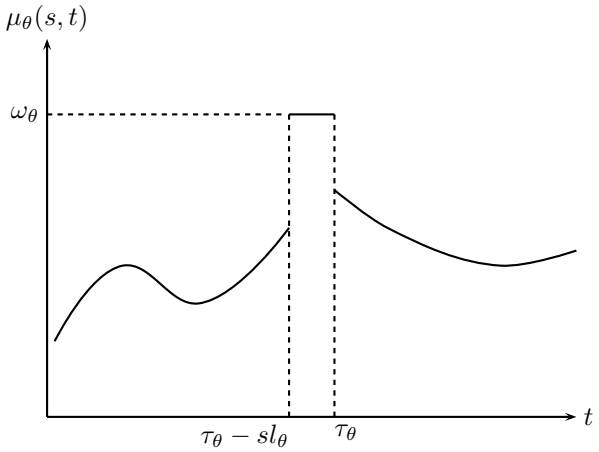

Figure 4.3. A depiction of the control variation associated to fixed interval needle variation data

useful thing to do only begins to become apparent in Section 5.4 when we approximate the reachable set with convex cones using needle variations.

The following result shows that fixed interval needle variations exist at Lebesgue points for $t \mapsto f(\xi(\mu, x_0, t_0, t), \mu(t))$. It will be convenient to denote by $\mathrm{Leb}(\mu, x_0, t_0, t)$ the set of Lebesgue points of $\tau \mapsto f(\xi(\mu, x_0, t_0, \tau), \mu(\tau))$ in the interval (t_0, t).

4.9 Proposition: (Existence of fixed interval needle variations) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$. If $\theta = (\tau_\theta, l_\theta, \omega_\theta)$ is fixed interval needle variation data with $\tau_\theta \in \mathrm{Leb}(\mu, x_0, t_0, t_1)$, then the fixed interval needle variation $v_\theta \in \mathbb{R}^n$ exists and is given by*

$$v_\theta = l_\theta(f(\xi(\mu, x_0, t_0, \tau_\theta), \omega_\theta) - f(\xi(\mu, x_0, t_0, \tau_\theta), \mu(\tau_\theta)).$$

Proof: Using the integral form of the differential equation we have

$$\begin{aligned}
\xi(\mu_\theta(s,\cdot), x_0, t_0, \tau_\theta) &= \xi(\mu_\theta(s,\cdot), x_0, t_0, \tau_\theta - sl_\theta) + \int_{\tau_\theta - sl_\theta}^{\tau_\theta} f(\xi(\mu_\theta(s,\cdot), x_0, t_0, t), \omega_\theta)\,\mathrm{d}t \\
&= \xi(\mu, x_0, t_0, \tau_\theta - sl_\theta) + \int_{\tau_\theta - sl_\theta}^{\tau_\theta} f(\xi(\mu_\theta(s,\cdot), x_0, t_0, t), \omega_\theta)\,\mathrm{d}t \\
&= \xi(\mu, x_0, t_0, \tau_\theta) - \int_{\tau_\theta - sl_\theta}^{\tau_\theta} f(\xi(\mu, x_0, t_0, t), \mu(t))\,\mathrm{d}t \\
&\quad + \int_{\tau_\theta - sl_\theta}^{\tau_\theta} f(\xi(\mu_\theta(s,\cdot), x_0, t_0, t), \omega_\theta)\,\mathrm{d}t.
\end{aligned}$$

Since $\tau_\theta \in \mathrm{Leb}(\mu, x_0, t_0, t_1)$ it holds that

$$\begin{aligned}
&\lim_{s\to 0} \frac{1}{s} \int_{\tau_\theta - sl_\theta}^{\tau_\theta} \|f(\xi(\mu, x_0, t_0, t), \mu(t)) - f(\xi(\mu, x_0, t_0, \tau_\theta), \mu(\tau_\theta))\|\,\mathrm{d}t = 0 \\
\implies \quad &\lim_{s\to 0} \frac{1}{s} \int_{\tau_\theta - sl_\theta}^{\tau_\theta} f(\xi(\mu, x_0, t_0, t), \mu(t))\,\mathrm{d}t = \lim_{s\to 0} \frac{1}{s} f(\xi(\mu, x_0, t_0, \tau_\theta), \mu(\tau_\theta)) sl_\theta \\
\implies \quad &\lim_{s\to 0} \frac{1}{s} \int_{\tau_\theta - sl_\theta}^{\tau_\theta} f(\xi(\mu, x_0, t_0, t), \mu(t))\,\mathrm{d}t = f(\xi(\mu, x_0, t_0, \tau_\theta), \mu(\tau_\theta)) l_\theta.
\end{aligned}$$

Since $t \mapsto \xi(\mu_\theta(s,\cdot), x_0, t_0, t)$ is continuous and since μ_θ is constant on $[\tau_\theta - sl_\theta, \tau_\theta]$,

$$\begin{aligned}
\lim_{s\to 0} \frac{1}{s} \int_{\tau_\theta - sl_\theta}^{\tau_\theta} f(\xi(\mu_\theta(s,\cdot), x_0, t_0, t), \omega_\theta)\,\mathrm{d}t &= f(\xi(\mu_\theta(0,\cdot), x_0, t_0, \tau_\theta), \omega_\theta) l_\theta \\
&= f(\xi(\mu, x_0, t_0, \tau_\theta), \omega_\theta) l_\theta.
\end{aligned}$$

Combining all of this gives

$$\frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} \xi(\mu_\theta(s,\cdot), x_0, t_0, \tau_\theta) = l_\theta(f(\xi(\mu, x_0, t_0, \tau_\theta), \omega_\theta) - f(\xi(\mu, x_0, t_0, \tau_\theta), \mu(\tau_\theta)),$$

as desired. ■

This gives the following useful property of the set of fixed interval needle variations at a point where they are defined.

4.10 Corollary: (The set of fixed interval needle variations is a cone) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$. If $\tau_\theta \in \mathrm{Leb}(\mu, x_0, t_0, t_1)$, then the set of fixed interval needle variations associated with fixed interval needle variation data of the form $\theta = (\tau_\theta, l_\theta, \omega_\theta)$ is a cone in $\mathbb{R}^n$.*

Proof: If v_θ is the fixed interval needle variation associated with the fixed interval needle variation data $(\tau_\theta, l_\theta, \omega_\theta)$ and if $\lambda \in \mathbb{R}_{\geq 0}$, then λv_θ is the fixed interval needle variation associated with the fixed interval needle variation data $(\tau_\theta, \lambda l_\theta, \omega_\theta)$. ■

Motivated by the corollary, if $\theta = (\tau_\theta, l_\theta, \omega_\theta)$ is fixed interval needle variation data and if $\lambda \in \mathbb{R}_{\geq 0}$, then we denote by $\lambda\theta$ the fixed interval needle variation data $\lambda\theta = (\tau_\theta, \lambda l_\theta, \omega_\theta)$. The corollary shows that, if the fixed interval needle variation associated with θ exists, then the fixed interval needle variation associated with $\lambda\theta$ exists, and we have $v_{\lambda\theta} = \lambda v_\theta$.

4.3. Multi-needle variations

Now we extend the analysis from the preceding section to allow for the effects of multiple needle variations. The reason for this is simple. The set of fixed interval needle variations at a given Lebesgue point form a cone by Corollary 4.10. However, these will not generally form a convex set, for example, because the control set U need not be convex (see Exercise E4.2). To generate a convex set of variations from needle variations we allow the times of the needle variations to vary.

Again the presentation breaks into fixed and free intervals cases, and we consider the fixed interval case in detail, saving for Section 4.4 the development needed for the free interval problem.

The notion needed in the fixed interval case is the following.

4.11 Definition: (Fixed interval multi-needle variation) Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, let $x_0 \in \mathcal{X}$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$.

(i) ***Fixed interval multi-needle variation data*** is a collection $\Theta = \{\theta_1, \dots, \theta_k\}$ of fixed interval needle variation data $\theta_j = (\tau_j, l_j, \omega_j)$, $j \in \{1, \dots, k\}$, such that the times $\tau_1, \dots, \tau_k$ are distinct.

(ii) The ***control variation*** of μ associated to the fixed interval multi-needle variation data $\Theta = \{\theta_1, \dots, \theta_k\}$ is the map $\mu_\Theta \colon J \times [t_0, t_1] \to U$ defined by

$$\mu_\Theta(s,t) = \begin{cases} \omega_j, & t \in [\tau_j - s l_j, \tau_j],\ j \in \{1, \dots, k\}, \\ \mu(t), & \text{otherwise}, \end{cases}$$

and where $J = [0, s_0]$ is an interval sufficiently small that $\mu_\Theta(s, \cdot) \colon t \mapsto \mu_\Theta(s,t)$ is a well-defined admissible control for each $s \in J$.

(iii) The ***fixed interval multi-needle variation*** associated with the control μ, the trajectory $\xi(\mu, x_0, t_0, \cdot)$, the time $t \in [t_0, t_1]$, $t > \tau_j$, $j \in \{1, \dots, k\}$, and the fixed interval multi-needle variation data $\Theta = \{\theta_1, \dots, \theta_k\}$ is the element $v_\Theta(t) \in \mathbb{R}^n$ defined by

$$v_\Theta(t) = \frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} \xi(\mu_\Theta(s, \cdot), x_0, t_0, t),$$

when the limit exists. •

The idea here is not much different than in Definition 4.8 except that we are performing the same sort of construction at multiple times. The following result gives conditions for the existence of fixed interval multi-needle variations.

4.12 Proposition: (Existence of fixed interval multi-needle variations) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, let $x_0 \in \mathcal{X}$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$. If $\Theta = \{\theta_1, \dots, \theta_k\}$ is fixed interval multi-needle variation date such that $\tau_1, \dots, \tau_k \in$*

$\mathrm{Leb}(\mu, x_0, t_0, t)$ *with* $\tau_j < t$, $j \in \{1, \dots, k\}$, *then the fixed interval multi-needle variation* $v_\Theta(t) \in \mathbb{R}^n$ *exists and is given by*

$$v_\Theta(t) = \Phi(\mu, x_0, t_0, \tau_1, t) \cdot v_{\theta_1} + \cdots + \Phi(\mu, x_0, t_0, \tau_k, t) \cdot v_{\theta_k}.$$

Proof: Just as in the proof of Proposition 4.9 we have

$$\xi(\mu_\Theta(s, \cdot), x_0, t_0, \tau_1) = \xi(\mu, x_0, t_0, \tau_1) + s v_{\theta_1} + o(s).$$

Since the control $\mu_\Theta(s, \cdot)$ on $[\tau_2 - s l_2, \tau_2]$ is just μ, it follows from Proposition 4.3 that

$$\xi(\mu_\Theta(s, \cdot), x_0, t_0, \tau_2 - s l_2) = \xi(\mu, x_0, t_0, \tau_2 - s l_2) + s\Phi(\mu, x_0, t_0, \tau_1, \tau_2 - s l_2) \cdot v_{\theta_1} + o(s).$$

Now one can perform a computation, just like that in the proof of Proposition 4.9, to show that

$$\xi(\mu_\Theta(s, \cdot), x_0, t_0, \tau_2) = \xi(\mu_\Theta(s, \cdot), x_0, t_0, \tau_2 - s l_2) + s l_2 f(\xi(\mu_\Theta(s, \cdot), x_0, t_0, \tau_2 - s l_2), \omega_2) + o(s).$$

Again following the computations in the proof of Proposition 4.9, and using the fact that τ_2 is a Lebesgue point, we have

$$\xi(\mu, x_0, t_0, \tau_2 - s l_2) = \xi(\mu, x_0, t_0, \tau_2) - s l_2 f(\xi(\mu, x_0, \tau_2), \mu(\tau_2)) + o(s).$$

Putting all of this together gives

$$\xi(\mu_\Theta(s, \cdot), x_0, t_0, \tau_2) = \xi(\mu, x_0, t_0, \tau_2) + s\Phi(\mu, x_0, t_0, \tau_1, \tau_2 - s l_2) \cdot v_{\theta_1} + s v_{\theta_2} + o(s).$$

Carrying on in this way (i.e., using induction) we arrive at

$$\begin{aligned}\xi(\mu_\Theta(s, \cdot), x_0, t_0, t) = \xi(\mu, x_0, t_0, t) + s\Phi(\mu, x_0, t_0, \tau_1, t) \cdot v_{\theta_1} + \dots \\ + s\Phi(\mu, x_0, t_0, \tau_k, t) \cdot v_{\theta_k} + o(s). \quad (4.2)\end{aligned}$$

Differentiation with respect to s at $s = 0$ gives the result. ■

One can see in the proof why it is that we require the times for a fixed interval multi-needle variation to be distinct. It is possible to consider different needle variations based at the same time, but one has to be more careful in computing the form of the variation. However, it is not necessary here to use this degree of generality.

One of the important observations one can make about a fixed interval multi-needle variation involves convex combinations. In order to express this, the following notation is helpful. If $\Theta = \{\theta_1, \dots, \theta_k\}$ is fixed interval multi-needle variation data and if $\lambda = \{\lambda_1, \dots, \lambda_k\} \subset \mathbb{R}_{\geq 0}$ then we denote $\lambda\Theta = \{\lambda_1\theta_1, \dots, \lambda_k\theta_k\}$, where we use the notation introduced following Corollary 4.10.

4.13 Corollary: (Coned convex combinations of fixed interval multi-needle variations) *Let* $\Sigma = (\mathcal{X}, f, U)$ *be a control system, let* $t_0, t_1 \in \mathbb{R}$ *satisfy* $t_0 < t_1$, *let* $x_0 \in \mathcal{X}$, *and let* $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$. *If* $\Theta = \{\theta_1, \dots, \theta_k\}$ *is a fixed interval multi-needle variation such that* $\tau_1, \dots, \tau_k \in \mathrm{Leb}(\mu, x_0, t_0, t)$ *with* $t > \tau_j$, $j \in \{1, \dots, k\}$, *and if* $\lambda = \{\lambda_1, \dots, \lambda_k\} \subset \mathbb{R}_{\geq 0}$ *then*

$$v_{\lambda\Theta}(t) = \lambda_1\Phi(\mu, x_0, t_0, \tau_1, t) \cdot v_{\theta_1} + \cdots + \lambda_k\Phi(\mu, x_0, t_0, \tau_k, t) \cdot v_{\theta_k}.$$

Moreover, if

$$\sum_{j=1}^{k} \lambda_j = 1,$$

then the limit $\frac{\mathrm{d}}{\mathrm{d}s}\big|_{s=0}\xi(\mu_{\lambda\Theta}(s,\cdot),x_0,t_0,t)$ *exists uniformly in* λ.

Proof: The first assertion follows immediately from the fact that $v_{\lambda_j\theta_j} = \lambda_j v_{\theta_j}$ for $j \in \{1,\dots,k\}$. The final assertion concerning uniform convergence with respect to λ follows from a compactness argument that we leave to the reader. ■

4.14 Remark: (Why needle variations?) Having now discussed a little bit about needle variations, we should really ask ourselves why these are useful to look at. Variations in general provide us with a way to look at trajectories "nearby" a given trajectory. Needle variations do this in a very specific way. What a needle variation does is isolate the effects of changing the control from its nominal value to a different value around a single instant. The notion of a multi-needle variation encodes the effects of doing this at various different times. Thus the way to think of the set of multi-needle variations is this: It represents the effects at a given time t of instantaneously altering the value of the control around almost all times (specifically at Lebesgue points) preceding t. •

4.4. Free interval variations

Fixed interval multi-needle variations are defined relative to a control which specifies the time interval for the control variation. Now we allow the length of the time interval to vary. We jump straight to multi-needle variations.

4.15 Definition: (Free interval multi-needle variation) Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, let $x_0 \in \mathcal{X}$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$.

(i) ***Free interval multi-needle variation data*** is a pair (Θ, Ψ) where $\Theta = \{\theta_1, \dots, \theta_k\}$ is fixed interval multi-needle variation data and where $\Psi = (\tau, \delta\tau) \in [t_0, t_1]\times\mathbb{R}$ satisfies $\tau > \tau_j$, $j \in \{1,\dots,k\}$.

(ii) The ***control variation*** of μ associated to free interval multi-needle variation data (Θ, Ψ) is the map $(s,t) \mapsto \mu_{(\Theta,\Psi)}(s,t)$ defined by

$$\mu_{(\Theta,\Psi)}(s,t) = \begin{cases} \mu_\Theta(t), & t \in [t_0, \tau], \\ \mu(\tau), & t \in (\tau, \tau + s\delta\tau], \\ \mu_\Theta(t - s\delta\tau), & t \in (\tau + s\delta\tau, t_1 + s\delta\tau] \end{cases}$$

when $\delta\tau \geq 0$ and by

$$\mu_{(\Theta,\Psi)}(s,t) = \begin{cases} \mu_\Theta(t), & [t_0, \tau + s\delta\tau], \\ \mu_\Theta(t - s\delta\tau), & t \in (\tau + s\delta\tau, t_1 + s\delta\tau] \end{cases}$$

when $\delta\tau < 0$. (Note that the domain of $\mu_{(\Theta,\Psi)}(s,\cdot)$ depends on s.)

(iii) The ***free interval multi-needle variation*** associated with the control μ, the trajectory $\xi(\mu, x_0, t_0, \cdot)$, the time $t > \tau$, and the free interval multi-needle variation data (Θ, Ψ) is the element $v_{(\Theta,\Psi)} \in \mathbb{R}^n$ defined by

$$v_{(\Theta,\Psi)}(t) = \frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} \xi(\mu_{(\Theta,\Psi)}(s, \cdot), x_0, t_0, t),$$

when the limit exists. •

In Figure 4.4 we show a free interval multi-needle variation for $\delta\tau > 0$. When $\delta\tau < 0$

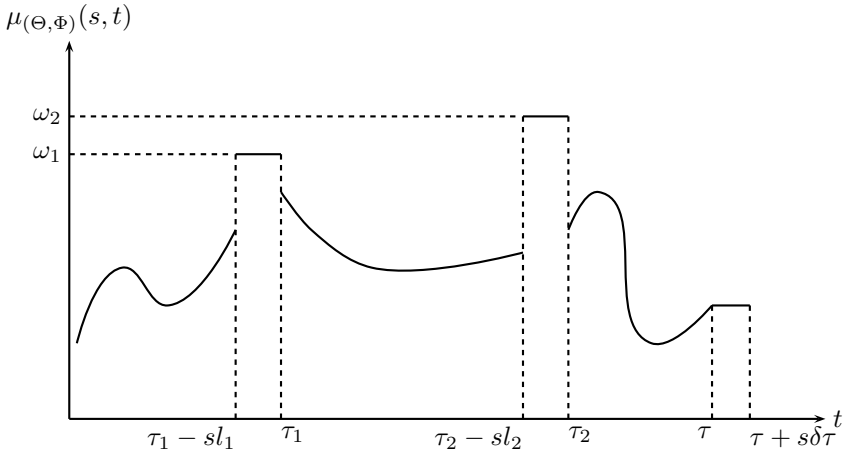

Figure 4.4. A free interval multi-needle variation for $\delta\tau > 0$

then the right endpoint of the interval simply gets shifted to the left, "deleting" that part of the control defined on $[\tau + s\delta\tau, \tau]$.

Let us do as we did for fixed interval multi-needle variations and show that free interval multi-needle variations exist with suitable assumptions on the data.

4.16 Proposition: (Existence of free interval multi-needle variations) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, let $x_0 \in \mathcal{X}$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$. If (Θ, Ψ) is free interval multi-needle variation data with $\Theta = \{\theta_1, \dots, \theta_k\}$ and $\Psi = (\tau, \delta\tau)$ such that $\tau_1, \dots, \tau_k, \tau \in \mathrm{Leb}(\mu, x_0, t_0, t)$, then the free interval multi-needle variation $v_{(\Theta,\Psi)} \in \mathbb{R}^n$ exists at $t > \tau$ and is given by*

$$v_{(\Theta,\Psi)}(t) = \Phi(\mu, x_0, t_0, \tau_1, t) \cdot v_{\theta_1} + \cdots + \Phi(\mu, x_0, t_0, \tau_k, t) \cdot v_{\theta_k} + \delta\tau\Phi(\mu, x_0, t_0, \tau, t) \cdot f(\xi(\mu, x_0, t_0, \tau), \mu(\tau)).$$

Proof: Since τ is a Lebesgue point for $t \mapsto f(\xi(\mu, x_0, t_0, t), \mu(t))$ it is also a Lebesgue point for $t \mapsto f(\xi(\mu_{(\Theta,\Psi)}(s, \cdot), x_0, t_0, t), \mu_{(\Theta,\Psi)}(t))$. Therefore, emulating the arguments made in the proof of Proposition 4.9, we obtain

$$\xi(\mu_{(\Theta,\Psi)}(s, \cdot), x_0, t_0, \tau + s\delta\tau) = \xi(\mu_{(\Theta,\Psi)}(s, \cdot), x_0, t_0, \tau) + s\delta\tau f(\xi(\mu_{(\Theta,\Psi)}(s, \cdot), x_0, t_0, \tau), \mu_{(\Theta,\Psi)}(s, \tau)) + o(s).$$

We also have, in similar fashion,

$$\xi(\mu, x_0, t_0, \tau + s\delta\tau) = \xi(\mu, x_0, t_0, \tau) + s\delta\tau f(\xi(\mu, x_0, t_0, \tau), \mu(\tau)) + o(s).$$

Now, by continuous dependence of solutions on initial conditions, for s sufficiently small there exists $g_s \colon [t_0, t_1] \to \mathbb{R}^n$ such that $\lim_{s\to 0} g_s(t) = 0$ uniformly in t and such that

$$f(\xi(\mu_{(\Theta,\Psi)}(s,\cdot), x_0, t_0, t), \mu_{(\Theta,\Psi)}(s,t)) = f(\xi(\mu, x_0, t_0, t), \mu(t)) + g_s(t), \qquad t \in [t_0, t_1].$$

Therefore,

$$\begin{aligned}\int_\tau^{\tau+s\delta\tau} f(\xi(\mu_{(\Theta,\Psi)}(s,\cdot), x_0, t_0, t), \mu_{(\Theta,\Psi)}(s,t))\,\mathrm{d}t &= \int_\tau^{\tau+s\delta\tau} f(\xi(\mu, x_0, t_0, t), \mu(t))\,\mathrm{d}t + o(s)\\ &= s\delta\tau f(\xi(\mu, x_0, t_0, \tau), \mu(\tau)) + o(s).\end{aligned}$$

Using the integral form of the differential equation then gives

$$\xi(\mu_{(\Theta,\Psi)}(s,\cdot), x_0, t_0, \tau + s\delta\tau) = \xi(\mu_{(\Theta,\Psi)}(s,\cdot), x_0, t_0, \tau) + s\delta\tau f(\xi(\mu, x_0, t_0, \tau), \mu(\tau)) + o(s).$$

From the proof of Proposition 4.12, particularly (4.2), we now have

$$\begin{aligned}\xi(\mu_{(\Theta,\Psi)}(s,\cdot), x_0, t_0, \tau + s\delta\tau) &= s\delta\tau f(\xi(\mu, x_0, t_0, \tau), \mu(\tau)) + \xi(\mu, x_0, t_0, \tau)\\ &\quad + s\Phi(\mu, x_0, t_0, \tau_1, \tau) \cdot v_{\theta_1} + \cdots + s\Phi(\mu, x_0, t_0, \tau_k, \tau) \cdot v_{\theta_k} + o(s).\end{aligned}$$

This then gives

$$\begin{aligned}\xi(\mu_{(\Theta,\Psi)}(s,\cdot), x_0, t_0, t + s\delta\tau) &= s\delta\tau\Phi(\mu, x_0, t_0, \tau, t) \cdot f(\xi(\mu, x_0, t_0, \tau), \mu(\tau)) + \xi(\mu, x_0, t_0, t)\\ &\quad + s\Phi(\mu, x_0, t_0, \tau_1, t) \cdot v_{\theta_1} + \cdots + s\Phi(\mu, x_0, t_0, \tau_k, t) \cdot v_{\theta_k} + o(s).\end{aligned}$$

Differentiation with respect to s at $s = 0$ now gives the result. ■

As with fixed interval multi-needle variations, free interval multi-needle variations are closed under coned convex combinations. To generate the notation to succinctly express this, let (Θ, Ψ) be free interval multi-needle variation data with $\Theta = \{\theta_1, \ldots, \theta_k\}$ and $\Psi = (\tau, \delta\tau)$ and let $\lambda = \{\lambda_0, \lambda_1, \ldots, \lambda_k\} \subset \mathbb{R}_{\geq 0}$. We then take $\lambda(\Theta, \Psi)$ to be the free interval multi-needle variation data $(\{\lambda_1\theta_1, \ldots, \lambda_k\theta_k\}, (\tau, \lambda_0\delta\tau))$. We then have the following result.

4.17 Corollary: (Coned convex combinations of free interval multi-needle variations) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, let $x_0 \in \mathcal{X}$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$. If (Θ, Ψ) is free interval multi-needle variation data with $\Theta = \{\theta_1, \ldots, \theta_k\}$ and $\Psi = (\tau, \delta\tau)$ such that $\tau_1, \ldots, \tau_k, \tau \in \mathrm{Leb}(\mu, x_0, t_0, t)$, and if $\lambda = \{\lambda_0, \lambda_1, \ldots, \lambda_k\} \subset \mathbb{R}_{\geq 0}$, then*

$$\begin{aligned}v_{\lambda(\Theta,\Psi)}(t) &= \lambda_1\Phi(\mu, x_0, t_0, \tau_1, t) \cdot v_{\theta_1} + \cdots + \lambda_k\Phi(\mu, x_0, t_0, \tau_k, t) \cdot v_{\theta_1}\\ &\quad + \lambda_0\delta\tau\Phi(\mu, x_0, t_0, \tau, t) \cdot f(\xi(\mu, x_0, t_0, \tau), \mu(\tau)).\end{aligned}$$

Proof: This exactly follows as does the proof of Corollary 4.13. ■

Exercises

E4.1 Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, let $x_0 \in \mathcal{X}$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$. For $\tau_0, \tau_1, \tau_2 \in [t_0, t_1]$ show that

1. $\Phi(\mu, x_0, t_0, \tau_1, \tau_1) = I_n$,
2. $\Phi(\mu, x_0, t_0, \tau_1, \tau_2) = \Phi(\mu, x_0, t_0, \tau_2, \tau_1)^{-1}$, and
3. $\Phi(\mu, x_0, t_0, \tau_1, \tau_2) \circ \Phi(\mu, x_0, t_0, \tau_0, \tau_1) = \Phi(\mu, x_0, t_0, \tau_0, \tau_2)$.

E4.2 Let $\mathcal{X} = \mathbb{R}^2$, let $m = 2$, let

$$U = \{(-1, 0), (0, -1), (0, 0), (1, 0), (0, 1)\},$$

and define

$$f((x^1, x^2), (u^1, u^2)) = (u^1, u^2).$$

Define $u(t) = (0, 0)$ for $t \in [t_0, t_1]$. Show that the set of fixed interval needle variations at any point $t \in [t_0, t_1]$ is not convex.

Chapter 5

The reachable set and approximation of its boundary by cones

In this chapter we carry out one of the important steps in understanding the Maximum Principle: the elucidation of the relationship between convex cones of multi-needle variations with the boundary of the reachable set. The development here is a little intricate and so will take some time to absorb on a first encounter.

The reader should be warned that the material in this chapter is best understood while forgetting about the fact that we are interested in optimal control. Only in Chapter 6 will we bring the optimal control problem back into the picture. Thus this chapter is dealing with the "in control" rather than the "in optimal control" aspect of the title for these notes.

5.1. Definitions

In this section we define precisely the various notions of reachable set that we shall use.

5.1 Definition: (Reachable set) Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, and let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$.

(i) The ***reachable set*** from x_0 at t_0 in time $t_1 - t_0$ is

$$\mathcal{R}(x_0, t_0, t_1) = \{\xi(\mu, x_0, t_0, t_1) \mid \mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])\}.$$

(ii) The ***reachable set*** from x_0 at t_0 is

$$\mathcal{R}(x_0, t_0) = \cup_{t_1 \in [t_0, \infty)} \mathcal{R}(x_0, t_0, t_1). \qquad \bullet$$

Note that since f is time-independent,

$$\mathcal{R}(x_0, t_0, t_1) = \mathcal{R}(x_0, 0, t_1 - t_0)$$

(with similar statements holding for all other variants of the reachable set). However, it will be necessary to sometimes consider trajectories naturally defined as starting at nonzero times, so we allow arbitrary initial times in our definition.

As we shall see, the Maximum Principle has a lot to do with the reachable set. Before we get to this, we need to define the various flavours of cones that we shall use to approximate the reachable set at its boundary.

5.2. The fixed interval tangent cone

We saw in Corollary 4.10 that the fixed interval needle variations at a Lebesgue point for $t \mapsto f(\xi(\mu, x_0, t_0, t), \mu(t))$ form a cone. We wish to consider unions of needle variations over all Lebesgue points, so let us establish some notation for this.

5.2 Definition: (Fixed interval tangent cone) Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$. For $t \in [t_0, t_1]$ we denote by $K(\mu, x_0, t_0, t)$ the closure of the coned convex hull of the set

$$\bigcup \{\Phi(\mu, x_0, t_0, \tau, t) \cdot v \mid \tau \in \mathrm{Leb}(\mu, x_0, t_0, t),\ v \text{ is a fixed interval needle variation at } \tau\}.$$

We call $K(\mu, x_0, t_0, t)$ the ***fixed interval tangent cone*** at t. •

The idea of the fixed interval tangent cone is that it should be a set of "directions" from which trajectories of the system emanate. This will be made precise in Lemma 5.10 below. In proving that lemma a key rôle will be played by the following notion.

5.3 Definition: (Fixed interval tangent simplex cone) Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$.

(i) ***Fixed interval tangent r-simplex cone data*** is a collection $\{\Theta_1, \dots, \Theta_r\}$ of fixed interval multi-needle variation data such that:
   (a) $\Theta_a = \{\theta_{a1}, \dots, \theta_{ak_a}\}$ (so defining the notation for the fixed interval needle variation data θ_{aj}, $j \in \{1, \dots, k_a\}$, $a \in \{1, \dots, r\}$);
   (b) $\theta_{aj} = (\tau_{aj}, l_{aj}, \omega_{aj})$, $j \in \{1, \dots, k_a\}$, $a \in \{1, \dots, r\}$ (so defining the notation $(\tau_{aj}, l_{aj}, \omega_{aj})$, $j \in \{1, \dots, k_a\}$, $a \in \{1, \dots, r\}$);
   (c) the times τ_{aj}, $j \in \{1, \dots, k_a\}$, $a \in \{1, \dots, r\}$, are distinct and all in $\mathrm{Leb}(\mu, x_0, t_0, t_1)$;
   (d) if $v_{\Theta_a}(t)$, $a \in \{1, \dots, r\}$, are the fixed interval multi-needle variations associated to Θ_a, $a \in \{1, \dots, r\}$, at time $t > \tau_{aj}$, $j \in \{1, \dots, k_a\}$, $a \in \{1, \dots, r\}$, then the coned convex hull of $\{v_{\Theta_1}, \dots, v_{\Theta_r}\}$ is an r-simplex cone.

(ii) The r-simplex cone defined by fixed interval tangent r-simplex cone data $\{\Theta_1, \dots, \Theta_r\}$ is a ***fixed interval tangent r-simplex cone*** at time t. •

The usefulness of fixed interval tangent simplex cones comes from the following simple result.

5.4 Lemma: (Property of fixed interval tangent simplex cones) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$. Let $\Theta = \{\Theta_1, \dots, \Theta_r\}$ be a fixed interval tangent r-simplex cone data, let $\lambda = \{\lambda_1, \dots, \lambda_r\} \subset \mathbb{R}_{\geq 0}$, and let $t > \tau_{aj}$, $j \in \{1, \dots, k_a\}$, $a \in \{1, \dots, r\}$. Denote by $v_{\Theta_a}(t)$, $a \in \{1, \dots, r\}$, the fixed interval multi-needle variations at time t.*

Then, for $s_0 \in \mathbb{R}_{>0}$ sufficiently small and for each $s \in [0, s_0]$, there exists a control $\mu_{\lambda,\Theta}(s,\cdot) \in \mathscr{U}(x_0, t_0, [t_0, t_1])$ such that

$$\frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} \xi(\mu_{\lambda,\Theta}(s,\cdot), x_0, t_0, t) = \lambda_1 v_{\Theta_1}(t) + \cdots + \lambda_r v_{\Theta_r}(t).$$

Proof: This follows from Corollary 4.13 since $\{\theta_{a,j} \mid j \in \{1,\dots,k_a\},\ a \in \{1,\dots,r\}\}$ is a fixed interval multi-needle variation, this itself being a consequence of the distinctness of the times τ_{aj}, $j \in \{1,\dots,k_a\}$, $a \in \{1,\dots,r\}$. ■

The point is that all directions in $K(\mu, x_0, t_0, t)$ that are contained in fixed interval tangent simplex cones are generated by a fixed interval multi-needle variation. The nonobvious thing is that all interior points in $K(\mu, x_0, t_0, t)$ have this property. The following result records this, along with another useful characterisation of the fixed interval tangent cone.

5.5 Lemma: (Characterisations of fixed interval tangent cone) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$. For a subset $C \subset \mathbb{R}^n$ and for $t \in [t_0, t_1]$ the following statements are equivalent:*

(i) $C = K(\mu, x_0, t_0, t)$;

(ii) *C is the closure of the union of the set of fixed interval tangent r-simplex cones at time t where $r = \dim(K(\mu, x_0, t_0, t))$;*

(iii) $C = \mathrm{cl}\big(\bigcup \{\Phi(\mu, x_0, t_0, \tau, t) \cdot K(\mu, x_0, t_0, \tau) \mid \tau \in \mathrm{Leb}(\mu, x_0, t_0, t)\}\big)$.

Proof: (ii)$\subset$(i) Note that any fixed interval tangent simplex cone at time t is, by definition, contained in $K(\mu, x_0, t_0, t)$. Thus the closure of the set of all fixed interval tangent simplex cones at time t is contained in $K(\mu, x_0, t_0, t)$ since the latter is closed.

(i)$\subset$(ii) Let $r = \dim(K(\mu, x_0, t_0, t))$ and let $v \in \mathrm{rel\,int}(K(\mu, x_0, t_0, t))$. By Proposition B.17 there exists $v_1, \dots, v_r \in K(\mu, x_0, t_0, t)$ such that $\mathrm{conv\,cone}(\{v_1, \dots, v_r\})$ is an r-simplex cone with v in its relative interior. Thus there exists $\lambda_1, \dots, \lambda_r \in \mathbb{R}_{>0}$ such that

$$v = \lambda_1 v_1 + \cdots + \lambda_r v_r.$$

Let $a \in \{1,\dots,r\}$. Since $v_a \in K(\mu, x_0, t_0, t)$ there exists fixed interval needle variation data $\theta_{aj} = (\tau_{aj}, l_{aj}, \omega_{aj})$ and $\lambda_{aj} \in \mathbb{R}_{>0}$, $j \in \{1,\dots,k_a\}$, $a \in \{1,\dots,r\}$, such that

$$v_a = \lambda_{a1}\Phi(\mu, x_0, t_0, \tau_{\theta_{a1}}, t) \cdot v_{\theta_{a1}} + \cdots + \lambda_{ak_a}\Phi(\mu, x_0, t_0, \tau_{\theta_{ak_a}}, t) \cdot v_{\theta_{ak_a}}.$$

Now take $\Theta_a = \{\lambda_{a1}\theta_{a1}, \dots, \lambda_{ak_a}\theta_{ak_a}\}$ and note that $v_a = v_{\Theta_a}(t)$.

If the times τ_{aj}, $j \in \{1,\dots,k_a\}$, $a \in \{1,\dots,r\}$, are distinct then $\{\Theta_1, \dots, \Theta_r\}$ is fixed interval tangent r-simplex cone data and

$$v \in \mathrm{int}(\mathrm{conv\,cone}(\{v_{\Theta_1}(t), \dots, v_{\Theta_r}(t)\})).$$

However, there is no reason for the times τ_{aj}, $j \in \{1,\dots,k_a\}$, $a \in \{1,\dots,r\}$, to be distinct. So let $\epsilon > 0$. The times τ_{aj}, $j \in \{1,\dots,k_a\}$, $a \in \{1,\dots,r\}$, can be modified to be distinct Lebesgue points τ'_{aj}, $j \in \{1,\dots,k_a\}$, $a \in \{1,\dots,r\}$, in such a way that

1. the total difference between the original times and the modified times is at most ϵ and
2. the total of the differences $\|f(\xi(\tau'_{aj}), \mu(\tau'_{aj})) - f(\xi(\tau_{aj}), \mu(t_{aj}))\|$ is at most ϵ.

From this it follows that $v_{\theta_{aj}}$ and $v_{\theta'_{aj}}$ can also be made arbitrarily close. Since $t \mapsto \Phi(\mu, x_0, t_0, \tau, t)$ is continuous in t it follows that $\Phi(\mu, x_0, t_0, \tau_{aj}, t) \cdot v_{\theta_{aj}}$ and $\Phi(\mu, x_0, t_0, \tau'_{aj}, t) \cdot v_{\theta'_{aj}}$ can be made arbitrarily close. From this we can assert that the times τ'_{aj}, $j \in \{1, \dots, k_a\}$, $a \in \{1, \dots, r\}$, can be made distinct in such a way that the points $v'_1, \dots, v'_r$ defined by

$$v'_a = \lambda_{a1}\Phi(\mu, x_0, t_0, \tau'_{\theta_{a1}}, t) \cdot v_{\theta'_{a1}} + \cdots + \lambda_{ak_a}\Phi(\mu, x_0, t_0, \tau'_{\theta_{ak_a}}, t) \cdot v_{\theta'_{ak_a}}, \qquad a \in \{1, \dots, r\},$$

are arbitrarily close to the points $v_1, \dots, v_r$. In particular, since $v \in \operatorname{rel\,int}(\operatorname{conv\,cone}(\{v_1, \dots, v_r\}))$ we also have $v \in \operatorname{rel\,int}(\operatorname{conv\,cone}(\{v'_1, \dots, v'_r\}))$. Thus we can write

$$v = \lambda'_1 v'_1 + \cdots + \lambda'_r v'_r$$

for $\lambda'_1, \dots, \lambda'_r \in \mathbb{R}_{>0}$. Then we may argue as above to define

$$\begin{aligned} \theta'_{aj} &= (\tau'_{aj}, \lambda'_{aj} l_{aj}, \omega'_{aj}), \qquad j \in \{1, \dots, k'_a\},\ a \in \{1, \dots, r\}, \\ \Theta'_a &= \{\theta'_{a1}, \dots, \theta'_{ak_a}\}, \qquad a \in \{1, \dots, r\}, \end{aligned}$$

with the property that $v'_a = v_{\Theta'_a}(t)$, $a \in \{1, \dots, r\}$. Therefore,

$$v \in \operatorname{rel\,int}(\operatorname{conv\,cone}(\{v_{\Theta'_1}(t), \dots, v_{\Theta'_r}(t)\})).$$

Thus $\operatorname{rel\,int}(K(\mu, x_0, t_0, t))$ is contained in union of the fixed interval tangent r-simplex cones, and thus the closure of the former is contained in the closure of the latter, as desired.

(i)$\subset$(iii) It is clear from the definition of the fixed interval tangent cone and from Exercise E4.1 that

$$\Phi(\mu, x_0, t_0, \tau, t) \cdot K(\mu, x_0, t_0, \tau) \subset K(\mu, x_0, t_0, t) \tag{5.1}$$

for each $\tau \in \operatorname{Leb}(\mu, x_0, t_0, t)$. More generally,

$$\Phi(\mu, x_0, t_0, \tau_1, t) \cdot K(\mu, x_0, t_0, \tau_1) \subset \Phi(\mu, x_0, t_0, \tau_2, t) \cdot K(\mu, x_0, t_0, \tau_2)$$

whenever $\tau_1, \tau_2 \in \operatorname{Leb}(\mu, x_0, t_0, t)$ satisfy $\tau_1 < \tau_2$. This shows that the family of sets

$$\Phi(\mu, x_0, t_0, \tau, t) \cdot K(\mu, x_0, t_0, \tau), \qquad \tau \in \operatorname{Leb}(\mu, x_0, t_0, t),$$

is an increasing family of subsets of $K(\mu, x_0, t_0, t)$ with respect to inclusion and the natural order on $\operatorname{Leb}(\mu, x_0, t_0, t)$. Moreover, each member of this family of sets is a closed convex cone by Exercise EB.3, and so their union is a convex cone. Thus the coned convex hull of

$$\operatorname{cl}\Big(\bigcup \{\Phi(\mu, x_0, t_0, \tau, t) \cdot K(\mu, x_0, t_0, \tau) \mid \tau \in \operatorname{Leb}(\mu, x_0, t_0, t)\}\Big)$$

is a closed convex cone in $K(\mu, x_0, t_0, t)$ which contains the set

$$\bigcup \{\Phi(\mu, x_0, t_0, \tau, t) \cdot v \mid \tau \in \operatorname{Leb}(\mu, x_0, t_0, t),\ v \text{ is a fixed interval needle variation at } \tau\}.$$

Since $K(\mu, x_0, t_0, t)$ is the smallest such closed convex cone this shows that

$$K(\mu, x_0, t_0, t) \subset \operatorname{cl}\Big(\bigcup \{\Phi(\mu, x_0, t_0, \tau, t) \cdot K(\mu, x_0, t_0, \tau) \mid \tau \in \operatorname{Leb}(\mu, x_0, t_0, t)\}\Big).$$

(iii)$\subset$(i) This is clear from (5.1), along with the fact that $K(\mu, x_0, t_0, t)$ is closed. ■

5.3. The free interval tangent cone

The procedure for constructing the free interval tangent cone is a little different than that for the fixed interval tangent cone since we begin with multi-needle variations. But the end effect is the same and many of the details we omit since they follow closely the fixed interval constructions.

5.6 Definition: (Free interval tangent cone) Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$. For $t \in [t_0, t_1]$ we denote by $K^{\pm}(\mu, x_0, t_0, t)$ the closure of the coned convex hull of the set

$$\bigcup \{\Phi(\mu, x_0, t_0, \tau, t)\cdot v \mid \tau \in \mathrm{Leb}(\mu, x_0, t_0, t),\ v \text{ is a free interval multi-needle variation at } \tau\}.$$

We call $K^{\pm}(\mu, x_0, t_0, t)$ the ***free interval tangent cone*** at t. •

In order to elucidate the meaning of the free interval tangent cone, we will proceed as in the fixed interval case, and introduce the notion of a tangent simplex cone.

5.7 Definition: (Free interval tangent simplex cone) Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$.

(i) ***Free interval tangent r-simplex cone data*** is a collection $\{(\Theta_1, \Psi_1), \dots, (\Theta_r, \Psi_r)\}$ of free interval multi-needle variation data such that:

  (a) $\Theta_a = \{\theta_{a1}, \dots, \theta_{ak_a}\}$ and $\Psi_a = (\tau_a, \delta\tau_a)$, $a \in \{1, \dots, r\}$ (so defining the notation for the fixed interval needle variation data θ_{aj}, $j \in \{1, \dots, k_a\}$, $a \in \{1, \dots, r\}$, and the pairs $(\tau_a, \delta\tau_a)$, $a \in \{1, \dots, r\}$);

  (b) $\theta_{aj} = (\tau_{aj}, l_{aj}, \omega_{aj})$, $j \in \{1, \dots, k_a\}$, $a \in \{1, \dots, r\}$ (so defining the notation $(\tau_{aj}, l_{aj}, \omega_{aj})$, $j \in \{1, \dots, k_a\}$, $a \in \{1, \dots, r\}$);

  (c) the times τ_a, $a \in \{1, \dots, r\}$, and τ_{aj}, $j \in \{1, \dots, k_a\}$, $a \in \{1, \dots, r\}$, are distinct and all in $\mathrm{Leb}(\mu, x_0, t_0, t_1)$;

  (d) if $v_{(\Theta_a, \Psi_a)}(t)$, $a \in \{1, \dots, r\}$, are the free interval multi-needle variations associated to (Θ_a, Ψ_a), $a \in \{1, \dots, r\}$, at time $t > \tau_{aj}$, $j \in \{1, \dots, k_a\}$, $a \in \{1, \dots, r\}$, then the coned convex hull of $\{v_{(\Theta_1, \Psi_1)}(t), \dots, v_{(\Theta_r, \Psi_r)}(t)\}$ is an r-simplex cone.

(ii) The r-simplex cone defined by free interval tangent r-simplex cone data $\{(\Theta_1, \Psi_1), \dots, (\Theta_r, \Psi_r)\}$ is a ***free interval tangent r-simplex cone*** at time t. •

The following result clarifies the importance of free interval tangent simplex cones.

5.8 Lemma: (Property of free interval tangent simplex cones) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$. Let $(\Theta, \Psi) = \{(\Theta_1, \Psi_1), \dots, (\Theta_r, \Psi_r)\}$ be free interval tangent r-simplex cone data, let $\lambda = \{\lambda_1, \dots, \lambda_r\} \subset \mathbb{R}_{\geq 0}$, and let $t \in [t_0, t_1]$ satisfy $t > \tau_a$, $a \in \{1, \dots, r\}$ and $t > \tau_{aj}$, $j \in \{1, \dots, k_a\}$, $a \in \{1, \dots, r\}$. Denote by $v_{(\Theta_a, \Psi_a)}(t)$, $a \in \{1, \dots, r\}$, the free interval multi-needle variations at time t.*

Then, for $s_0 \in \mathbb{R}_{>0}$ sufficiently small and for each $s \in [0, s_0]$, there exists a control $\mu_{\lambda,(\Theta,\Psi)}(s, \cdot) \in \mathscr{U}(x_0, t_0, [t_0, t_1])$ such that

$$\frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} \xi(\mu_{\lambda,(\Theta,\Psi)}(s, \cdot), x_0, t_0, t) = \lambda_1 v_{(\Theta_1, \Psi_1)}(t) + \cdots + \lambda_r v_{(\Theta_r, \Psi_r)}(t).$$

Proof: This follows from Corollary 4.17, along with some (at this point) standard computations along the lines of those performed in the proofs of Propositions 4.9, 4.12, and 4.16. We leave the fairly tedious details to the reader. ■

The following lemma now shows the relationship between free interval tangent simplex cones and the free interval tangent cone, along with another useful characterisation of the latter.

5.9 Lemma: (Characterisations of free interval tangent cone) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$. For a subset $C \subset \mathbb{R}^n$ and for $t \in [t_0, t_1]$ the following statements are equivalent:*

(i) $C = K^{\pm}(\mu, x_0, t_0, t)$;

(ii) *C is the closure of the union of the set of free interval tangent r-simplex cones at time t where $r = \dim(K^{\pm}(\mu, x_0, t_0, t))$;*

(iii) $C = \mathrm{cl}\big(\bigcup \{\Phi(\mu, x_0, t_0, \tau, t) \cdot K^{\pm}(\mu, x_0, t_0, \tau) \mid \tau \in \mathrm{Leb}(\mu, x_0, t_0, t)\}\big)$.

Proof: (ii)⊂(i) The argument here is like that from the corresponding part of the proof of Lemma 5.5.

(i)⊂(ii) The argument here is like that from the corresponding part of the proof of Lemma 5.5.

(i)⊂(iii) We first claim that

$$\Phi(\mu, x_0, t_0, \tau_1, \tau_2) \cdot K^{\pm}(\mu, x_0, t_0, \tau_1) \subset K^{\pm}(\mu, x_0, t_0, \tau_2)$$

whenever $\tau_1, \tau_2 \in \mathrm{Leb}(\mu, x_0, t_0, t)$ satisfy $\tau_1 < \tau_2$. For brevity denote $\xi = \xi(\mu, x_0, t_0, \cdot)$. Since (5.1) holds we need only show that

$$\Phi(\mu, x_0, t_0, \tau_1, \tau_2) \cdot f(\xi(\tau_1), \mu(\tau_1)) \in K^{\pm}(\mu, x_0, t_0, \tau_2).$$

We do this by proving the following lemma which is of general interest in any case.

1 Lemma: *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$. Let $\tau_1, \tau_2 \in \mathrm{Leb}(\mu, x_0, t_0, t_1)$ then*

$$\Phi(\mu, x_0, t_0, \tau_1, \tau_2) \cdot f(\xi(\mu, x_0, t_0, \tau_1), \mu(\tau_1)) = f(\xi(\mu, x_0, t_0, \tau_2), \mu(\tau_2)).$$

Proof: Define $\gamma_{\tau_1}(s) = \xi(\mu, x_0, t_0, \tau_1 + s)$ for $s \in [-\epsilon, \epsilon]$ and $\epsilon \in \mathbb{R}_{>0}$ sufficiently small. Note that

$$\frac{\mathrm{d}}{\mathrm{d}s}\gamma_{\tau_1}(s) = \boldsymbol{D}_4\xi(\mu, x_0, t_0, \tau_1 + s) = f(\xi(\mu, x_0, t_0, \tau_1 + s), \mu(\tau_1 + s)),$$

provided that the derivative exists. Since τ_1 is a Lebesgue point we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0}\gamma_{\tau_1}(s) &= \lim_{s\to 0} s^{-1}(\gamma_{\tau_1}(s) - \gamma_{\tau_1}(0)) = \lim_{s\to 0} s^{-1}\int_0^s \frac{\mathrm{d}}{\mathrm{d}\sigma}\gamma_{\tau_1}(\sigma)\,\mathrm{d}\sigma \\ &= \int_{\tau_1}^{\tau_1+s} f(\xi(\mu, x_0, t_0, \tau_1 + \sigma), \mu(\tau_1 + \sigma))\,\mathrm{d}\sigma = f(\xi(\mu, x_0, t_0, \tau_1), \mu(\tau_1)).\end{aligned}$$

Thus γ_{τ_1} is differentiable at 0. Therefore, $\gamma_{\tau_2} \colon s \mapsto \xi(\mu, \gamma_{\tau_1}(s), \tau_1, \tau_2 + s)$ is also differentiable by the Chain Rule. Moreover, by the composition property of flows of differential equations we have

$$\xi(\mu, \gamma_{\tau_1}(s), \tau_1, \tau_2 + s) = \xi(\mu, x_0, t_0, \tau_2 + s).$$

Since τ_2 is a Lebesgue point we have, exactly as in our preceding computation,

$$\frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0}\gamma_{\tau_2}(s) = f(\xi(\mu, x_0, t_0, \tau_2), \mu(\tau_2)).$$

However, by Proposition 4.3 we also have

$$\frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0}\gamma_{\tau_2}(s) = \Phi(\mu, x_0, t_0, \tau_1, \tau_2)\frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0}\gamma_{\tau_1}(s) = \Phi(\mu, x_0, t_0, \tau_1, \tau_2)\cdot f(\xi(\mu, x_0, t_0, \tau_1), \mu(\tau_1)),$$

as desired. ▼

This part of the proposition now follows immediately.

(iii)⊂(i) The argument here is like that from the corresponding part of the proof of Lemma 5.5. ■

5.4. Approximations of the reachable set by cones

Now that we have introduced the cones that we shall use to approximate the reachable set, let us make precise the actual relationship between these cones and the reachable set.

5.4.1. Approximation by the fixed interval tangent cone. For the fixed interval case we have the following result.

5.10 Lemma: (Points interior to fixed interval tangent cones are "in" the fixed time reachable set) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$. If $v_0 \in \operatorname{int}(K(\mu, x_0, t_0, t))$ for $t \in [t_0, t_1]$, then there exists a cone $K \subset \operatorname{int}(K(\mu, x_0, t_0, t))$ and $r \in \mathbb{R}_{>0}$ such that*

(i) $v_0 \in \operatorname{int}(K)$ and

(ii) such that

$$\{\xi(\mu, x_0, t_0, t) + v \mid v \in K,\ \|v\| < r\} \subset \mathcal{R}(x_0, t_0, t).$$

Proof: In the proof we refer implicitly to the material in Section B.4. By Proposition B.17 let $v_1, \ldots, v_n \in K(\mu, x_0, t_0, t)$ be convex combinations of fixed interval multi-needle variations at t having the property that $v_0 \in \operatorname{int}(\operatorname{conv}\operatorname{cone}(\{v_1, \ldots, v_n\}))$. We may assume, as in the proof of Lemma 5.5, that the Lebesgue times for all fixed interval multi-needle variations contributing to the determination of $v_1, \ldots, v_n$ are distinct. Let $K' = \operatorname{conv}\operatorname{cone}(\{v_1, \ldots, v_n\})$. Since the fixed interval multi-needle variations form a cone we may assume that $v_1, \ldots, v_n$ lie in the hyperplane in $\mathbb{R}^n$ passing through v_0 and orthogonal to v_0. This hyperplane is then the affine hull of $\{v_1, \ldots, v_n\}$. We may define barycentric coordinates for K' which we denote by (l, λ). Thus every point $v \in K'$ can be written as

$$v = l(\lambda_1 v_1 + \cdots + \lambda_n v_n)$$

for some unique $(l, (\lambda_1, \ldots, \lambda_n))$ with $l \geq 0$ and $\lambda_1, \ldots, \lambda_n \in \mathbb{R}_{\geq 0}$ which sum to 1.

By Lemma 5.4, for each point $v \in K'$ with coordinates $(l, \lambda_1, \ldots, \lambda_n)$ and for $s > 0$ sufficiently small we can define a control $\mu_v(s, \cdot)$ such that

$$\begin{aligned}\xi(\mu_v(s,\cdot), x_0, t_0, t) &= \xi(\mu_v(s,\cdot), x_0, t_0, t) + sv + o(s)\\ &= \xi(\mu_v(s,\cdot), x_0, t_0, t) + sl(v)(\lambda_1(v)v_1 + \cdots + \lambda_n(v)v_n) + o(s).\end{aligned}$$

Let us denote by $s(v)$ the largest value of s such that the control $\mu_v(s,\cdot)$ explicitly defined in Definition 4.8 makes sense. If one considers the form of the control variation $\mu_v(s)$ as given in Definition 4.8, one sees that $s(\lambda v) = \lambda^{-1}s(v)$. This holds essentially because in the expression for the control variation associated with fixed interval needle variation data θ, s and l_θ appear only as sl_θ which is used to determine the length of the interval on which the control ω_θ is applied. Moreover, a similar examination yields $\mu_v(\lambda s,\cdot) = \mu_{\lambda v}(s,\cdot)$ for $\lambda \in \mathbb{R}_{>0}$ having the property that $\lambda s \in [0, s(v))$. The upshot of this is that, if we think of s as fixed and $l(v)$ sufficiently small, we can write

$$\xi(\mu_v(s,\cdot),x_0,t_0,t) = \xi(\mu_v(s,\cdot),x_0,t_0,t) + sl(v)(\lambda_1(v)v_1 + \cdots + \lambda_n(v)v_n) + o(l(v)),$$

i.e., we can swap the limit as $s \to 0$ for the limit as $l(v) \to 0$. Since $\lambda_1,\dots,\lambda_n$ take values in the standard n-simplex which is compact, we note that we do indeed have

$$\lim_{l(v)\to 0} l(v)^{-1}\big(\xi(\mu_v(s,\cdot),x_0,t_0,t) - \xi(\mu_v(s,\cdot),x_0,t_0,t)\big) = 0,$$

uniformly in $\lambda_1,\dots,\lambda_n$.

Now we do take $s_0 > 0$ as fixed and think of the map

$$v \mapsto s_0^{-1}\big(\xi(\mu_v(s,\cdot),x_0,t_0,t) - \xi(\mu,x_0,t_0,t)\big) = l(v)(\lambda_1(v)v_1 + \cdots + \lambda_n(v)v_n) + o(l(v)) \quad (5.2)$$

as a map from an appropriately small tip of the cone K' to $\mathbb{R}^n$; let us denote this tip by C'. The image of this map can be thought of as being a piece of the reachable set around the point $\xi(\mu,x_0,t_0,t)$ which has been shifted by $-\xi(\mu,x_0,t_0,t)$ and then scaled by s_0^{-1}. If $l(v)$ is sufficiently small, the image will be a subset of the half-space of $\mathbb{R}^n$ defined by $l > 0$. Thus a cone whose tip is in the image of (5.2) defines a cone with a vertex at $\xi(\mu,x_0,t_0,t)$ whose tip is contained in the reachable set. The lemma will thus be proved if we can show that there is a subcone of K' whose tip is contained in the image of the map (5.2).

Having used barycentric coordinates to define the map (5.2), we now discard them. Instead let us use as coordinates for $\mathbb{R}^n$ the coordinates (l,ρ) where l measures the distance from the hyperplane orthogonal to v_0 passing through the origin and ρ is a vector in this hyperplane. Thus we write a point $v \in \mathbb{R}^n$, either in K' or not, as

$$v = l(v)\frac{v_0}{\|v_0\|} + \rho(v).$$

In these coordinates the map (5.2) is given by

$$v \mapsto l(v)\frac{v_0}{\|v_0\|} + \rho(v) + o(l(v)).$$

Having introduced these coordinates, let us use them but to define the coordinates we shall actually use in our construction. These coordinates we denote by (l,r) where l is as previously and where $r = \frac{\rho}{l}$. Thus these coordinates are only valid in the half-space where $l > 0$, but this is all we shall be concerned with in any case. Let K'' be a subcone of K' with the property that in the coordinates (l,r) we have

$$K'' = \{(l,r) \mid \|r\| \le a\}$$

for some $a \in \mathbb{R}_{>0}$ sufficiently small. Let $P(a,b)$ be a tip of K'' represented in the coordinates (l,r) by

$$P(a,b) = \{(l,r) \mid \ l \in [0,b],\ \|r\| \in [0,a]\}$$

for $b \in \mathbb{R}_{>0}$ sufficiently small. The idea is that the cone K'' is represented by a "cylinder" in these coordinates, its tip being "blown up." In any case, the map (5.2) in these coordinates we denote by

$$(l,r) \mapsto (L(l,r), R(l,r)),$$

and we note that

$$L(l,r) = l + o(l), \quad R(l,r) = r + o(1).$$

By defining $L(0,r) = 0$ and $R(0,r) = r$ this gives a representation of the map (5.2) in the coordinates (l,r) as being a map f from the set $P(a,b)$ to the half-plane where $L > 0$. Since $\lim_{l\to 0}\|R(l,r) - r\| = 0$ uniformly in r, we can choose b sufficiently small that $\|R(l,r) - r\| < \frac{a}{4}$ for all $(l,r) \in P(a,b)$. Since $L(l,r) - l = o(l)$ uniformly in r, we have $\lim_{l\to 0}(l^{-1}L(l,r) - 1) = 0$ meaning that we can choose b sufficiently small that $|b^{-1}L(b,r) - 1| < \frac{1}{4}$ or $|L(b,r) - b| < \frac{b}{4}$. Let us further choose $a = 2b$. Now consider a point $(l_0, r_0) \in P(a,b)$ with $l_0 \in (0, \frac{b}{4})$ and $\|r_0\| < \frac{a}{4}$.

We will now use Lemma C.3 to show that $(l_0, r_0) \in \text{image}(f)$. To do so we must show that for all $(l,r) \in \text{bd}(P(a,b))$ we have

$$\|(L(l,r), R(l,r)) - (l,r)\| < \|(l,r) - (l_0,r_0)\|. \tag{5.3}$$

First let us consider boundary points of the form $(0,r)$. Here we simply have

$$\|(L(0,r), R(0,r)) - (0,r)\| = 0 < \|(0,r) - (l_0,r_0)\|,$$

and so (5.3) is immediately verified. Next consider boundary points of the form (b,r). Here we have

$$\|(L(b,r), R(b,r)) - (b,r)\|^2 = |L(b,r) - b|^2 + \|R(b,r) - r\|^2 < \frac{b^2}{16} + \frac{a^2}{16}$$

and

$$\|(b,r) - (l_0,r_0)\|^2 = |b - l_0|^2 + \|r - r_0\|^2 > \frac{9b^2}{16}.$$

Given our assumptions about a and b, the inequality (5.3) holds in this case as well. Finally, we consider the boundary points of the form (l,r) where $\|r\| = a$. In this case we have

$$\|(L(l,r), R(l,r)) - (l,r)\|^2 = |L(l,r) - l|^2 + \|R(l,r) - r\|^2 < \frac{b^2}{16} + \frac{a^2}{16}$$

and

$$\|(l,r) - (l_0,r_0)\|^2 = |l - l_0|^2 + \|r - r_0\|^2 > |\|r\| - \|r_0\|| > \frac{9a^2}{16}.$$

Again, given our assumptions on a and b, it follows that (5.3) holds. Thus the point (l_0, r_0) is in the image of the map f. Thus a cylinder around the line segment

$$\{(l,0) \mid \ |l| < \tfrac{b}{4}\}$$

is contained in the image of f. In the undeformed coordinates this cylinder corresponds to the "tip" of a convex cone, and so this proves the lemma.

In Figure 5.1 we illustrate the idea behind the last part of the proof. ■

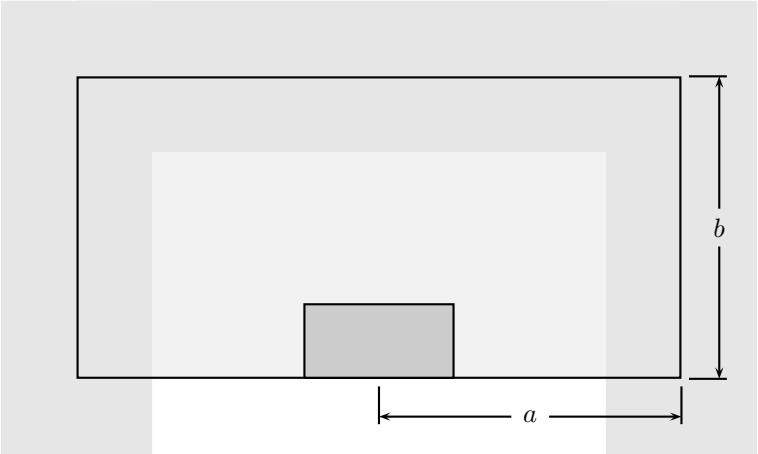

Figure 5.1. An illustration of the constructions in the last part of the proof of Lemma 5.10. The large box with the solid boundary represents $P(a, b)$. The dark shaded smaller box with the solid boundary represents where the points $(l_0, r_0) \in \text{image}(f)$ live. The medium shaded region represents where the boundary of $P(a, b)$ is mapped under f.

It is important to understand what the lemma is *not* saying, as this will bear on our subsequent results. Most importantly, the lemma says nothing when $\text{int}(K(\mu, x_0, t_0, t)) = \emptyset$. In particular, one cannot replace the condition that $v_0 \in \text{int}(K(\mu, x_0, t_0, t))$ with the condition that $v_0 \in \text{rel int}(K(\mu, x_0, t_0, t))$ as is demonstrated in Exercise E5.2. We shall have more to say about this somewhat important idea in Section 8.5 where we will encounter an instance where it comes up in a tangible way. For the moment let us merely say that for a reader wishing to understand the rôle of the Maximum Principle in control theory, this is a point that they would do well to appreciate.

5.4.2. Approximation by the free interval tangent cone. For the free interval case the result we desire is the following.

5.11 Lemma: (Points interior to free interval tangent cones are "in" the reachable set) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$. If $v_0 \in \text{int}(K^{\pm}(\mu, x_0, t_0, t))$ for $t \in [t_0, t_1]$ then there exists a cone $K \subset \text{int}(K^{\pm}(\mu, x_0, t_0, t))$ and $r \in \mathbb{R}_{>0}$ such that*

(i) $v_0 \in \text{int}(K)$ and

(ii) such that

$$\{x_0 + v \mid v \in K, \ \|v\| < r\} \subset \mathcal{R}(x_0, t_0).$$

Proof: This follows from Lemmata 5.8 and 5.9 in the same way that Lemma 5.10 follows from Lemmata 5.4 and 5.5. ■

5.5. The connection between tangent cones and the Hamiltonian

At a first glance, the appearance of the Hamiltonian in the statement of the Maximum Principle is rather mysterious. Its rôle is illuminated in the proof. However, since pointwise maximisation of the Hamiltonian is connected solely with the tangent cones, and since we have just now constructed these tangent cones, it seems as if there is no better time than the present to explain the relationship between the Hamiltonian and the tangent cones. The reader will note that the results in this section do not depend *a priori* on any sort of optimal control problem or on any statements about the reachable set; they are a consequence only of properties of tangent cones.

Recall from Definition 3.1 the notion of the Hamiltonian. The following completely trivial lemma actually gives significant insight into the rôle of the Hamiltonian and the maximum Hamiltonian.

5.12 Lemma: (A property of the maximum Hamiltonian) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system and let $(x, p, \bar{u}) \in \mathcal{X} \times \mathbb{R}^n \times U$. Then $H_\Sigma(x, p, \bar{u}) = H^{\max}_\Sigma(x, p)$ if and only if*

$$\langle p, v\rangle \leq 0, \qquad v \in \{f(x, u) - f(x, \bar{u}) \mid u \in U\}.$$

Proof: We have

$$\begin{aligned}
& H_\Sigma(x, p, \bar{u}) = H^{\max}(x, p), \\
\iff \quad & H_\Sigma(x, p, \bar{u}) \geq H_\Sigma(x, p, u), \qquad u \in U, \\
\iff \quad & H_\Sigma(x, p, \bar{u}) - H_\Sigma(x, p, u) \geq 0, \qquad u \in U, \\
\iff \quad & \langle p, f(x, u) - f(x, \bar{u})\rangle \leq 0, \qquad u \in U, \\
\iff \quad & \langle p, v\rangle \leq 0, \qquad v \in \{f(x, u) - f(x, \bar{u}) \mid u \in U\},
\end{aligned}$$

as desired. ■

For a system $\Sigma = (\mathcal{X}, f, U)$ and for $x \in \mathcal{X}$ let us denote

$$F_\Sigma(x) = \{f(x, u) \mid u \in U\} \subset \mathbb{R}^n.$$

With this notation, in Figure 5.2 we illustrate how one can think of the condition that the Hamiltonian be maximised: the vector $p \in \mathbb{R}^n$ is orthogonal to a support hyperplane for $F_\Sigma(x)$. The existence of such a p implies that the control maximising the Hamiltonian is on the boundary of $\operatorname{conv}(F_\Sigma(x))$.

The essentially insightful result is the following.

5.13 Lemma: (The Hamiltonian and the fixed interval tangent cone) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$. For each $t \in [t_0, t_1]$ let $K_t \subset \mathbb{R}^n$ be a convex cone such that $K(\mu, x_0, t_0, t) \subset K_t$, and suppose that at some $\tau \in [t_0, t_1]$ there exists $\lambda(\tau) \in \mathbb{R}^n$ such that*

$$\langle \lambda(\tau), v\rangle \leq 0, \qquad v \in K_\tau.$$

Let $t \mapsto \lambda(t)$ be the adjoint response for Σ along $(\xi(\mu, x_0, t_0, \cdot), \mu)$ equal to $\lambda(\tau)$ at time τ.

Then, for any $t \in \operatorname{Leb}(\mu, x_0, t_0, \tau)$, it holds that

$$H_\Sigma(\xi(\mu, x_0, t_0, t), \lambda(t), \mu(t)) = H^{\max}_\Sigma(\xi(\mu, x_0, t_0, t), \lambda(t)).$$

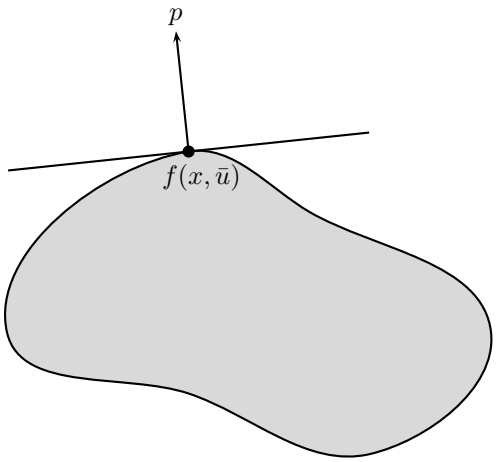

Figure 5.2. A depiction of the Hamiltonian maximisation condition: the shaded region represents $F_\Sigma(x)$

Proof: Let us abbreviate $\xi = \xi(\mu, x_0, t_0, \cdot)$. Let $t \in \text{Leb}(\mu, x_0, t_0, \tau)$ and consider the fixed interval needle variation data $(t, 1, \omega)$ for some $\omega \in U$. By Lemma 5.5 we then have

$$\Phi(\mu, x_0, t_0, t, \tau) \cdot \big(f(\xi(t), \omega) - f(\xi(t), \mu(t))\big) \in K(\mu, x_0, t_0, \tau) \subset K_\tau.$$

Therefore,

$$\langle \lambda(\tau), \Phi(\mu, x_0, t_0, t, \tau) \cdot f(\xi(t), \omega) \rangle - \langle \lambda(\tau), \Phi(\mu, x_0, t_0, t, \tau) \cdot f(\xi(t), \mu(t)) \rangle \leq 0.$$

Using Proposition 4.7 gives

$$\langle \lambda(t), f(\xi(t), \omega) \rangle - \langle \lambda(t), f(\xi(t), \mu(t)) \rangle \leq 0,$$

or, in what amounts to the same thing by Lemma 5.12,

$$H_\Sigma(\xi(t), \lambda(t), \omega) \leq H_\Sigma(\xi(t), \lambda(t), \mu(t)).$$

Since this must hold for every $\omega \in U$, the lemma follows. ■

The idea of the lemma is roughly this. The maximisation of the Hamiltonian almost everywhere in $[t_0, \tau]$ is implied by the existence of a support hyperplane for the free interval tangent cone at τ. The existence of such a support hyperplane is saying something about the control being "extremal" in some sense. Somewhat more precisely, by applying needle variations, trajectories at time τ can be made to move only "one way." Lemmata 5.12 and 5.13 say that this implies that the control must also have an "extremal" property at almost all instants preceding τ. (The preceding discussion really only has poignancy when the free interval tangent cone has a nonempty interior.)

The following elementary result is of a somewhat similar flavour to the preceding one in that it relates a property of the maximum Hamiltonian to a property of a family of cones along a controlled trajectory.

5.14 Lemma: (A condition for the maximum Hamiltonian to be zero) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1))$, and let $\tau \in [t_0, t_1]$. For each $t \in [t_0, t_1]$ let $K_t \subset \mathbb{R}^n$ be a convex cone such that $\operatorname{span}_{\mathbb{R}}(f(\xi(\mu, x_0, t_0, \tau), \mu(\tau))) \subset K_\tau$. Suppose that there exists a map $t \mapsto \lambda(t)$ such that*

$$\langle \lambda(\tau), v \rangle \leq 0, \qquad v \in K_\tau.$$

Then it holds that

$$H_\Sigma(\xi(\mu, x_0, t_0, \tau), \lambda(\tau), \mu(\tau)) = 0.$$

Proof: Since $\operatorname{span}_{\mathbb{R}}(f(\xi(\mu, x_0, t_0, \tau), \mu(\tau))) \subset K_\tau$ we have

$$\langle \lambda(\tau), \alpha f(\xi(\mu, x_0, t_0, \tau), \mu(\tau)) \rangle \leq 0$$

for every $\alpha \in \mathbb{R}$. Taking $\alpha = 1$ and then $\alpha = -1$ then gives

$$\langle \lambda(\tau), f(\xi(\mu, x_0, t_0, \tau), \mu(\tau)) \rangle = H_\Sigma(\xi(\mu, x_0, t_0, \tau), \lambda(\tau), \mu(\tau)) = 0,$$

as desired. ■

When the control is bounded, the conclusions of the preceding results can be strengthened to show that the maximum Hamiltonian is constant. We state this in the following form.

5.15 Lemma: (Constancy of the maximum Hamiltonian) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1]) \cap \mathscr{U}_{\mathrm{bdd}}([t_0, t_1])$. Suppose that $\lambda \colon [t_0, t_1] \to \mathbb{R}^n$ is an adjoint response for Σ along $(\xi(\mu, x_0, t_0, \cdot), \mu)$ satisfying*

$$H_\Sigma(\xi(\mu, x_0, t_0, t), \lambda(t), \mu(t)) = H_\Sigma^{\max}(\xi(\mu, x_0, t_0, t), \lambda(t))$$

for almost every $t \in [t_0, t_1]$. Then the function

$$t \mapsto H_\Sigma^{\max}(\xi(\mu, x_0, t_0, t), \lambda(t))$$

is constant.

Proof: Let us abbreviate $\xi = \xi(\mu, x_0, t_0, \cdot)$. Let $B = \mathrm{cl}(\mathrm{image}(\mu))$ so that $B \subset \mathrm{cl}(U)$ is then compact. Define $h_\Sigma^{\max} \colon \mathcal{X} \times \mathbb{R}^n \to \mathbb{R}$ by

$$h_\Sigma^{\max}(x, p) = \sup\{H_\Sigma(x, p, u) \mid u \in B\},$$

and note that $h_\Sigma^{\max}$ is bounded and that

$$h_\Sigma^{\max}(x, p) \leq H_\Sigma^{\max}(x, p) \tag{5.4}$$

for all $(x, p) \in \mathcal{X} \times \mathbb{R}^n$. Our hypotheses ensure that

$$h_\Sigma^{\max}(\xi(t), \lambda(t)) = H_\Sigma^{\max}(\xi(t), \lambda(t))$$

for almost every $t \in [t_0, t_1]$.

We now show that $t \mapsto h_\Sigma^{\max}(\xi(t), \lambda(t))$ is absolutely continuous. Since $x \mapsto f(x, u)$ is of class C^1 for each $u \in \mathrm{cl}(U)$ it follows that $(x, p) \mapsto H_\Sigma(x, p, u)$ is of class C^1 for each

$u \in \mathrm{cl}(U)$. Let $A \subset \mathcal{X} \times \mathbb{R}^n$ be compact and be such that $(\xi(t), \lambda(t)) \in A$ for every $t \in [t_0, t_1]$. Let $u \in B$. Then, following Lemma 1 in the proof of Theorem C.1, there exists $M \in \mathbb{R}_{>0}$ such that

$$|H_\Sigma(x, p, u) - H_\Sigma(x', p', u)| < M\|(x, p) - (x', p')\|$$

for all $(x, p), (x', p') \in A$. The constant M may be chosen independently of u since u lies in the compact set B. Now, by compactness of B and continuity of the map $u \mapsto H_\Sigma(x, p, u)$, let $u, u' \in B$ be such that

$$h_\Sigma^{\max}(x, p) = H_\Sigma(x, p, u), \quad h_\Sigma^{\max}(x', p') = H_\Sigma(x', p', u')$$

for $(x, p), (x', p') \in A$. Since

$$H_\Sigma(x, p, u') \le H_\Sigma(x, p, u), \quad H_\Sigma(x', p', u) \le H_\Sigma(x', p', u')$$

we have

$$\begin{aligned} -M\|(x, p) - (x', p')\| &\le H_\Sigma(x, p, u') - H_\Sigma(x', p', u') \\ &\le H_\Sigma(x, p, u) - H_\Sigma(x', p', u) \le M\|(x, p) - (x', p')\|. \end{aligned}$$

From this we conclude that

$$|h_\Sigma^{\max}(x, p) - h_\Sigma^{\max}(x', p')| \le M\|(x, p) - (x', p')\|,$$

so showing that $h_\Sigma^{\max}$ is Lipschitz. Recall that a function $t \mapsto f(t)$ on $[t_0, t_1]$ is absolutely continuous if and only if, for each $\epsilon \in \mathbb{R}_{>0}$, there exists $\delta \in \mathbb{R}_{>0}$ such that, if $\{(a_j, b_j)\}_{j \in \{1,\dots,k\}}$ is a finite collection of disjoint open intervals for which

$$\sum_{j=1}^{k} |b_j - a_j| < \delta, \tag{5.5}$$

then

$$\sum_{j=1}^{k} |f(b_j) - f(a_j)| < \epsilon.$$

We now show that $t \mapsto h_\Sigma^{\max}(\xi(t), \lambda(t))$ is absolutely continuous. Let $\epsilon \in \mathbb{R}_{>0}$ and take $\delta = M^{-1}\epsilon$. Then, if (5.5) is satisfied for a finite collection $\{(a_j, b_j)\}_{j \in \{1,\dots,k\}}$ of disjoint open intervals,

$$\sum_{j=1}^{k} |h_\Sigma^{\max}(b_j) - h_\Sigma^{\max}(a_j)| \le \sum_{j=1}^{k} M|b_j - a_j| < \epsilon.$$

Thus $t \mapsto h_\Sigma^{\max}(\xi(t), \lambda(t))$ is absolutely continuous, as desired.

Now we show that the derivative of $t \mapsto h_\Sigma^{\max}(\xi(t), \lambda(t))$ is almost everywhere zero, which will show that the function is constant. Suppose that $t' \in [t_0, t_1]$ is a point at which $t \mapsto h_\Sigma^{\max}(\xi(t), \lambda(t))$, $t \mapsto \xi(t)$, and $t \mapsto \lambda(t)$ are all differentiable. The complement of such t's has measure zero. For $t > t'$ we have

$$h_\Sigma^{\max}(\xi(t), \lambda(t)) \ge H_\Sigma(\xi(t), \lambda(t), \mu(t'))$$

so that

$$h_\Sigma^{\max}(\xi(t),\lambda(t)) - h_\Sigma^{\max}(\xi(t'),\lambda(t')) \\ \geq H_\Sigma(\xi(t),\lambda(t),\mu(t')) - H_\Sigma(\xi(t),\lambda(t'),\mu(t')) \\ + H_\Sigma(\xi(t),\lambda(t'),\mu(t')) - H_\Sigma(\xi(t'),\lambda(t'),\mu(t')).$$

Then

$$\frac{\mathrm{d}}{\mathrm{d}t}h_\Sigma^{\max}(\xi(t'),\lambda(t')) \geq \boldsymbol{D}_1 H_\Sigma(\xi(t'),\lambda(t'),\mu(t'))\cdot\dot{\xi}(t') + \boldsymbol{D}_2 H_\Sigma(\xi(t'),\lambda(t'),\mu(t'))\cdot\dot{\lambda}(t').$$

A direct computation using the definition of H_Σ then gives

$$\frac{\mathrm{d}}{\mathrm{d}t}h_\Sigma^{\max}(\xi(t'),\lambda(t')) \geq 0.$$

It can similarly be shown, by considering $t < t'$, that

$$\frac{\mathrm{d}}{\mathrm{d}t}h_\Sigma^{\max}(\xi(t'),\lambda(t')) \leq 0.$$

Thus $t \mapsto h_\Sigma^{\max}(\xi(t),\lambda(t))$ has zero derivative almost everywhere, as desired. Thus there exists $C \in \mathbb{R}$ such that $h_\Sigma^{\max}(\xi(t),\lambda(t)) = C$ for every $t \in [t_0,t_1]$.

Now we show the constancy of $H_\Sigma^{\max}$. We use a lemma that we state in a rather more general form than is necessary, but which illustrates why $H_\Sigma^{\max}$ is lower semicontinuous.

1 Sublemma: *Let X be a topological space, let J be an index set, and let $\{f_j\}_{j\in J}$ be a family of continuous $\mathbb{R}$-valued functions on X. If we define $f^{\max}\colon X \to \overline{\mathbb{R}}$ by*

$$f^{\max}(x) = \sup\{f_j(x) \mid j \in J\},$$

then $f^{\max}$ is lower semicontinuous, i.e., if $x_0 \in X$ and if $\epsilon \in \mathbb{R}_{>0}$, then there exists a neighbourhood $\mathcal{U}$ of x_0 such that $f^{\max}(x) > f^{\max}(x_0) - \epsilon$.

Proof: Let $a \in \mathbb{R}$ and define

$$\mathcal{A}_a = \{x \in X \mid f^{\max}(x) \leq a\}, \quad \mathcal{A}_{j,a} = \{x \in X \mid f_j(x) \leq a\}.$$

We claim that

$$\mathcal{A}_a = \cap_{j\in J}\mathcal{A}_{j,a}.$$

If $x \in \mathcal{A}_a$ then

$$f_j(x) \leq f^{\max}(x) \leq a,$$

implying that $x \in \mathcal{A}_{j,a}$ for each $j \in J$. Conversely, let $x \in \cap_{j\in J}\mathcal{A}_{j,a}$ so that $f_j(x) \leq a$ for every $j \in J$. Let $\epsilon \in \mathbb{R}_{>0}$. Then there exists $j_0 \in J$ such that $f_{j_0}(x) > f^{\max}(x) - \epsilon$. Thus

$$f^{\max}(x) - \epsilon < f_{j_0}(x) \leq a$$

This gives $f^{\max}(x) - \epsilon < a$ for every $\epsilon \in \mathbb{R}_{>0}$, and so $f^{\max}(x) \leq a$ and so $x \in \mathcal{A}_a$.

The above arguments show that $(f^{\max})^{-1}((a,\infty])$ is open for each $a \in \mathbb{R}$. In particular, for $\epsilon \in \mathbb{R}_{>0}$ the set $(f^{\max})^{-1}((f^{\max}(x_0)-\epsilon,\infty])$ is open. Thus there exists a neighbourhood $\mathcal{U}$ about x_0 such that $f^{\max}(x) > f^{\max}(x_0) - \epsilon$. ▼

By the lemma, $H_\Sigma^{\max}$ is lower semicontinuous. Therefore, for every $t' \in [t_0, t_1]$ and $\epsilon \in \mathbb{R}_{>0}$ there exists $\delta \in \mathbb{R}_{>0}$ such that $|t - t'| < \delta$ implies that

$$H_\Sigma^{\max}(\xi(t'), \lambda(t')) < H_\Sigma^{\max}(\xi(t), \lambda(t)) + \epsilon.$$

Since

$$h_\Sigma^{\max}(\xi(t), \lambda(t)) = H_\Sigma^{\max}(\xi(t), \lambda(t))$$

for almost every $t \in [t_0, t_1]$ it follows that there exists t such that this equality holds and such that $|t - t'| < \delta$. Therefore, for this t,

$$H_\Sigma^{\max}(\xi(t'), \lambda(t')) < h_\Sigma^{\max}(\xi(t), \lambda(t)) + \epsilon = C + \epsilon.$$

Since this holds for every $\epsilon > 0$ we have

$$H_\Sigma^{\max}(\xi(t'), \lambda(t')) \leq C = h_\Sigma^{\max}(\xi(t'), \lambda(t'))$$

for every $t' \in [t_0, t_1]$. By (5.4) it follows that $t \mapsto H_\Sigma^{\max}(\xi(t), \lambda(t))$ is constant. ■

The proof of the lemma is rather technical, so it is difficult to glean any insight from it.

5.6. Controlled trajectories on the boundary of the reachable set

In this section we assimilate our (not inconsiderable) efforts put forward thus far in this chapter to prove theorems that will be crucial to the proof of the Maximum Principle. These results lie at the core of why the Maximum Principle and the developments surrounding it are so important, not just in optimal control, but in control theory in general.

5.6.1. The fixed interval case. The main result here is the following. Much of the work in the proof has already been done as a result of our efforts to understand the tangent cones and their relationship with the Hamiltonian.

5.16 Theorem: (A characterisation of trajectories steered to the boundary of the fixed time reachable set) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $\mu_* \in \mathscr{U}(x_0, t_0, [t_0, t_1])$. If $\xi(\mu_*, x_0, t_0, t_1) \in \mathrm{bd}(\mathcal{R}(x_0, t_0, t_1))$ then there exists an adjoint response $\lambda_* \colon [t_0, t_1] \to \mathbb{R}^n$ for Σ along $(\xi(\mu_*, x_0, t_0, \cdot), \mu_*)$ such that*

$$H_\Sigma(\xi(\mu_*, x_0, t_0, t), \lambda_*(t), \mu_*(t)) = H_\Sigma^{\max}(\xi(\mu_*, x_0, t_0, t), \lambda_*(t))$$

for almost every $t \in [t_0, t_1]$. Moreover, if additionally $\mu_* \in \mathscr{U}_{\mathrm{bdd}}([t_0, t_1]$, then the function

$$t \mapsto H_\Sigma^{\max}(\xi(\mu_*, x_0, t_0, t), \lambda_*(t))$$

is constant.

Proof: Let us abbreviate $\xi_* = \xi(\mu_*, x_0, t_0, \cdot)$. Since $\xi_*(t_1) \in \mathrm{bd}(\mathcal{R}(x_0, t_0, t_1))$ there exists a sequence $\{x_j\}_{j \in \mathbb{Z}_{>0}}$ in $\mathcal{X} \setminus \mathrm{cl}(\mathcal{R}(x_0, t_0, t_1))$ which converges to $\xi_*(t_1)$. Let $v_j = \frac{x_j - \xi_*(t_1)}{\|x_j - \xi_*(t_1)\|}$ so that $\{v_j\}_{j \in \mathbb{Z}_{>0}}$ is a sequence in $\mathbb{S}^{n-1}$. By the Bolzano–Weierstrass Theorem there exists a subsequence $\{v_{j_k}\}_{k \in \mathbb{Z}_{>0}}$ which converges to some unit vector v_0. We claim that $v_0 \notin \mathrm{int}(K(\mu_*, x_0, t_0, t_1))$. Indeed, were this not the case, then for some sufficiently large N it

would hold that $v_{j_k} \in \text{int}(K(\mu_*, x_0, t_0, t_1))$ for $k \geq N$. By Lemma 5.10 this implies that $x_{j_k} \in \text{int}(\mathcal{R}(x_0, t_0, t_1))$ for all k sufficiently large. This violates the assumption that $x_{j_N} \notin \text{cl}(\mathcal{R}(x_0, t_0, t_1))$. Thus either $v_0 \in \text{bd}(K(\mu_*, x_0, t_0, t_1))$ or $v_0 \notin \text{bd}(K(\mu_*, x_0, t_0, t_1))$ since $K(\mu_*, x_0, t_0, t)$ is closed. By Corollary B.20 it follows that there exists a hyperplane $P(t_1)$ such that v_0 is contained in one of the closed half-spaces defined by $P(t_1)$ and $K(\mu, x_0, t_0, t_1)$ is contained in the other closed half-space. Let $\lambda_*(t_1)$ be a vector orthogonal to $P(t_1)$ contained in a half-space not containing $K(\mu, x_0, t_0, t)$. Let λ_* be the unique adjoint response for Σ along (ξ_*, μ_*) having the property that it takes the value $\lambda_*(t_1)$ at time t_1. Note that

$$\langle \lambda_*(t_1), v \rangle \leq 0, \qquad v \in K(\mu_*, x_0, t_0, t_1).$$

The theorem now follows from Lemmata 5.13 and 5.15. ■

In Figure 5.3 we illustrate the idea behind the proof of the preceding theorem. The

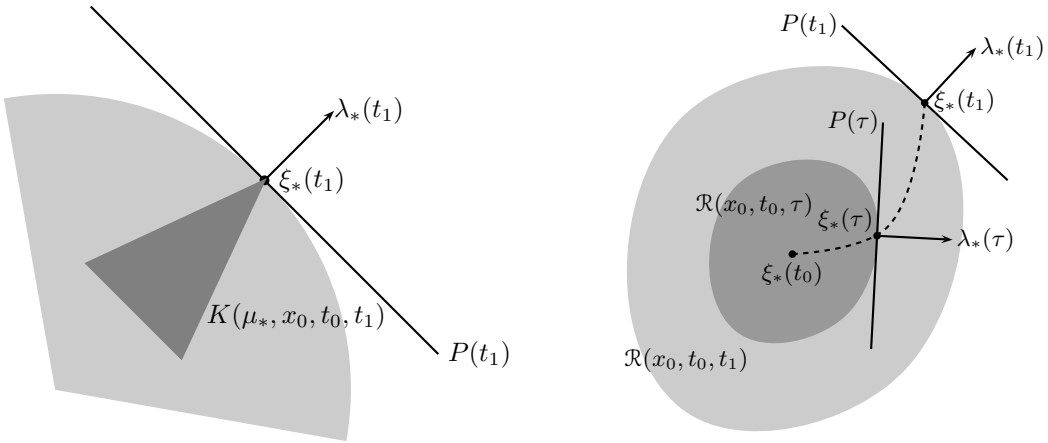

Figure 5.3. A depiction of the proof of Theorem 5.16. The cartoon on the left shows the tangent cone at the terminal time and the cartoon on the right shows the transport of hyperplanes along the trajectory.

idea behind the picture on the left is where the proof starts, and so this is perhaps the most important thing to understand. The intuition is that, since the final point of the trajectory is on the boundary, any variation of the trajectory will move "inwards" and so the cone $K(\mu_*, x_0, t_0, t_1)$ will point "into" the reachable set. There is, therefore, a hyperplane separating the cone from a vector $\lambda_*(t_1)$ which points "out" of the reachable set. This is what is depicted on the left in Figure 5.3. With the final value of the adjoint response given by $\lambda_*(t_1)$, this can be translated backwards along the trajectory. Moreover, at each point on the trajectory, the picture on the left in Figure 5.3 will be duplicated by virtue of Proposition 4.5 and Lemma 5.5. This is the situation depicted on the right in Figure 5.3. With the adjoint response defined, one can now define the Hamiltonian, which is perhaps the most mysterious aspect in the statement of Theorem 5.16. But the properties of the Hamiltonian are shown by Lemma 5.13 to follow from the picture on the left in Figure 5.3.

Corresponding to our comments following the statement of Lemma 5.10, one should be careful about what Theorem 5.16 is saying when $\text{int}(K(\mu_*, x_0, t_0, t_1)) = \emptyset$. While the

theorem applies in this case, it just may not be giving the sort of information that one expects based on the "standard" picture in Figure 5.3. The interested reader can think about this a little in Exercise E5.2. Here we shall only point out that this sort of discussion is where the deep connections between controllability theory and optimal control arise.

It is interesting to note that if a controlled trajectory hits the boundary of the reachable set at some time, it must have been at the boundary for all preceding times. The following result is equivalent to this, although it is stated in different terms.

5.17 Proposition: (Trajectories in the interior of the fixed time reachable set remain in the interior) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$. Suppose that for some $\tau \in [t_0, t_1)$ we have $\xi(\mu, x_0, t_0, \tau) \in \operatorname{int}(\mathcal{R}(x_0, t_0, \tau))$. Then $\xi(\mu, x_0, t_0, t) \in \operatorname{int}(\mathcal{R}(x_0, t_0, t))$ for all $t \in (\tau, t_1]$.*

Proof: Let $\mathcal{U}$ be a neighbourhood of $\xi(\mu, x_0, t_0, \tau)$ contained in $\mathcal{R}(x_0, t_0, \tau)$. For each $x \in \mathcal{U}$ there then exists a control $\mu_x \in \mathscr{U}(x_0, t_0, [t_0, \tau])$ such that $\xi(\mu_x, x_0, t_0, \tau) = x$. For $t' \in [\tau, t_1]$ extend μ_x to a control in $\mathscr{U}(x_0, t_0, [t_0, t'])$ as follows:

$$\mu_x(t) = \begin{cases} \mu_x(t), & t \in [t_0, \tau], \\ \mu(t), & t \in (\tau, t'], \end{cases}$$

where we make a slight abuse of notation. We then define a map $F \colon \mathcal{U} \to \mathcal{R}(x_0, t_0, t')$ by $x \mapsto \xi(\mu_x, x_0, t_0, t')$. Thus F sends x to the solution at t' of the initial value problem

$$\dot{\xi}(t) = f(\xi(t), \mu_x(t)), \quad \xi(\tau) = x.$$

This map is then a diffeomorphism onto its image, and so $F(\mathcal{U})$ is open. Moreover, it clearly contains $\xi(\mu, x_0, t_0, t')$, and so $\xi(\mu, x_0, t_0, t') \in \operatorname{int}(\mathcal{R}(x_0, t_0, t'))$. ■

5.6.2. The free interval case. The discussion above carries over more or less verbatim to characterisations of the boundary of the reachable set (as opposed to the fixed time reachable set). Thus we shall merely state the results and omit detailed proofs and discussion.

5.18 Theorem: (A characterisation of trajectories steered to the boundary of the reachable set) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $\mu_* \in \mathscr{U}(x_0, t_0, [t_0, t_1])$. If $\xi(\mu_*, x_0, t_0, t_1) \in \operatorname{bd}(\mathcal{R}(x_0, t_0))$ then there exists an adjoint response $\lambda_* \colon [t_0, t_1] \to \mathbb{R}^n$ for Σ along $(\xi(\mu_*, x_0, t_0, \cdot), \mu_*)$ such that*

$$H_\Sigma(\xi(\mu_*, x_0, t_0, t), \lambda_*(t), \mu_*(t)) = H_\Sigma^{\max}(\xi(\mu_*, x_0, t_0, t), \lambda_*(t)) = 0$$

for almost every $t \in [t_0, t_1]$. Moreover, if additionally $\mu_* \in \mathscr{U}_{\text{bdd}}([t_0, t_1]$, then the function

$$t \mapsto H_\Sigma^{\max}(\xi(\mu_*, x_0, t_0, t), \lambda_*(t))$$

is everywhere zero.

Proof: The existence of an adjoint response such that the Hamiltonian is maximised almost everywhere follows in the same manner as the proof of Theorem 5.16 except that one uses $K^{\pm}(\mu_*, x_0, t_0, t)$ in place of $K(\mu_*, x_0, t_0, t)$ and $\mathcal{R}(x_0, t_0)$ in place of $\mathcal{R}(x_0, t_0, t_1)$. That the Hamiltonian and the maximum Hamiltonian are almost everywhere zero follows from Lemmata 5.13 and 5.14, noting that $K(\mu_*, x_0, t_0, t) \subset K^{\pm}(\mu_*, x_0, t_0, t)$ for every $t \in [t_0, t_1]$ and that $\operatorname{span}_{\mathbb{R}}(f(\xi_*(t), \mu_*(t))) \subset K^{\pm}(\mu_*, x_0, t_0, t)$ for almost every $t \in [t_0, t_1]$. The constancy of the maximum Hamiltonian follows from Lemma 5.15. ■

We also have the statement that trajectories in the interior of the reachable set remain in the interior.

5.19 Proposition: (Trajectories in the interior of the reachable set remain in the interior) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $x_0 \in \mathcal{X}$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$. Suppose that for some $\tau \in [t_0, t_1)]$ we have $\xi(\mu, x_0, t_0, \tau) \in \operatorname{int}(\mathcal{R}(x_0, t_0))$. Then $\xi(\mu, x_0, t_0, t) \in \operatorname{int}(\mathcal{R}(x_0, t_0))$ for all $t \in (\tau, t_1]$.*

Exercises

E5.1 Take $\mathsf{M} = \mathbb{R}^2$, $m = 1$, $U = [-1, 1]$, and define $f((x^1, x^2), u) = (u, (x^2)^2)$. Show that the reachable sets $\mathcal{R}((0,0), 0, T)$ and $\mathcal{R}((0,0), 0)$ are as depicted in Figure E5.1, where the left and right boundaries for $\mathcal{R}((0,0), 0)$ are given by the graph of the

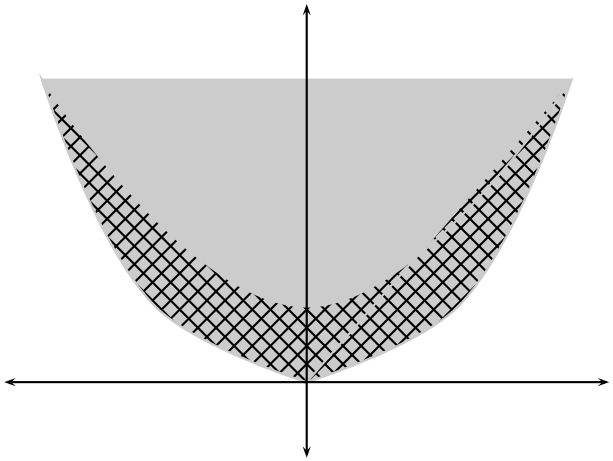

Figure E5.1. Reachable sets: the shaded area represents $\mathcal{R}((0,0), 0)$ and the hatched area represents $\mathcal{R}((0,0), 0, T)$ for some $T > 0$.

function $x^2(x^1) = \frac{1}{3}|x^1|^3$, and that the upper boundary for $\mathcal{R}((0,0), 0, T)$ is given by the graph of the function

$$x^2(x^1) = -\frac{|x^1|^3}{4} + \frac{T|x^1|^2}{4} + \frac{T^2|x^1|}{4} + \frac{T^3}{12}.$$

In particular, $\mathcal{R}((0,0), 0, T)$ is not convex.

In the next exercise you will explore what can happen when the interior of the fixed interval tangent cone is empty.

E5.2 Let $\mathcal{X} = \mathbb{R}^2$, let $m = 1$, define

$$f((x^1, x^2), u) = (-x^2 u, x^1 u),$$

and take $U = [-1, 1]$. For the system $\Sigma = (\mathcal{X}, f, U)$, answer the following questions.

(a) Show that for any admissible control u the resulting trajectory with initial condition (x_0^1, x_0^2) takes values in the circle of radius $\|(x_0^1, x_0^2)\|$.

(b) Show that $K(u, (x_0^1, x_0^2), 0, t) \subset \operatorname{span}_{\mathbb{R}}(-\xi^2(t), \xi^1(t))$ for any admissible control u, where $(\xi^1(t), \xi^2(t)) = \xi(u, (x_0^1, x_0^2), 0, t)$. Note that $\operatorname{int}(K(u, (x_0^1, x_0^2), 0, t)) = \emptyset$.

(c) Show that

$$(\xi(u, (x_0^1, x_0^2), 0, t) + K(u, (x_0^1, x_0^2), 0, t)) \cap \mathcal{R}((x_0^1, x_0^2), 0, t) = \emptyset.$$

(d) Let $t \mapsto (\lambda^1(t), \lambda^2(t))$ be an adjoint response satisfying

$$\langle (\lambda^1(t_0), \lambda^2(t_0)), (-\xi^2(t_0), \xi^1(t_0)) \rangle = 0.$$

Show that

$$\langle (\lambda^1(t), \lambda^2(t)), (-\xi^2(t), \xi^1(t)) \rangle = 0$$

for all t.

(e) Show that for *any* controlled trajectory, the conclusions of Theorem 5.16 or Theorem 5.18 hold with an adjoint response in part (d).

(f) What is the point of this exercise?

Chapter 6

A proof of the Maximum Principle

In this chapter we complete the proof of the Maximum Principle. As we shall see, much of the heavy lifting has been done in Chapter 5. We will break the proof up into various components. We will prove Theorems 3.4 and 3.5 simultaneously since their proofs differ only in a few places. We shall be careful to point out the places where the two cases need to be considered separately.

6.1. The extended system

It is very helpful in optimal control in general to include the cost as a variable in the problem, thereby extending the state space.

6.1 Definition: (Extended system) Let $\Sigma = (\mathcal{X}, f, U)$ be a control system and let L be a Lagrangian for Σ. The ***extended system*** is the system $\hat{\Sigma} = (\hat{\mathcal{X}}, \hat{f}, U)$ defined by asking that

(i) $\hat{\mathcal{X}} = \mathbb{R} \times \mathcal{X}$ and

(ii) $\hat{f}((x^0, x), u) = (L(x, u), f(x, u))$. •

Note that the equations governing the extended system are

$$\begin{aligned} \dot{\xi}^0(t) &= L(\xi(t), \mu(t)), \\ \dot{\xi}(t) &= f(\xi(t), \mu(t)). \end{aligned}$$

This immediately gives

$$\xi^0(t) = \int_{t_0}^{t} L(\xi(\tau), \mu(\tau))\, \mathrm{d}\tau,$$

so that $\xi^0(t)$ is the cost accrued by the controlled trajectory up to time t. Thus the extended system essentially has the cost added as an extra variable.

We will adopt the convention of placing a "hat" on everything associated to the extended system. Thus, for example, the state for $\hat{\Sigma}$ will be denoted by $\hat{x} = (x^0, x)$, a trajectory for $\hat{\Sigma}$ will be denoted by $\hat{\xi}$, the reachable set for $\hat{\Sigma}$ will be denoted by $\hat{\mathcal{R}}(x_0, t_0)$, and $\hat{K}(\mu, \hat{x}_0, t_0, t)$ will denote a fixed interval tangent cone for the extended system. We will not bother to explicitly define all of the new symbols arising from this convention as it should be patently obvious what they mean.

6.2. Optimal trajectories lie on the boundary of the reachable set of the extended system

The first part of the proof of the Maximum Principle makes the connection between optimal control and the results concerning the reachable set in Chapter 5. This connection lies at the heart of the reason why the Maximum Principle has to do with so much more than just optimal control.

The main result is the following.

6.2 Lemma: (Optimal trajectories lie on the boundary of the reachable set of the extended system) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let L be a Lagrangian for Σ, and let $S_0, S_1 \subset \mathcal{X}$ be subsets. Suppose that $(\xi_*, \mu_*) \in \mathscr{P}(\Sigma, L, S_0, S_1, [t_0, t_1])$ or that $(\xi_*, \mu_*) \in \mathscr{P}(\Sigma, L, S_0, S_1)$ is defined on $[t_0, t_1]$, respectively. Then $\hat{\xi}_*(t_1) \in \mathrm{bd}(\hat{\mathcal{R}}(\hat{\xi}_*(t_0), t_0, t_1))$ or $\hat{\xi}_*(t_1) \in \mathrm{bd}(\hat{\mathcal{R}}(\hat{\xi}_*(t_0), t_0))$, respectively.*

Proof: Let us first consider the case where $(\xi_*, \mu_*) \in \mathscr{P}(\Sigma, L, S_0, S_1, [t_0, t_1])$. We claim that $\hat{\xi}_*(t_1) = (\xi_*^0(t_1), \xi_*(t_1))$ has the property that

$$\xi_*^0(t_1) = \inf\{x^0 \in \mathbb{R} \mid (x^0, \xi_*(t_1)) \in \hat{\mathcal{R}}(\hat{\xi}(t_0), t_0, t_1)\}.$$

That is to say, the final cost $\xi_*^0(t_1)$ is the lowest possible among all elements of $(\xi, \mu) \in \mathrm{Carc}(\Sigma, L, S_0, S_1, [t_0, t_1])$ that steer $\xi_*(t_0)$ to $\xi_*(t_1)$. Said this way, the claim is obvious since (ξ_*, μ_*) is assumed to be optimal. To see that this implies that $\hat{\xi}_*(t_1) \in \mathrm{bd}(\hat{\mathcal{R}}(\hat{\xi}_*(t_0), t_0, t_1))$, let $\mathcal{U}$ be a neighbourhood of $\hat{\xi}_*(t_1)$ in $\hat{\mathcal{X}}$. Since any such neighbourhood will contain points of the form $(x^0, \xi_*(t_1))$ with $x^0 < \xi_*^0(t_1)$, and since such points are not in $\hat{\mathcal{R}}(\hat{\xi}(t_0), t_0, t_1)$, this means that $\hat{\xi}_*(t_1) \in \mathrm{bd}(\hat{\mathcal{R}}(\hat{\xi}_*(t_0), t_0, t_1))$, as desired.

The argument in the case where $(\xi_*, \mu_*) \in \mathscr{P}(\Sigma, L, S_0, S_1)$ follows similarly, the only difference being that the final times are free both in the optimal control problem and in the reachable set. ■

In Figure 6.1 we depict the idea behind the proof of the preceding lemma.

6.3. The properties of the adjoint response and the Hamiltonian

Using Lemma 6.2 it is easy to deduce the Hamiltonian components of the Maximum Principle. We begin be asserting the existence of an adjoint response having the properties asserted in the Maximum Principle.

6.3 Lemma: (The adjoint response) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let L be a Lagrangian for Σ, and let $S_0, S_1 \subset \mathcal{X}$ be subsets. Suppose that $(\xi_*, \mu_*) \in \mathscr{P}(\Sigma, L, S_0, S_1)$ is defined on $[t_0, t_1]$ or that $(\xi_*, \mu_*) \in \mathscr{P}(\Sigma, L, S_0, S_1, [t_0, t_1])$. Then there exists an absolutely continuous map $\lambda_* \colon [t_0, t_1] \to \mathbb{R}^n$ and $\lambda_*^0 \in \{0, -1\}$ with the following properties:*

(i) either $\lambda_*^0 = -1$ or $\lambda_*(t_0) \neq 0$;

(ii) λ_* is an adjoint response for $(\Sigma, \lambda_*^0 L)$ along (ξ_*, μ_*);

(iii) $H_{\Sigma, \lambda_*^0 L}(\xi_*(t), \lambda_*(t), \mu_*(t)) = H^{\max}_{\Sigma, \lambda_*^0 L}(\xi_*(t), \lambda_*(t))$ for almost every $t \in [t_0, t_1]$.

Proof: Let us first consider the fixed interval problem. We shall refine slightly the first steps in the proofs of Theorem 5.16 since in this case we have the additional structure of

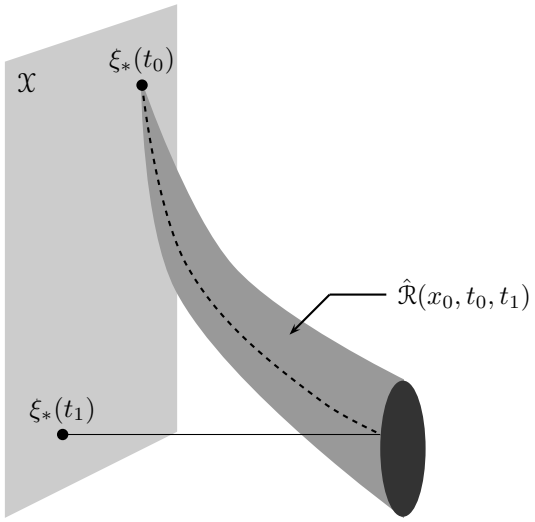

Figure 6.1. The idea of the proof of Lemma 6.2

our system being the extended system associated with an optimal control problem. The key element of the refinement is the observation that the vector $(-1, 0) \in \mathbb{R} \oplus \mathbb{R}^n$ cannot lie in the interior of $\hat{K}(\mu_*, \hat{x}_0, t_0, t_1)$. Indeed, if this were not the case then, by Lemma 5.10, there must be points in $\hat{\mathcal{R}}(\hat{\xi}_*(t_0), t_0, t_1)$ whose final cost would be lower than $\xi^0_*(t_1)$, and this would violate the optimality of (ξ_*, μ_*). Therefore, there exists a hyperplane $\hat{P}(t_1)$ such that $(-1, 0)$ is contained in one of the closed half-spaces defined by $\hat{P}(t_1)$ and $\hat{K}(\mu_*, \hat{x}_0, t_0, t_1)$ is contained in the other closed half-space. We take $\hat{\lambda}_*(t_1)$ to be a vector orthogonal to $\hat{P}(t_1)$ and contained in the half-space not containing $\hat{K}(\mu_*, \hat{x}_0, t_0, t_1)$, and note that

$$\begin{aligned} &\langle \hat{\lambda}_*(t_1), (-1, 0) \rangle \geq 0, \\ &\langle \hat{\lambda}_*(t_1), \hat{v} \rangle \leq 0, \qquad \hat{v} \in \hat{K}(\mu_*, \hat{x}_0, t_0, t_1). \end{aligned}$$

Note that this implies that $\lambda^0_*(t_1) \leq 0$. We then define $\hat{\lambda}_*$ to be the adjoint response equal to $\hat{\lambda}_*(t_1)$ at time t_1. From the equations for the adjoint response we immediately have $\dot{\lambda}^0_*(t) = 0$ (since $\hat{f}$ is independent of x^0) and so λ^0_* is constant and nonpositive. If $\lambda^0_* \neq 0$ then we can redefine $\hat{\lambda}_*$ to be $-(\lambda^0_*)^{-1}\hat{\lambda}_*$, and this ensures that $\hat{\lambda}_*(t) = (\lambda^0_*, \lambda_*(t))$ with $\lambda^0_* \in \{0, -1\}$.

Since

$$\hat{H}_{\hat{\Sigma}}((x^0, x), (p^0, p), u) = \langle p, f(x, u) \rangle + p^0 L(x, u) = H_{\Sigma, p^0 L}(x, p, u),$$

it follows from Theorem 5.16 that

$$H_{\Sigma, \lambda^0_* L}(\xi_*(t), \lambda_*(t), \mu_*(t)) = H^{\max}_{\Sigma, \lambda^0_* L}(\xi_*(t), \lambda_*(t))$$

for almost every $t \in [t_0, t_1]$.

The condition that $\lambda^0_* = -1$ or $\lambda_*(t_0) \neq 0$ follows since $\hat{\lambda}_*(t) \neq 0$ for every $t \in [t_0, t_1]$ by virtue of linearity of the adjoint equation. This gives the lemma in the fixed interval case.

The argument in the free interval case is essentially the same, but with $\hat{K}^{\pm}(\mu_*, \hat{x}_0, t_0, t_1)$ replacing $\hat{K}(\mu_*, \hat{x}_0, t_0, t_1)$ and with $\hat{\mathcal{R}}(\hat{\xi}_*(t_0), t_0)$ replacing $\hat{\mathcal{R}}(\hat{\xi}_*(t_0), t_0, t_1)$. ■

The lemma, along with Theorems 5.16 and 5.18, give the following corollary.

6.4 Corollary: (Constancy of the Hamiltonian) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let L be a Lagrangian for Σ, and let $S_0, S_1 \subset \mathcal{X}$ be subsets. Suppose that $(\xi_*, \mu_*) \in \mathscr{P}(\Sigma, L, S_0, S_1, [t_0, t_1])$ or that $(\xi_*, \mu_*) \in \mathscr{P}(\Sigma, L, S_0, S_1)$ is defined on $[t_0, t_1]$, respectively. Also suppose that $\mu_* \in \mathscr{U}_{\text{bdd}}([t_0, t_1])$. If $\lambda_*^0 \in \{0, -1\}$ and $\lambda_* \colon [t_0, t_1] \to \mathbb{R}^n$ are as in Lemma 6.3, then*

$$t \mapsto H^{\max}_{\Sigma, \lambda_*^0 L}(\xi_*(t), \lambda_*(t))$$

is constant or everywhere zero, respectively.

6.4. The transversality conditions

The verification of the transversality conditions takes a significant amount of additional work. Our discussion will begin by taking a rather general form, ostensibly having nothing to do with optimal control. For readers familiar with the notion of a manifold with boundary, it is possible to skip the first rather clumsy steps which attempt to mimic this idea without going through the complete programme.

6.5 Definition: (Edged set) An ***edged set*** in $\mathbb{R}^n$ is a subset $\mathcal{E}$ such that $\mathcal{E} = \phi(\mathcal{U})$ where

(i) $\mathcal{U} \subset \mathbb{R}^k$ has the form

$$\mathcal{U} = \mathcal{U}' \cap \{(y^1, \dots, y^k) \mid y^k \geq 0\}$$

where $\mathcal{U}'$ is a neighbourhood of 0, and where

(ii) $\phi \colon \mathcal{U}' \to \mathbb{R}^n$ is a homeomorphism onto its image with $\boldsymbol{D}\phi(y)$ injective for each $y \in \mathcal{U}'$.

The ***boundary*** of $\mathcal{E}$ is the set

$$\text{bd}(\mathcal{E}) = \{\phi(y^1, \dots, y^{k-1}, 0) \mid (y^1, \dots, y^{k-1}, 0) \in \mathcal{U}\}.$$

The ***dimension*** of such an edged set is $\dim(\mathcal{E}) = k$. •

Note that the boundary of an edged set $\mathcal{E}$ will not generally agree with the boundary of $\mathcal{E}$ as a subset of $\mathbb{R}^n$. This might cause confusion but for the fact that in our discussion we will always mean the boundary of an edged set $\mathcal{E}$ to be as in Definition 6.5.

The picture one might have in mind for an edged set is depicted in Figure 6.2. At points $x \in \mathcal{E}$ that are not in the boundary of $\mathcal{E}$, the usual notion of tangent space applies, i.e., the tangent space is the image of $\boldsymbol{D}\phi(y)$ where $\phi(y) = x$. At points on the boundary we use the following notion.

6.6 Definition: (Tangent half space) Let $\mathcal{E}$ be an edged set with $\mathcal{E} = \phi(\mathcal{U})$ and with $\mathcal{U}$ and ϕ as in Definition 6.5. For $x \in \text{bd}(\mathcal{E})$ let $y \in \mathcal{U}$ have the property that $\phi(y) = x$. The ***tangent half-space*** to $\mathcal{E}$ at x is

$$\mathsf{T}^+_x\mathcal{E} = \boldsymbol{D}\phi(y)(\{(v^1, \dots, v^k) \mid v^k \geq 0\}).$$ •

The idea behind the tangent half-space is depicted in Figure 6.3.

Let us now suppose that S_0 is a smooth constraint set being defined by $S_0 = \Phi_0^{-1}(0)$. We denote $\mathsf{T}_x S_0 = \ker(\boldsymbol{D}\Phi_0(x))$ the tangent space to S_0 at a point $x \in S_0$. Now let us

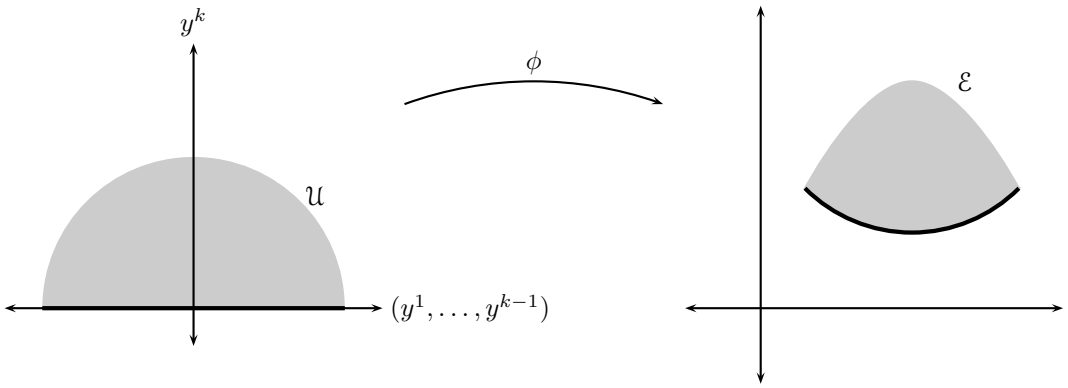

Figure 6.2. A depiction of an edged set

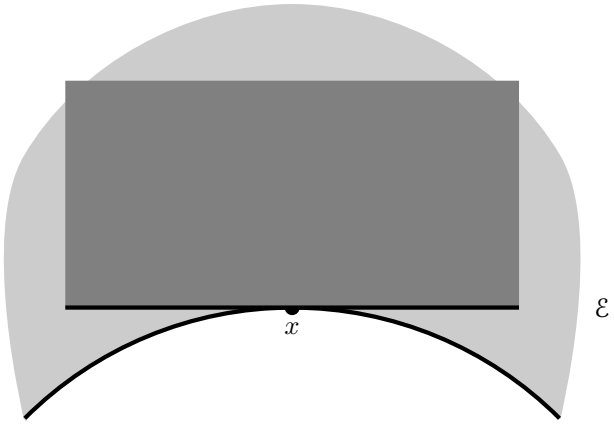

Figure 6.3. A depiction of the tangent half-space

suppose that we have an initial point $x_0 \in S_0$, a control $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$, and a point $\tau \in (t_0, t_1) \cap \text{Leb}(\mu, x_0, t_0, t_1)$. For brevity denote $\xi = \xi(\mu, (0, x_0), t_0, \cdot)$. Let us denote

$$\mathscr{K}(\mu, x_0, t_0, t) = \text{cl}(\text{conv cone}(\Phi(\mu, x_0, t_0, t_0, t)(\mathsf{T}_{\xi(t_0)} S_0) \cup K(\mu, x_0, t_0, t))).$$

With these ideas at hand, we state a result which is somewhat analogous to Lemma 5.10. The lemma is stated in a way that it has nothing *a priori* to do with optimal control, just as is the case with Lemma 5.10.

6.7 Lemma: (Fixed interval tangent cones and tangent half-spaces that are not separable) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $S_0 = \Phi^{-1}(0)$ be a constraint set, let $x_0 \in S_0$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$, and let $\tau \in (t_0, t_1) \cap \text{Leb}(\mu, x_0, t_0, t_1)$. Let $\mathcal{E}$ be an edged set with $\xi(\mu, x_0, t_0, \tau) \in \text{bd}(\mathcal{E})$ and suppose that the tangent half-space of $\mathcal{E}$ at $\xi(\mu, x_0, t_0, \tau)$ and the cone $\mathscr{K}(\mu, x_0, t_0, \tau)$ are not separable. Then there exists $x_0' \in S_0$ such that the set*

$$(\mathcal{E} \setminus \text{bd}(\mathcal{E})) \cap \mathcal{R}(x_0', t_0, t_1)$$

is nonempty.

Proof: We will be slightly sketchy about the proof here since it has many similarities to the proof of Lemma 5.10 which we presented in some detail.

Let $\xi = \xi(\mu, x_0, t_0, \cdot)$. Since $\mathsf{T}^+_{\xi(\tau)}\mathcal{E}$ and $\mathscr{K}(\mu, x_0, t_0, \tau)$ are not separable there can be no proper subspace of $\mathbb{R}^n$ containing both by Theorem B.22. That is,

$$\mathbb{R}^n = \operatorname{span}_{\mathbb{R}}(\mathsf{T}^+_{\xi(\tau)}\mathcal{E}) + \operatorname{span}_{\mathbb{R}}(\mathscr{K}(\mu, x_0, t_0, \tau)).$$

Moreover, since the cones $\mathsf{T}^+_{\xi(\tau)}\mathcal{E}$ and $\mathscr{K}(\mu, x_0, t_0, \tau)$ are not separable, their relative interiors have a nonempty intersection by Theorem B.22. Therefore,

$$\operatorname{span}_{\mathbb{R}}(\mathsf{T}^+_{\xi(\tau)}\mathcal{E}) \cap \operatorname{span}_{\mathbb{R}}(\mathscr{K}(\mu, x_0, t_0, \tau)) \neq \{0\}.$$

Thus we can choose a subcone H_τ of $\mathsf{T}^+_{\xi(\tau)}\mathcal{E}$ such that

$$\mathbb{R}^n = H_\tau + \operatorname{span}_{\mathbb{R}}(\mathscr{K}(\mu, x_0, t_0, \tau)), \quad \dim(H_\tau \cap \mathscr{K}(\mu, x_0, t_0, \tau)) = 1.$$

Let us now choose coordinates for $\mathbb{R}^n$ in order to simplify the problem. First, by an orthogonal change of coordinates we can suppose that

$$H_\tau \cap \mathscr{K}(\mu, x_0, t_0, \tau) = \operatorname{span}_{\mathbb{R}}((1, 0, \ldots, 0)).$$

Let P_τ denote the orthogonal complement to $(1, 0, \ldots, 0)$ in $\operatorname{aff}(\mathscr{K}(\mu, x_0, t_0, \tau)$. By a linear change of coordinates we can suppose that P_τ is orthogonal to H_τ. We suppose that our coordinates are chosen such that the first k basis vectors span the subspace generated by H_τ and the last $n-k$ basis vectors span P_τ. We depict the situation in the new coordinates in Figure 6.4.

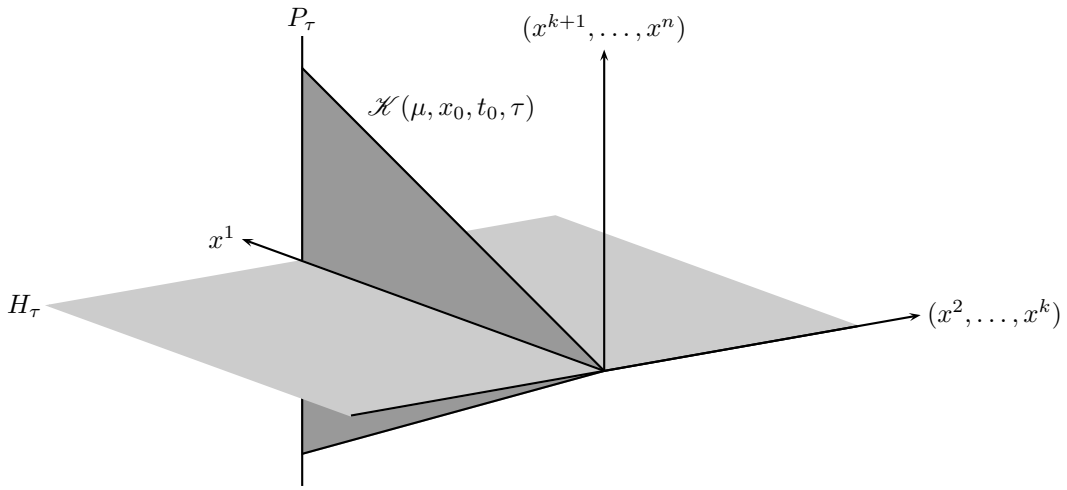

Figure 6.4. A depiction of the constructions used in the proof of Lemma 6.7

Now let $v_0 \in \operatorname{rel\,int}(H_\tau) \cap \operatorname{rel\,int}(\mathscr{K}(\mu, x_0, t_0, \tau))$. We define two simplex cone neighbourhoods of v_0, one in H_τ and another in $\mathscr{K}(\mu, x_0, t_0, \tau)$. The existence of such neighbourhoods

is ensured by Proposition B.17. The simplex cone neighbourhood K_1 in H_τ we define as follows. For $x \in \mathcal{E}$ let us denote by $\tilde{f}_1^{-1}(x)$ the orthogonal component of $x - \xi(\tau)$ in the subspace generated $\mathsf{T}^+_{\xi(\tau)}\mathcal{E}$. In a sufficiently small simplex cone neighbourhood $\tilde{K}_1$ of v_0 in $\mathsf{T}^+_{\xi(\tau)}\mathcal{E}$ the notation makes sense and $\tilde{f}_1$ defines a homeomorphism from a tip C_1' of the cone K_1' to a neighbourhood of $\xi(\tau)$ in $\mathcal{E}$. We then take K_1' to be a simplex cone neighbourhood of v_0 in H_τ that is contained in $\tilde{K}_1 \cap H_\tau$. We take K_2' to be any simplex cone neighbourhood of v_0 in $\mathscr{K}(\mu, x_0, t_0, \tau)$. We may choose the vectors generating the cone K_2' to lie in the plane P_τ passing through v_0.

Let us now consider the effects of perturbing the initial condition from x_0 in S_0. To do this, we parameterise S_0 in a neighbourhood $\mathcal{U}_0$ of x_0 in S_0 as follows. We let $\mathsf{T}_{x_0}S_0$ be the tangent space to S_0 at x_0. We then write $x \in S_0$ as $x = x_0 + v(x) + u(x)$ for some uniquely defined $v(x) \in \mathsf{T}_{x_0}S_0$ and $u(x)$ orthogonal to $\mathsf{T}_{x_0}S_0$. If $x \in \mathcal{U}_0$ and if $\mathcal{U}_0$ is sufficiently small, then the map $x \mapsto v(x)$ is a diffeomorphism, essentially by the Implicit Function Theorem (we leave the details of this for the reader to sort out). If $x \in \mathcal{U}_0$ and if $s \in \mathbb{R}$ is sufficiently small, then we have

$$\xi(\mu, x, t_0, \tau) = \xi(\tau) + s\Phi(\mu, x_0, t_0, t_0, \tau) \cdot v(x) + o(s).$$

In arriving at this formula, one uses the fact that τ is a Lebesgue point, along with computations like those used in the proofs of Propositions 4.9, 4.12, and 4.16. If, additionally, Θ is fixed interval needle variation data, then we have, again following our familiar computations,

$$\xi(\mu_\Theta(s, \cdot), x, t_0, \tau) = \xi(\tau) + sv_\Theta(\tau) + s\Phi(\mu, x_0, t_0, t_0, \tau) \cdot v(x) + o(s), \tag{6.1}$$

where $v_\Theta(\tau)$ is the fixed interval multi-needle variation for Θ at time τ. The equation (6.1) then defines (after some details which we leave to the reader; but cf. the proof of Lemma 5.10) a map $\tilde{f}_2$ from a tip C_2' of the cone K_2' to $\mathbb{R}^n$.

Now, just as in Lemma 5.10, we use coordinates (l, r_1) and (l, r_2) to parameterise the cones K_1' and K_2'. Under this parameterisations, the tips C_1' and C_2' of these cones become cubes in $\mathbb{R}^k$ and $\mathbb{R}^{n-k+1}$, respectively. We also use coordinates (L_1, R_1) and (L_2, R_2) for the codomain of the maps $\tilde{f}_1$ and $\tilde{f}_2$ in the coordinates (l, r_1) and (l, r_2), respectively. This notation mirrors that in Lemma 5.10. We let f_1 and f_2 denote the representations of the maps $\tilde{f}_1$ and $\tilde{f}_2$ in these coordinates. We then have f_1 and f_2 given by

$$f_1(l, r_1) = (L_1, R_1), \quad f_2(l, r_2) = (L_2, R_2)$$

where

$$L_1(l, r_1) = l, \quad R_1(l, r_1) = r_1 + o(1), \quad L_2(l, r_2) = l + o(l), \quad R_2(l, r_2) + o(1).$$

We can extend these maps to be defined at $l = 0$ by taking

$$L_1(0, r_1) = 0, \quad R_1(0, r_1) = r_1, \quad L_2(0, r_1) = 0, \quad R_2(0, r_2) = r_2.$$

Now, for $a \in \mathbb{R}_{>0}$ define

$$C(a) = \{x + \tfrac{a}{2}(1, 0, \dots, 0) \mid \max\{|x^1|, \dots, |x^n|\} \leq a\}$$

to be the cube with sides of length a shifted by $\frac{a}{2}$ in the direction of v_0. For $b \in \mathbb{R}_{>0}$ define

$$P_1(a) = \{(x^1, \dots, x^n) \in C(a) \mid x^{k+1} = \cdots = x^n = 0\},$$
$$P_2(a,b) = \{(x^1, \dots, x^n) \in C(a) \mid x^1 = b,\ x^2 = \cdots = x^k = 0\}.$$

For a sufficiently small, $C(a)$ is in the domain of f_1 and f_2. Moreover, a can be chosen sufficiently small (cf. the proof of Lemma 5.10) that

$$\|f_1(x_1) - x_1\| < \tfrac{a}{4}, \quad \|f_2(x_2) - x_2\| < \tfrac{a}{4}$$

for $x_1 \in P_1(a)$ and $x_2 \in P_2(a, \frac{1}{2}))$. By Lemma C.4, $f_1(P_1(a)) \cap f_2(P_2(a, \frac{a}{2})) \neq \emptyset$. Note that points of the form

$$f_1(x^1, \dots, x^k, 0, \dots, 0), \qquad x^1 \in \mathbb{R}_{>0},$$

are in $\mathcal{E} \setminus \mathrm{bd}(\mathcal{E})$. Also, points in $f_2(P_2(a, \frac{a}{2}))$ are in $\mathcal{R}(x_0', t_0, \tau)$ for some $x_0' \in S_0$. This proves the lemma. ■

Of course, there is a free interval version of the lemma which we merely state. Its proof is a simple adaptation of the previous proof. To state the result we denote

$$\mathscr{K}^{\pm}(\mu, x_0, t_0, \tau) = \mathrm{cl}(\mathrm{conv}\,\mathrm{cone}(\Phi(\mu, x_0, t_0, t_0, \tau)(\mathsf{T}_{\xi(t_0)} S_0) \cup K^{\pm}(\mu, x_0, t_0, \tau))),$$

the obvious adaptation of the fixed interval definition.

6.8 Lemma: (Free interval tangent cones and tangent half-spaces that are not separable) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let $S_0 = \Phi^{-1}(0)$ be a constraint set, let $x_0 \in S_0$, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, let $\mu \in \mathscr{U}(x_0, t_0, [t_0, t_1])$, and let $\tau \in (t_0, t_1) \cap \mathrm{Leb}(\mu, x_0, t_0, t_1)$. Let $\mathcal{E}$ be an edged set with $\xi(\mu, x_0, t_0, \tau) \in \mathrm{bd}(\mathcal{E})$ and suppose that the tangent half-space of $\mathcal{E}$ at $\xi(\mu, x_0, t_0, \tau)$ and the cone $\mathscr{K}^{\pm}(\mu, x_0, t_0, \tau)$ are not separable. Then there exists $x_0' \in S_0$ such that the set*

$$(\mathcal{E} \setminus \mathrm{bd}(\mathcal{E})) \cap \mathcal{R}(x_0', t_0)$$

is nonempty.

Now let us apply the preceding lemmata to the transversality conditions of the Maximum Principle. To do so we need some notation. We will now resume using the notation from the statement of the Maximum Principle, since we are assuming that we have an optimal trajectory. We denote

$$\hat{\mathsf{T}}_x S_a = \{(0, v) \in \mathbb{R} \oplus \mathbb{R}^n \mid v \in \mathsf{T}_x S_a\}, \qquad a \in \{1, 2\},$$
$$\hat{S}_1 = \{(x^0, x) \in \hat{\mathcal{X}} \mid x^0 \leq \xi_*^0(t_1),\ x \in S_1\},$$
$$\hat{S}_\tau = \{\hat{\xi}(\tau) \mid \dot{\hat{\xi}}(t) = \hat{f}(\hat{\xi}(t), \mu_*(t)),\ \hat{\xi}(t_1) \in \hat{S}_1\}, \qquad \tau \in (t_0, t_1).$$

Note that $\hat{S}_1$ and (therefore) $\hat{S}_\tau$ are edged sets. More precisely, in a neighbourhood of $\hat{\xi}_*(t_1)$ (resp. $\hat{\xi}_*(\tau)$), $\hat{S}_1$ (resp. $\hat{S}_\tau$) is an edged set. Note that

$$\mathsf{T}^+_{\hat{\xi}_*(t_1)} \hat{S}_1 = \mathrm{conv}\,\mathrm{cone}(\{(-1, 0)\} \cup \hat{\mathsf{T}}_{\hat{\xi}_*(t_1)} S_1),$$
$$\mathsf{T}^+_{\hat{\xi}_*(\tau)} \hat{S}_\tau = \mathrm{conv}\,\mathrm{cone}(\{(-1, 0)\} \cup \hat{\Phi}(\mu_*, \hat{x}_0, t_0, t_1, \tau)(\hat{\mathsf{T}}_{\hat{\xi}_*(t_1)} \hat{S}_1)),$$

the latter equality holding since $(-1,0)$ is transported to $(-1,0)$ by the variational equation for the extended system. We also denote $\hat{\mathscr{K}}(\mu_*,\hat{x}_0,t_0,t)$ (resp. $\hat{\mathscr{K}}^{\pm}(\mu_*,\hat{x}_0,t_0,t)$) as the closed convex hull of $\hat{\Phi}(\mu_*,\hat{x}_0,t_0,t_0,t)(\hat{\mathsf{T}}_{x_0}S_0)$ and $\hat{K}(\mu_*,x_0,t_0,t)$ (resp. $\hat{K}^{\pm}(\mu_*,x_0,t_0,t)$).

With this notation, we have the following result.

6.9 Lemma: (Separation of cones for transversality conditions) *Let $\Sigma=(\mathcal{X},f,U)$ be a control system, let L be a Lagrangian for Σ, and let $S_0,S_1\subset\mathcal{X}$ be constraint sets. Suppose that $(\xi_*,\mu_*)\in\mathscr{P}(\Sigma,L,S_0,S_1,[t_0,t_1])$ or that $(\xi_*,\mu_*)\in\mathscr{P}(\Sigma,L,S_0,S_1)$ is defined on $[t_0,t_1]$, respectively. Then the cones $\hat{\mathscr{K}}(\mu_*,\hat{x}_0,t_0,t_1)$ (resp. $\hat{\mathscr{K}}^{\pm}(\mu_*,\hat{x}_0,t_0,t_1)$) and $\mathsf{T}^{+}_{\hat{\xi}_*(t_1)}\hat{S}_1$ are separable.*

Proof: We will carry out the proof in the fixed interval case, the free interval case following along entirely similar lines.

Suppose that $\hat{\mathscr{K}}(\mu_*,\hat{x}_0,t_0,t_1)$ and $\mathsf{T}^{+}_{\hat{\xi}_*(t_1)}\hat{S}_1$ are not separable. Since

$$\hat{\mathscr{K}}(\mu_*,\hat{x}_0,t_0,t_1)=\cup_{t\in(t_0,t_1)}\hat{\mathscr{K}}(\mu_*,\hat{x}_0,t_0,t),$$

there exists $\tau\in\mathrm{Leb}(\mu_*,x_0,t_0,t_1)$ such that $\hat{\Phi}(\mu_*,\hat{x}_0,t_0,\tau,t_1)(\hat{\mathscr{K}}(\mu_*,\hat{x}_0,t_0,\tau))$ and $\mathsf{T}^{+}_{\hat{\xi}_*(t_1)}\hat{S}_1$ are not separable. Thus $\hat{\mathscr{K}}(\mu_*,\hat{x}_0,t_0,\tau)$ and $\hat{\Phi}(\mu_*,\hat{x}_0,t_0,t_1,\tau)(\mathsf{T}^{+}_{\hat{\xi}_*(t_1)}\hat{S}_1)$ are also not separable. The latter cone is simply $\mathsf{T}^{+}_{\hat{\xi}_*(\tau)}S_\tau$. We now apply Lemma 6.7 to conclude that there is a control $\tilde{\mu}$ defined on $[t_0,\tau]$ and a point $x_0'\in S_0$ such that $\xi(\tilde{\mu},x_0',t_0,\tau)\in\hat{S}_\tau\setminus\mathrm{bd}(\hat{S}_\tau)$. One can extend $\tilde{\mu}$ to a control μ defined on $[t_0,t_1]$ by having it agree with μ_* on $[\tau,t_1]$. The resulting control will steer x_0' to a point in $\hat{S}_1\setminus\mathrm{bd}(\hat{S}_1)$. This contradicts the optimality of (ξ_*,μ_*). ■

Finally, we show that the preceding lemma implies that the transversality conditions can be met. We work in the fixed interval setting; the free interval case follows in a similar vein. By Lemma 6.9 we know that the cones $\hat{\mathscr{K}}(\mu_*,\hat{x}_0,t_0,t_1)$ and $\mathsf{T}^{+}_{\hat{\xi}_*(t_1)}\hat{S}_1$ are separable. Choose $\hat{\lambda}_*(t_1)=(\lambda_*^0,\lambda_*(t_1))$ such that

$$\begin{aligned}\langle\hat{\lambda}_*(t_1),v\rangle&\le 0,\qquad \hat{v}\in\hat{\mathscr{K}}(\mu_*,\hat{x}_0,t_0,t_1),\\ \langle\hat{\lambda}_*(t_1),v\rangle&\ge 0,\qquad \hat{v}\in\mathsf{T}^{+}_{\hat{\xi}_*(t_1)}\hat{S}_1.\end{aligned}$$

Since $\hat{K}(\mu_*,\hat{x}_0,t_0,t_1)\subset\hat{\mathscr{K}}(\mu_*,\hat{x}_0,t_0,t_1)$ the adjoint response $t\mapsto\lambda_*(t)$ defined such that λ_* agrees with $\lambda_*(t_1)$ at time t_1 satisfies the conclusions of the Maximum Principle. Since $\hat{\mathsf{T}}_{\hat{\xi}(t_1)}S_1\subset\mathsf{T}^{+}_{\hat{\xi}_*(t_1)}\hat{S}_1$ we have

$$\langle\lambda_*(t_1),v\rangle\le 0,\qquad v\in\mathsf{T}_{\xi_*(t_1)}S_1.$$

Since $\mathsf{T}_{\xi_*(t_1)}S_1$ is a subspace, this means that $\lambda_*(t_1)$ is orthogonal to $\mathsf{T}_{\xi_*(\tau)}S_1$ which is the transversality condition at the terminal point. Since

$$\hat{\Phi}(\mu_*,\hat{x}_0,t_0,t_1)(\hat{\mathsf{T}}_{\hat{x}_0}S_0)\subset\hat{\mathscr{K}}(\mu_*,\hat{x}_0,t_0,t_1),$$

a similar argument shows that $\lambda_*(t_1)$ is orthogonal to $\Phi(\mu_*,x_0,t_0,t_1)(\mathsf{T}_{x_0}S_0)$. Now Proposition 4.5 implies that $\lambda_*(t_0)$ is orthogonal to $\mathsf{T}_{x_0}S_0$. This is the transversality condition at the initial point.

Chapter 7

A discussion of the Maximum Principle

Now that we have stated and proved the Maximum Principle, plus gone to some lengths to describe the elements in the proof, we are in the pleasant position of being able to say some things about what the Maximum Principle "means." There is much that one can say here, and we merely touch on a few of the more "obvious" things.

7.1. Normal and abnormal extremals

Extremals can have distinguishing characteristics that are interesting to study. Here we focus only on two aspects of a possible characterisation. The first concerns the distinction between "normal" and "abnormal." This has to do with the seemingly mysterious constant λ^0 appearing in the statement of the Maximum Principle.

7.1 Definition: (Normal and abnormal extremals) Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let L be a Lagrangian, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $S_0, S_1 \subset \mathcal{X}$ be sets.

(i) A controlled extremal (ξ, μ) for $\mathscr{P}(\Sigma, L, S_0, S_1)$ (resp. for $\mathscr{P}(\Sigma, L, S_0, S_1, [t_0, t_1])$) is ***normal*** if it is possible to satisfy the necessary conditions of Theorem 3.4 (resp. Theorem 3.5) with the constant λ^0 taken to be -1.

(ii) A controlled extremal (ξ, μ) for $\mathscr{P}(\Sigma, L, S_0, S_1)$ (resp. for $\mathscr{P}(\Sigma, L, S_0, S_1, [t_0, t_1])$) is ***possibly abnormal*** if it is possible to satisfy the necessary conditions of Theorem 3.4 (resp. Theorem 3.5) with the constant λ^0 taken to be 0.

(iii) A controlled extremal (ξ, μ) for $\mathscr{P}(\Sigma, L, S_0, S_1)$ (resp. for $\mathscr{P}(\Sigma, L, S_0, S_1, [t_0, t_1])$) is ***abnormal*** if it is *only* possible to satisfy the necessary conditions of Theorem 3.4 (resp. Theorem 3.5) with the constant λ^0 taken to be 0.

An extremal ξ for $\mathscr{P}(\Sigma, L, S_0, S_1)$ (resp. for $\mathscr{P}(\Sigma, L, S_0, S_1, [t_0, t_1])$) is ***normal***, ***possibly abnormal***, or ***abnormal*** if there exists an admissible control μ such that (ξ, μ) is a normal, possibly abnormal, or abnormal, respectively, controlled extremal for $\mathscr{P}(\Sigma, L, S_0, S_1)$ (resp. for $\mathscr{P}(\Sigma, L, S_0, S_1, [t_0, t_1])$). •

One needs to be a little careful to understand what the definition says. It does not say that an extremal is abnormal if the constant λ^0 in the statement of the Maximum Principle is zero; for the extremal to be abnormal the constant λ^0 *must* be zero. Some authors refer to

what we call "possibly abnormal" as "abnormal" and what we call "abnormal" as "strictly abnormal."

The existence of abnormal extremals seems a little tough to swallow at first glance. For example, an abnormal extremal ξ, along with its corresponding adjoint response λ, satisfy the differential equation

$$\begin{aligned}\dot{\xi}(t) &= f(\xi(t), \mu(t)),\\ \dot{\lambda}(t) &= -\boldsymbol{D}_1 f^T(\xi(t), \mu(t)) \cdot \lambda(t),\end{aligned}$$

since $\lambda^0 = 0$. The peculiar thing is that this differential equation, while being a necessary condition for the optimal control problem with Lagrangian L, is itself not dependent on L in any way. Despite the seeming implausibility of this, the notion of an abnormal extremal is actually easy to interpret. The proof of the following result is achieved by understanding the proof of Lemma 6.3. We encourage the reader to do just this in Exercise E7.1.

7.2 Proposition: (Characterisation of abnormality) *A controlled extremal (ξ, μ) for $\mathscr{P}(\Sigma, L, S_0, S_1, [t_0, t_1])$ (resp. for $\mathscr{P}(\Sigma, S_0, S_1)$ and defined on $[t_0, t_1]$) is abnormal if and only if $(-1, 0) \in \mathrm{bd}(\hat{K}(\mu, x_0, t_0, t_1))$ (resp. $(-1, 0) \in \mathrm{bd}(\hat{K}^{\pm}(\mu, x_0, t_0, t_1))$).*

It can happen that abnormal extremals may be actually optimal. We shall see an instance of this in Section 9.3. In Section 8.5 we consider the problem of abnormality in linear quadratic optimal control. In this section the reader can observe that possibly abnormal extremals can arise in trivial ways. We also observe that a phenomenon entirely similar to abnormality occurs in constrained minimisation in finite dimensions using the Lagrange multiplier method. We explore this in Exercise E3.2.

7.2. Regular and singular extremals

The next classification of extremals we consider accounts for the fact that one may not be able to determine the extremal controls explicitly just from the Maximum Principle. When this happens, the extremal is said to be "singular." However, it turns out that there are many possible flavours of singularity, and one must really make some sort of choice as to what one means by singular in a given instance, often guided by particular features of the problem one is considering.

But the essential idea of singularity is as follows. Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let L be a Lagrangian, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $S_0, S_1 \subset \mathcal{X}$ be subsets. Let (ξ, μ) be a controlled extremal for $\mathscr{P}(\Sigma, L, S_0, S_1)$ (resp. for $\mathscr{P}(\Sigma, L, S_0, S_1, [t_0, t_1])$) with $\lambda^0 \in \{0, -1\}$ and $\lambda\colon [t_0, t_1] \to \mathbb{R}^n$ as in the statement of Theorem 3.4 (resp. Theorem 3.5). At almost every time $t \in [t_0, t_1]$ the extremal control must satisfy

$$H_{\Sigma,L}(\xi(t), \lambda(t), \mu(t)) = H^{\max}_{\Sigma,L}(\xi(t), \lambda(t)).$$

The notion of singularity deals with the nature of the problem of solving the equation

$$H_{\Sigma,L}(x, p, \mu) = H^{\max}_{\Sigma,L}(x, p) \tag{7.1}$$

for the extremal control μ. If one is able to do this, then this gives the control as a function of x and p. In some cases, one can examine the matter of solving the equation (7.1) by

differentiating with respect to u and using the Implicit Function Theorem. However, in many optimal control problems differentiation is not valid, usually because the control set has a boundary and the extremal control may lie on the boundary. Thus one is led to consider the map, defined for $(x,p) \in \mathcal{X} \times \mathbb{R}^n$,

$$\begin{aligned} h_{\Sigma,L}(x,p)\colon\ U &\to \mathbb{R} \\ u &\mapsto H_{\Sigma,L}(x,p,u), \end{aligned}$$

and the corresponding problem of solving the equation

$$h_{\Sigma,L}(x,p) \cdot u = H^{\max}_{\Sigma,L}(x,p).$$

For $\alpha \in \mathbb{R}$ let us denote

$$\ker_\alpha(h_{\Sigma,L}(x,p)) = \{u \in U \mid\ h_{\Sigma,L}(x,p) \cdot u = \alpha\}.$$

Thus, for example, $\ker_\alpha(h_{\Sigma,L}(\xi(t),\lambda(t)))$ is the collection of possible values of the extremal control in the case when $\alpha = H^{\max}_{\Sigma,L}(\xi(t),\lambda(t))$.

The following example shows that it might trivially be the case that the extremal control will *never* be uniquely determined by asking that it lie in $\ker_\alpha(h_{\Sigma,L}(t))$, no matter the value of α.

7.3 Example: (An example where all extremals are "singular") Let $\mathcal{X} = \mathbb{R}^2$, $m = 2$, and take

$$f((x^1,x^2),(u^1,u^2)) = (x^1, x^2 + u^1), \quad L((x^1,x^2),(u^1,u^2)) = \tfrac{1}{2}((x^1)^2 + (x^2)^2) + \tfrac{1}{2}(u^1)^2.$$

Since f and L do not depend on u^2, it follows that for any $(x,p) \in \mathcal{X} \times \mathbb{R}^2$ and for any $\alpha \in \mathbb{R}$ we will have

$$\{(0,u^2) \mid\ u^2 \in \mathbb{R}\} \subset \ker_\alpha(h_{\Sigma,L}(x,p)).$$

Thus we will never be able to solve uniquely for the extremal control. •

This example is obviously stupid and contrived; we cannot solve uniquely for the extremal control because the control set is "too large." However, it does point the way towards some sort of reasonable notion of singularity. For $x \in \mathcal{X}$ we define a map

$$\begin{aligned} F_\Sigma(x)\colon\ U &\to \mathbb{R}^n \\ u &\mapsto f(x,u), \end{aligned}$$

and for $v \in \mathbb{R}^n$ define

$$\ker_v(F_\Sigma(x)) = \{u \in U \mid\ F_\Sigma(x) \cdot u = v\}.$$

With this notation we make the following definition of singularity.

7.4 Definition: (Regular and singular extremals) Let $\Sigma = (\mathcal{X}, f, U)$ be a control system, let L be a Lagrangian, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $S_0, S_1 \subset \mathcal{X}$ be sets. Let (ξ,μ) be a controlled extremal for $\mathscr{P}(\Sigma, L, S_0, S_1)$ (resp. for $\mathscr{P}(\Sigma, L, S_0, S_1, [t_0,t_1])$) with $\lambda^0 \in \{0,-1\}$ and $\lambda\colon [t_0,t_1] \to \mathbb{R}^n$ as in the statement of Theorem 3.4 (resp. Theorem 3.5). Suppose that $t \in [t_0,t_1]$ has the property that

$$H_{\Sigma,\lambda^0 L}(\xi(t),\lambda(t),\mu(t)) = H^{\max}_{\Sigma,\lambda^0 L}(\xi(t),\lambda(t))$$

(note that this holds for almost every t).

(i) The controlled extremal (ξ, μ) is ***regular*** at t if

$$\ker_{H^{\max}_{\Sigma,L}(\xi(t),\lambda(t))} h_{\Sigma,L}(\xi(t), \lambda(t)) = \ker_{f(\xi(t),\mu(t))} F_\Sigma(\xi(t)).$$

(ii) The controlled extremal (ξ, μ) is ***singular*** at t if it is not regular at t.

(iii) If $I \subset [t_0, t_1]$ is a subinterval of $[t_0, t_1]$ then the controlled extremal (ξ, μ) is ***regular*** (resp. ***singular***) on I if it is regular (resp. singular) for almost every $t \in I$. •

The idea of a regular extremal, according to our definition, is that one can solve for the extremal control in a way that is unique in the sense that any two possible extremal controls will give the same dynamics for the system. For a singular extremal, on the other hand, one may have another extremal control for which the dynamics of the system are actually different. In these cases, to discriminate between different extremal controls, one must consider so-called higher-order necessary conditions; see, for example [Krener 1977]. A great deal has been done along these lines for second-order conditions by Agrachev and various coauthors (see, for example, [Agrachev and Sachkov 2004]).

While we have attempted to make our definition quite general, there are other possible notions of singularity that come up, particularly in time-optimal control. We refer the reader to [Bonnard and Chyba 2003] for a thorough discussion of singular extremals.

This is all we shall say here about the interesting and important notion of singularity.

7.3. Tangent cones and linearisation

Much of the Maximum Principle rests on the fact that the tangent cones $K(\mu, x_0, t_0, t)$ and $K^{\pm}(\mu, x_0, t_0, t)$ approximate the reachable set in the sense given by Lemmata 5.10 and 5.11, respectively. In order to get some sense of what this approximation means, we shall consider the tangent cones in a special case. The discussion in this section is quite unrelated to optimal control, and should rather be thought of as an attempt to understand the tangent cones.

We consider a control system $\Sigma = (\mathcal{X}, f, U)$ and we suppose that $0 \in U$. We also let $(x_0, 0) \in \mathcal{X} \times U$ be a ***controlled equilibrium***, by which we mean that $f(x_0, 0) = 0$. Thus $(t \mapsto x_0, t \mapsto 0)$ is a controlled trajectory for the system. Let us abbreviate the zero control, $t \mapsto 0$, merely by 0 and suppose this to be defined on an interval $[t_0, t_1]$. Note that the variational equation is

$$\dot{v}(t) = \boldsymbol{D}_1 f(x_0, 0) \cdot v(t),$$

which is simply the standard Jacobian linearisation of the system about x_0 subject to the zero control. Let us abbreviate $A = \boldsymbol{D}_1 f(x_0, 0)$ so that, following Example 1.3,

$$\Phi(0, x_0, t_0, \tau, t) \cdot v = \exp(A(t - \tau)) \cdot v.$$

A fixed interval needle variation at $\tau \in [t_0, t_1]$ will then have form $v_\theta = l_\theta f(x_0, u)$ for $u \in U$, using the fact that $f(x_0, 0) = 0$. Therefore, the fixed interval tangent cone $K(0, x_0, t_0, t)$ is the closed convex hull of the set of vectors of the form $\exp(A(t - \tau)) \cdot f(x_0, u)$ for $\tau \in [t_0, t]$ and for $u \in U$.

The following result encodes the character of the fixed interval tangent cone in the special case that is most commonly encountered.

7.5 Proposition: (Tangent cones and linearisation) *Suppose that*

$$\operatorname{conv}\operatorname{cone}(\{f(x_0,u) \mid u \in U\})$$

is a subspace. Then, with the constructions above, $K(0,x_0,t_0,t)$ is the smallest subspace containing $\{f(x_0,u) \mid u \in U\}$ and which is invariant under A. Explicitly,

$$K(0,x_0,t_0,t) = \operatorname{conv}\operatorname{cone}(\{A^j f(x_0,u) \mid j \in \mathbb{Z}_{\geq 0},\ u \in U\}).$$

Proof: We first prove a lemma which is sort of interesting on its own.

1 Lemma: *Let $V \subset \mathbb{R}^n$ be a subspace and let $A \in \mathrm{L}(\mathbb{R}^n;\mathbb{R}^n)$. Then the following statements are equivalent:*

(i) V is invariant under A;

(ii) there exists an interval $I \subset \mathbb{R}$ with nonempty interior such that V is invariant under $\exp(At)$ for each $t \in I$.

Proof: Suppose that V is A-invariant, let $x \in V$, and let $I \subset \mathbb{R}_{\geq 0}$ be an interval with nonempty interior. Then $A^k x \in V$ for each $k \in \mathbb{Z}_{\geq 0}$. Thus, for each $N \in \mathbb{Z}_{>0}$ and for each $t \in I$ we have

$$\sum_{k=1}^{N} \frac{t^k A^k}{k!} x \in V$$

since V is a subspace. Since V is closed we have

$$\exp(At)x = \lim_{N\to\infty} \sum_{k=1}^{N} \frac{t^k A^k}{k!} x \in V.$$

Now suppose that $I \subset \mathbb{R}_{\geq 0}$ is an interval with nonempty interior and that $\exp(At)x \in V$ for every $t \in I$ and $x \in V$. Let $t_0 \in \operatorname{int}(I)$ and let $x \in V$ so that $\exp(-At_0)x \in V$ (we use that fact that V is invariant under an invertible linear map L if and only if it is invariant under L^{-1}). Thus the curve $t \mapsto \exp(At)\exp(-At_0)x$ takes values in V for all t sufficiently near t_0. Since V is a subspace we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=t_0} \exp(At)\exp(-At_0)x = Ax \in V,$$

and so V is A-invariant. ▼

From the lemma, the proposition follows immediately since $K(0,x_0,t_0,t)$ is clearly the smallest subspace containing $\{f(x_0,u) \mid u \in U\}$ which is invariant under $\exp(A(\tau - t_0))$ for $\tau \in [t_0,t]$. ■

For control-affine systems one can make this result even more compelling. Thus we let $\Sigma = (\mathcal{X}, f, U)$ be a control-affine system, i.e.,

$$f(x,u) = f_0(x) + f_1(x) \cdot u,$$

for maps $f_0 \colon \mathcal{X} \to \mathbb{R}^n$ and $f_1 \colon \mathcal{X} \to \mathrm{L}(\mathbb{R}^m;\mathbb{R}^n)$. Given $x_0 \in \mathcal{X}$ such that $f_0(x_0) = 0$ we then have, using our notation from above, $A = \boldsymbol{D}f_0(x_0)$. Let us also define $B = f_1(x_0)$. We then have the following result which follows immediately from Proposition 7.5.

7.6 Corollary: (Tangent cones and linearisation for control-affine systems) *For a control-affine system with the constructions above and with*

$$0 \in \operatorname{int}(\operatorname{conv}(\{B(u) \mid u \in U\})),$$

$K(0, x_0, 0, t)$ *is the smallest* A*-invariant subspace of* $\mathbb{R}^n$ *containing* $\operatorname{image}(B)$.

For readers familiar with linear control theory, and more specifically the geometric point of view adopted in, say, [Wonham 1985], we comment that this means that $K(0, x_0, 0, t)$ is equal to the columnspace of the ***Kalman controllability matrix***

$$\left[\; B \mid AB \mid A^2B \mid \cdots \mid A^{n-1}B \; \right].$$

The bottom line for the whole discussion is the following somewhat imprecisely stated punchline.

> *The cone* $K(\mu, x_0, t_0, t_1)$ *(or* $K^{\pm}(\mu, x_0, t_0, t_1)$*) having a nonempty interior is a generalisation of the system having a controllable linearisation.*

7.7 Remark: (The relationship between controllability and optimal control) There is a saying that goes, "Necessary conditions for optimality are sufficient conditions for controllability, and vice versa." The punchline above illustrates what this saying might mean as concerns the necessary conditions of the Maximum Principle and the sufficient condition for controllability given by the controllability of the linearisation. Expansions of this saying beyond the Maximum Principle and beyond linearised controllability require so-called "higher-order" conditions. Such things are explored, for example, by Krener [1977]. Here let us merely remark that this is related to the notion of singular extremals. •

7.4. Differential geometric formulations

A much better conceptual understanding of the Maximum Principle is possible if all of the constructions surrounding it are carried out in the framework of differential geometry. Many of the constructions are simply more natural and intuitive when stated in terms of manifolds, tangent spaces, vector fields, one-forms, etc. We have stopped short of doing this explicitly in our development in order to make the presentation accessible (note easy, note) to those who do not know any differential geometry. In this section we shall sketch how one can translate at least some of the constructions into the differential geometric world. Readers familiar with differential geometry can then easily complete the process on their own. For readers unfamiliar with differential geometry, we can only recommend that, as a matter of course, they take steps to undo their plight.

Sussmann has devoted some effort to understanding the geometry behind the Maximum Principle. A good account of some part of this can be found in the paper [Sussmann 1997].

7.4.1. Control systems. The first abstraction one makes is to use a manifold M as the state space. If we keep the control set U as a subset of $\mathbb{R}^m$, then the map f giving the system dynamics becomes a map $f\colon \mathsf{M} \times \operatorname{cl}(U) \to \mathsf{TM}$ with the property that $f(x, u) \in \mathsf{T}_x\mathsf{M}$ for every $(x, u) \in \mathsf{M} \times \operatorname{cl}(U)$. The system equations then take the form

$$\xi'(t) = f(\xi(t), \mu(t)).$$

Different and/or more general notions of control systems are possible, but we do not elect to pursue this here. The definitions of various classes of controls and controlled trajectories can be made exactly as in our non-differential geometric formulation. One must be careful to understand what is meant by an absolutely continuous curve on a manifold, but this is easily done.

One can also define Lagrangians for control systems on manifolds in the same manner as we have done. This, of course, then allows one to formulate analogues of the standard optimal control problems, Problems 1.7 and 1.8.

7.4.2. The Maximum Principle. The Maximum Principle can also easily be stated for control systems and corresponding optimal control problems on manifolds. For example, the Hamiltonian can be defined exactly as we have done in our treatment. The only thing one needs to be aware of is that the manifold will not generally have an inner product, and so in the definition of the Hamiltonian, one should think of "p" as being an element of the cotangent space at $x \in \mathsf{M}$. Thus the Hamiltonian and the maximum Hamiltonian are functions on $\mathsf{T}^*\mathsf{M} \times U$ and $\mathsf{T}^*\mathsf{M}$, respectively. Some readers may be aware of the fact that the cotangent bundle has a canonical symplectic structure. This introduces a fairly deep connection with optimal control to symplectic geometry. Indeed, this connection lies behind the "derivation" of the Maximum Principle from the calculus of variations in Section 2.4.

The other part of the Maximum Principle that one must be careful to understand properly is the adjoint response. In the geometric formulation, the adjoint response is an absolutely continuous one-form field along the trajectory of the system. The differential equation for the adjoint response as given in Definition 3.2 is the Hamiltonian vector field associated with the system's natural Hamiltonian and the canonical cotangent bundle symplectic structure. We will see in a moment another interpretation of the differential equation for the adjoint response.

The final ingredient in the Maximum Principle is the transversality conditions. First of all, in the differential geometric formulation one should replace our clumsy notion of a "smooth constraint set" with the more natural notion of a submanifold. In this case, the transversality condition is that the adjoint response should annihilate the tangent spaces to S_0 and S_1 at the endpoints.

7.4.3. The variational and adjoint equations, and needle variations. The variational and adjoint equations have analogues in the differential geometric setting. To make this development precise requires a pile of notation, so we shall merely sketch things here. We refer to [Lewis and Tyner 2003, Sussmann 1997] for details.

The variational equation describes, as is made precise in Proposition 4.3, the linearisation of a control system along a trajectory. There are various ways to think about this in a geometric setting. Two natural ways are the following.

1. *The tangent lift:* Given a time-dependent vector field $X : \mathbb{R} \times \mathsf{M} \to \mathsf{TM}$ (such as might be obtained by substituting a specific control into the equations for a control system), one can define its ***tangent lift***, a time-dependent vector field on TM, in coordinates by

$$X^T(t, v_x) = X^i(t, x)\frac{\partial}{\partial x^i} + \frac{\partial X^i}{\partial x^j}(t, x)v^j\frac{\partial}{\partial v^i}.$$

The integral curves of this vector field are easily seen to be vector fields along integral curves of X. One can show that these integral curves are related to infinitesimal variations of integral curves of X exactly in the manner described by Proposition 4.3.

2. *The differential operator point of view:* One can think of the variational equation as a differential operator on the absolutely continuous vector fields along a trajectory of the system. There are many such possible choices of differential operator (for example, one such differential operator can be assigned in a natural way given an affine connection on M). There is, however, a natural one associated to the integral curve of a vector field. This natural one corresponds to what is often called the "Lie drag." Moreover, this natural one also agrees with the use of the tangent lift to give the variational equation.

In this way, after some technical work, one can understand the variational equation in the differential geometric sense.

For the adjoint equation, there are similarly (at least) two alternative and equivalent ways of thinking of it.

1. *The cotangent lift:* For a time-dependent vector field X on M its ***cotangent lift*** is the time-dependent vector field on $\mathsf{T}^*\mathsf{M}$ defined in coordinates by

$$X^{T^*}(t, \alpha_x) = X^i(t, x)\frac{\partial}{\partial x^i} - \frac{\partial X^j}{\partial x^i}(t, x)p_j\frac{\partial}{\partial p_i}.$$

This vector field can be defined intrinsically as the Hamiltonian vector field for the time-dependent Hamiltonian $H_X(t, \alpha_x) = \langle \alpha_x; X(t, x)\rangle$. Similarly to the tangent lift, the integral curves of the cotangent lift are one-form fields along integral curves of X. These one-form fields satisfy the adjoint equation in our sense in local coordinates.

2. *The differential operator point of view:* A differential operator on vector fields along a curve induces in a natural way a differential operator on tensor fields along the same curve. In particular, the variational equation induces a differential operator on the one-form fields along a curve. If one chooses the natural differential operator for the vector fields along a trajectory of a control system, the corresponding natural differential operator for one-form fields is exactly the adjoint equation.

With the variational equation understood, one can now define the various notions of needle variations just as we have done, and prove the similar results concerning their existence and form.

7.4.4. Tangent cones. The definitions of the tangent cones can be made pretty much exactly as we have done in our treatment. The only thing one needs to be aware of is that, in the geometric formulation, the tangent cones $K(\mu, x_0, t_0, t)$ and $K^{\pm}(\mu, x_0, t_0, t)$ are subsets of $\mathsf{T}_{\xi(t)}\mathsf{M}$ where $\xi = \xi(\mu, x_0, t_0, \cdot)$. This is one of the places in the development where the geometric formulation really helps to clarify what one is doing. In our non-geometric development, one get confused about where these cones really live, and this makes the interpretation of Lemmata 5.10 and 5.11 more difficult that it should be. Indeed, in Exercise E7.5 we invite the reader to provide geometric formulations of lemmata. We recommend that the reader do this in order to really understand what is really going on here. Once one has done this, then a deeper understanding of the very important Theorems 5.16 and 5.18 will also follow.

Exercises

E7.1 Prove Proposition 7.2.

E7.2 Consider the control system $\Sigma = (\mathcal{X}, f, U)$ defined by

(i) $\mathcal{X} = \mathbb{R}^2$,

(ii) $f(x, u) = (x^2, u)$,

(iii) $U = [-1, 1]$.

Define a Lagrangian for Σ by $L((x^1, x^2), u) = 1$. We consider the problem $\mathscr{P} = \mathscr{P}(\Sigma, L, \{(0,0)\}, \{(x_0^1, x_0^2)\})$ of steering the system from the origin to a point $(x_0^1, x_0^1) \in \mathcal{X}$ in minimum time.

While our main interest is in extremal trajectories emanating from $(0, 0)$, many of the questions below actually are true for more general extremal trajectories. Therefore, we ask that you take $(0, 0)$ as the original point only in those cases where you are explicitly asked to do so.

(a) Determine the extremal control u as a function of (x^1, x^2, p^1, p^2).

(b) Show that if a curve $t \mapsto (x^1(t), x^2(t))$ solves $\mathscr{P}$ then there exists a curve $t \mapsto (p^1(t), p^2(t))$ such that

$$\dot{x}^1 = x^2, \quad \dot{x}^2 = u, \quad \dot{p}^1 = 0, \quad \dot{p}^2 = -p^1.$$

(c) Based on your answer in (a), why does it make sense to call the set

$$S = \{(x^1, x^2, p^1, p^2) \in \mathcal{X} \times \mathbb{R}^2 \mid p_2 = 0\}$$

the ***switching surface***?

(d) Show that every solution to $\mathscr{P}$ intersects S at most once.

(e) Show that the extremals then satisfy the equations

$$\dot{x}^1 = x^2, \quad \dot{x}^2 = \pm 1.$$

Sketch the solutions to these differential equations for fixed sign of the control.

(f) For the initial condition $(0, 0)$ and a given final condition (x_0^1, x_0^2), indicate how to determine whether the control starts as $u = +1$ or $u = -1$, and determine the time at which a switch must occur.

(g) Show that all abnormal extremals are contained in the surface $p^1 x^2 + \operatorname{sign}(p^2) p^2 = 0$.

(h) Show that if t_s is the switching time for an abnormal extremal then $x^2(t_s) = 0$.

(i) Argue that any abnormal extremal originating from $(0, 0)$ contains no switches.

(j) Sketch the set of points reachable from $(0, 0)$ in time $T > 0$. Argue based on the material in Section 7.1 that the extremal trajectories originating from $(0, 0)$ are normal.

The following exercise exhibits the so-called "Fuller phenomenon," first pointed out by Fuller [1960].

E7.3 Consider the control system $\Sigma = (\mathcal{X}, f, U)$ defined by

(i) $\mathcal{X} = \mathbb{R}^2$,

(ii) $f(x, u) = (x^2, u)$,

(iii) $U = [-1, 1]$.

Define a Lagrangian for Σ by $L((x^1, x^2), u) = \frac{1}{2}(x^1)^2$. We consider the problem $\mathscr{P} \triangleq \mathscr{P}(\Sigma, L, \{(x_0^1, x_0^2)\}, \{(0, 0)\}, [0, T])$ of steering the system from a point $(x_0^1, x_0^2) \in \mathcal{X}$ to the origin in time T while minimising the integral of the square of the distance of x^1 from 0.

(a) Determine the extremal control u as a function of (x^1, x^2, p^1, p^2).

(b) Show that if a curve $t \mapsto (x^1(t), x^2(t))$ solves $\mathscr{P}$ then there exists a curve $t \mapsto (p^1(t), p^2(t))$ and $\lambda^0 \in \{0, -1\}$ such that

$$\dot{x}^1 = x^2, \quad \dot{x}^2 = u, \quad \dot{p}^1 = -\lambda^0 x^1, \quad \dot{p}^2 = -p^1.$$

(c) Based on your answer in (a), why does it make sense to call the set

$$S = \{(x^1, x^2, p^1, p^2) \in \mathcal{X} \times \mathbb{R}^2 \mid p^2 = 0\}$$

the ***switching surface***?

(d) Show that the extremals then satisfy the equations

$$\dot{x}^1 = x^2, \quad \dot{x}^2 = \pm 1.$$

Sketch the solutions to these differential equations for fixed sign of the control.

(e) Show that it is not possible for an extremal to start on $S \setminus \{((0, 0), (0, 0))\}$ and reach $((0, 0), (0, 0)) \in \mathcal{X} \times \mathbb{R}^2$ without again passing through $S \setminus \{((0, 0), (0, 0))\}$ (just consider the case when $\lambda^0 = -1$).

E7.4 Consider a control system $\Sigma = (\mathcal{X}, f, U)$,

$$f(x, u) = f_0(x) + f_1(x) \cdot u,$$

with $U = \mathbb{R}^m$ and with $f_1(x)$ injective for each $x \in \mathcal{X}$. Show that all abnormal extremals associated with a given Lagrangian L are singular.

E7.5 (For readers who know some differential geometry.) State Lemmata 5.10 and 5.11 in a differential geometric formulation for control systems.

Chapter 8

Linear quadratic optimal control

One of the more spectacular successes of the theory of optimal control concerns linear systems with quadratic Lagrangians. A certain variant of this problem, which we discuss in Section 8.4, leads rather surprisingly to a stabilising state feedback law for linear systems. This state feedback law is so effective in practice that it sees many applications in that place some refer to as "the real world."

We shall not attempt to be as exhaustive in our treatment in this section as we have been up to this point. Our intent is principally to illustrate the value of the Maximum Principle in investigating problems in optimal control. As we shall see, even in this case it is typical that work remains to be done after one applies the Maximum Principle.

The ideas in this chapter originate with the seminal paper of Kalman [1960], and have been extensively developed in the literature. They now form a part of almost any graduate course on "linear systems theory." There are many texts for such a course, indeed too many to make it sensible to mention even one (almost such as might be the case with, say, texts on calculus or linear algebra). We will let the reader discover which one they like best.

8.1. Problem formulation

We consider a linear system $\Sigma = (A, B, U)$ and take $U = \mathbb{R}^m$. Thus the dynamics of the system are governed by

$$\dot{\xi}(t) = A(\xi(t)) + B(\mu(t)).$$

In order to eliminate uninteresting special cases, we suppose make the following assumption throughout this section.

8.1 Assumption: (Property of B) The matrix B is full rank and $m \in \mathbb{Z}_{>0}$. •

We next consider symmetric matrices $Q \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ and $R \in \mathrm{L}(\mathbb{R}^m; \mathbb{R}^m)$ with R positive-definite (denoted $R > 0$). It is sometimes possible to relax the definiteness condition on R. However, this adds complications to the problem that we will simply not consider here. The Lagrangian we use is then given by

$$L(x, u) = \tfrac{1}{2} x^T Q x + \tfrac{1}{2} u^T R u.$$

Note that the cost for a controlled trajectory (ξ, μ) defined on $[t_0, t_1]$ is

$$J_{\Sigma,L}(\xi, \mu) = \int_{t_0}^{t_1} (\tfrac{1}{2} \xi(t)^T Q \xi(t) + \tfrac{1}{2} \mu(t)^T R \mu(t)) \, \mathrm{d}t,$$

which can be thought of as the sum of weighted L^2-norms of the state and the control. This way of thinking of the cost function opens the door to a rich function analytic way of thinking about linear quadratic optimal control. We will have nothing to say about this, but refer to [Dullerud and Paganini 1999] as a fairly recent reference in an area where much has been published.

For a given linear system $\Sigma = (A, B, \mathbb{R}^m)$ and given symmetric matrices Q and $R > 0$, we will be interested in the fixed interval optimal control problem of steering from an initial state $x_0 \in \mathbb{R}^n$ at time t_0 to an unspecified final state at time t_1 (thus the final constraint set is $S_1 = \mathbb{R}^n$). Let us denote the set of optimal controlled trajectories for this problem by $\mathscr{P}(A, B, Q, R, x_0, t_0, t_1)$.

8.2. The necessary conditions of the Maximum Principle

It is fairly easy to characterise the controlled extremals for the linear quadratic optimal control problem.

8.2 Proposition: (The Maximum Principle for linear quadratic optimal control) *Let $\Sigma = (A, B, \mathbb{R}^m)$ be a linear control system, let $Q \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ and $R \in \mathrm{L}(\mathbb{R}^m; \mathbb{R}^m)$ be symmetric with $R > 0$, let $x_0 \in \mathbb{R}^n$, and let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$. If $(\xi_*, \mu_*) \in \mathscr{P}(A, B, Q, R, x_0, t_0, t_1)$ then there exists a map $\lambda_* \colon [t_0, t_1] \to \mathbb{R}^n$ such that together ξ_* and λ_* satisfy the initial/final value problem*

$$\begin{bmatrix} \dot{\xi}_*(t) \\ \dot{\lambda}_*(t) \end{bmatrix} = \begin{bmatrix} A & -S \\ Q & -A^T \end{bmatrix} \begin{bmatrix} \xi_*(t) \\ \lambda_*(t) \end{bmatrix}, \quad \xi_*(t_0) = x_0, \ \lambda_*(t_1) = 0,$$

where $S = BR^{-1}B^T$.

Proof: We first note that we have $\lambda_*(t_1) = 0$ by the transversality conditions of the Maximum Principle. In this case, the fact that the total adjoint response must be nonzero ensures that $\lambda_*^0 = -1$. Thus the maximum Hamiltonian is

$$H_{\Sigma,-L}(x, p, u) = \langle p, Ax + Bu \rangle - (\tfrac{1}{2} x^T Q x + \tfrac{1}{2} u^T R u),$$

which is a quadratic function of u with a negative-definite second derivative. Thus the unique maximum occurs at the point where the derivative of the Hamiltonian with respect to u vanishes. That is to say, for almost every $t \in [t_0, t_1]$ we have

$$\mu_*(t) = -R^{-1} B^T \lambda_*(t),$$

as may be verified by a direct computation. Since the adjoint equations for the extended system are

$$\begin{aligned} \dot{\xi}^0(t) &= \tfrac{1}{2} \xi_*^T(t) Q \xi_*(t) + \tfrac{1}{2} \mu_*^T(t) R \mu_*(t), \\ \dot{\xi}(t) &= A(\xi(t)) + B(\mu(t)), \\ \dot{\lambda}^0(t) &= 0, \\ \dot{\lambda}(t) &= Q(\xi(t)) - A^T(\lambda(t)), \end{aligned}$$

one can substitute the form of the optimal control into the second of these equations to get the differential equations in the statement of the result. Obviously we must have $\xi_*(t_0) = x_0$. That $\lambda_*(t_1) = 0$ follows since the terminal condition is unspecified (i.e., $S_1 = \mathbb{R}^n$). ■

8.3. The rôle of the Riccati differential equation

In this section we introduce an important player in the theory of linear quadratic optimal control, the Riccati equation. For symmetric matrices Q and S the ***Riccati differential equation*** is the following differential equation for the $n \times n$ matrix function $F\colon I \to \mathrm{L}(\mathbb{R}^n;\mathbb{R}^n)$:

$$\dot{F}(t) + F(t)A + A^T F(t) - F(t)SF(t) + Q = 0.$$

This is a nonlinear differential equation so it is not trivial to characterise its solutions in a useful way. In particular, one must typically obtain solutions numerically. For us, the main question of interest will be, "When does the Riccati equation possess solutions?" As we shall see, this is intimately connected with the linear quadratic optimal control problem.

Indeed, we have the following theorem.

8.3 Theorem: (Characterisation of solutions of linear quadratic optimal control problem) *Let $\Sigma = (A, B, \mathbb{R}^m)$ be a linear control system, let $Q \in \mathrm{L}(\mathbb{R}^n;\mathbb{R}^n)$ and $R \in \mathrm{L}(\mathbb{R}^m;\mathbb{R}^m)$ be symmetric with $R > 0$, and let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$. Then the following statements are equivalent:*

(i) for each $t_0' \in [t_0, t_1]$ and $x_0 \in \mathbb{R}^n$, $\mathscr{P}(A, B, Q, R, x_0, t_0', t_1) \neq \emptyset$;

(ii) for each $t_0' \in [t_0, t_1]$ and $x_0 \in \mathbb{R}^n$, $\mathscr{P}(A, B, Q, R, x_0, t_0', t_1)$ is a singleton;

(iii) the solution of the Riccati differential equation exists and is bounded on $[t_0, t_1]$ when subject to the final condition $F(t_1) = 0_{n\times n}$;

(iv) the solution to the final value problem

$$\begin{bmatrix}\dot{\Xi}(t)\\ \dot{\Lambda}(t)\end{bmatrix} = \begin{bmatrix} A & -S\\ -Q & -A^T\end{bmatrix}\begin{bmatrix}\Xi(t)\\ \Lambda(t)\end{bmatrix}, \quad \Xi(t_1) = I_n,\ \Lambda(t_1) = 0_{n\times n}, \tag{8.1}$$

for matrices $\Xi, \Lambda \in \mathrm{L}(\mathbb{R}^n;\mathbb{R}^n)$ has the property that $\det \Xi(t) \neq 0$ for each $t \in [t_0, t_1]$.

Proof: The following "completing the squares" lemma, while simple, is integral to the problem.

1 Lemma: *Let $F\colon [t_0, t_1] \to \mathrm{L}(\mathbb{R};\mathbb{R})$ be absolutely continuous and let (ξ, μ) be a controlled trajectory for $(A, B, \mathbb{R}^m)$. Then*

$$\int_{t_0}^{t_1} \begin{bmatrix}\mu(t)\\ \xi(t)\end{bmatrix}^T \begin{bmatrix} 0_{n\times n} & B^T F(t)\\ F(t)B & \dot{F}(t) + A^T F(t) + F(t)A\end{bmatrix}\begin{bmatrix}\mu(t)\\ \xi(t)\end{bmatrix} \mathrm{d}t$$
$$= (\xi(t_1)^T F(t_1)\xi(t_1) - \xi(t_0)^T F(t_0)\xi(t_0)).$$

Proof: We have

$$\frac{\mathrm{d}}{\mathrm{d}t}(\xi(t)^T F(t)\xi(t)) = \dot{\xi}(t)^T F(t)\xi(t) + \xi(t)^T \dot{F}(t)\xi(t) + \xi(t)F(t)\dot{\xi}(t),$$

and the result follows by integrating both sides and substituting $\dot{\xi}(t) = A(\xi(t)) + B(\mu(t))$. ▼

(i) $\implies$ (ii) According to the proof of Proposition 8.2, if $(\xi_*, \mu_*) \in \mathscr{P}(A, B, Q, R, x_0, t_0', t_1)$ then

$$\mu_*(t) = -R^{-1}B^T\lambda_*(t),$$

where λ_* is determined by it, along with ξ_*, satisfying the initial/final value problem in the statement of Proposition 8.2.

(ii) $\implies$ (i) This is trivial.

(iii) $\implies$ (i) With F being the bounded solution to the Riccati differential equation with $F(t_1) = 0_{n\times n}$, we add the identity from Lemma 1 to the cost function, i.e., add zero to the cost function. After some straightforward manipulations we determine that the cost function for $\mathscr{P}(A, B, Q, R, x_0, t_0', t_1)$ is

$$\int_{t_0'}^{t_1} \|\mu(t) + R^{-1}B^TF(t)\xi(t)\|^2 \mathrm{d}t + \xi(t_0')^TF(t_0')\xi(t_0').$$

Thus we see that taking

$$\mu_*(t) = -R^{-1}B^TF(t)\xi_*(t),$$

immediately renders (ξ_*, μ_*) as an optimal trajectory.

(i) $\implies$ (iii) Suppose that there exists $t_0' \in [t_0, t_1]$ such that $\lim_{t\downarrow t_0'} \|F(t)\| = \infty$ where F is the solution to the Riccati differential equation with $F(t_1) = 0_{n\times n}$. We suppose that the solution is bounded on $[\tau, t_1]$ for any $\tau \in [t_0', t_1]$, which we may do using the local existence of solutions of the Riccati differential equation near t_1. We let $x_0 \in \mathbb{R}^n$ have the property that

$$\lim_{t\downarrow t_0'} |x_0^TF(t)x_0| = \infty.$$

Using our computations from the preceding part of the proof, if (ξ, μ) is a controlled trajectory defined on $[t_0', t_1]$ satisfying $\xi(t_0') = x_0$, then we have the cost from τ to t_1 as bounded below by $\xi(\tau)^TF(\tau)\xi(\tau)$. Thus

$$\lim_{\tau\downarrow t_0'} |\xi(\tau)^TF(\tau)\xi(\tau)| = \infty,$$

meaning that $\mathscr{P}(A, B, Q, R, x_0, t_0', t_1) = \emptyset$.

(i) $\implies$ (iv) We first use a lemma.

2 Lemma: *For curves $\xi, \lambda\colon [t_0, t_1] \to \mathbb{R}^n$, consider the following statements:*

(i) the curves are a solution to the initial/final value problem

$$\begin{bmatrix}\dot\xi(t)\\ \dot\lambda(t)\end{bmatrix} = \begin{bmatrix}A & -S\\ -Q & -A^T\end{bmatrix}\begin{bmatrix}\xi(t)\\ \lambda(t)\end{bmatrix}, \quad \xi(t_0) = x_0,\ \lambda(t_1) = 0;$$

(ii) $\xi(t) = \Xi(t)\xi(t_1)$ and $\lambda(t) = \Lambda(t)\xi(t_1)$ where Ξ and Λ satisfy (8.1).

Then (i) $\implies$ (ii).

Proof: With ξ and λ defined as satisfying the initial/final value problem from part (i), define $\hat\xi(t) = \Xi(t)\xi(t_1)$ and $\hat\lambda(t) = \Lambda(t)\xi(t_1)$ where Ξ and Λ satisfy the final value problem from part (ii). One can then verify by direct computation that ξ and λ together satisfy the same differential equations as $\hat\xi$ and $\hat\lambda$. Moreover, we immediately have $\hat\xi(\xi_1) = \xi(t_1)$ and $\hat\lambda(t_1) = 0$. Thus $\hat\xi$ and $\hat\lambda$ agree with ξ and λ at time t_1, and so agree by uniqueness of solutions of differential equations. ▼

Now, given $x_0 \in \mathbb{R}^n$ we will show that $x_0 \in \text{image}(\Xi(t_0'))$ for each $t_0' \in [t_0, t_1]$. Let $(\xi_*, \mu_*) \in \mathscr{P}(A, B, Q, R, x_0, t_0', t_1)$. By Lemma 2, $\xi_*(t) = \Xi(t)\xi_*(t_1)$ and the corresponding adjoint response λ_* defined in the proof of Proposition 8.2 satisfies $\lambda_*(t) = \Lambda(t)\xi_*(t_1)$ where Ξ and Λ are the solution of the final value problem (8.1). We then have

$$x_0 = \xi_*(t_0') = \Xi(t_0')\xi_*(t_1),$$

meaning that $\Xi(t_0')$ is surjective. Thus $\det \Xi(t_0') \neq 0$.

(iv) $\Longrightarrow$ (iii) We define $F(t) = \Lambda(t)\Xi(t)^{-1}$ and claim that F satisfies the Riccati differential equation with $F(t_1) = 0$. The final condition clearly holds, so we need only check that F satisfies the Riccati differential equation. Let $x' \in \mathbb{R}^n$ and let $t' \in [t_0, t_1]$. Define $x_1 \in \mathbb{R}^n$ by $\Xi(t')x_1 = x'$, this being possible since Ξ is invertible on $[t_0, t_1]$. Now let ξ and λ solve the final value problem

$$\begin{bmatrix} \dot{\xi}(t) \\ \dot{\lambda}(t) \end{bmatrix} = \begin{bmatrix} A & -S \\ -Q & -A^T \end{bmatrix} \begin{bmatrix} \xi(t) \\ \lambda(t) \end{bmatrix}, \quad \xi(t_1) = x_1,\ \lambda(t_1) = 0.$$

By Lemma 2 we have $\xi(t) = \Xi(t)\xi(t_1)$ and $\lambda(t) = \Lambda(t)\xi(t_1)$ where Ξ and Λ satisfy (8.1). Therefore, $\lambda(t) = F(t)\xi(t)$. We now have the four relations

$$\begin{aligned} \dot{\xi}(t) &= A(\xi(t)) - S(\lambda(t)), \\ \dot{\lambda}(t) &= -Q(\xi(t)) - A^T(\lambda(t)), \\ \lambda(t) &= F(t)\xi(t), \\ \dot{\lambda}(t) &= \dot{F}(t)\xi(t) + F(t)\dot{\xi}(t). \end{aligned} \tag{8.2}$$

Equating the right-hand sides of the second and fourth of these equations, and using the first and third of the equations, gives

$$(\dot{F}(t) + F(t)A + A^T F(t) - F(t)SF(t) + Q)\xi(t) = 0.$$

Evaluating at t' gives

$$(\dot{F}(t') + F(t')A + A^T F(t') - F(t')SF(t') + Q)x' = 0.$$

Since $t' \in [t_0, t_1]$ and $x' \in \mathbb{R}^n$ are arbitrary, this shows that F does indeed satisfy the Riccati differential equation. Thus the Riccati equation has a bounded solution by virtue of the fact that $\Lambda(t)\Xi(t)^{-1}$ is bounded, it being comprised from solutions to a linear differential equation with $\Xi(t)^{-1}$ being bounded. ■

8.4 Remark: (Solutions to the Riccati differential equation) Note that during the course of the proof we showed that the solution to the Riccati differential equation with final condition $F(t_1) = 0_{n\times n}$ is given by $F(t) = \Lambda(t)\Xi(t)^{-1}$, where Ξ and Λ solve the initial/final value problem from part (iv) of the theorem. Thus, while the Riccati differential equation is nonlinear, one can obtain its solution by solving a linear differential equation. •

8.5 Remark: (The rôle of controllability) It is useful to have a checkable sufficient condition to ensure that the equivalent conditions of Theorem 8.3 are met. Probably the most commonly encountered sufficient condition is that Q be positive-semidefinite and that the

system $(A, B, \mathbb{R}^m)$ be controllable, by which we mean that the Kalman controllability matrix

$$\left[\begin{array}{c|c|c|c|c} B & AB & A^2B & \cdots & A^{n-1}B \end{array}\right]$$

has full rank. This is discussed by Brockett [1970]. •

The following consequence of the theorem is the one that is most commonly emphasised in the theory.

8.6 Corollary: (Solutions to linear quadratic optimal control problems as state feedback) *Let $\Sigma = (A, B, \mathbb{R}^m)$ be a linear control system, let $Q \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ and $R \in \mathrm{L}(\mathbb{R}^m; \mathbb{R}^m)$ be symmetric with $R > 0$, and let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$. Suppose that the Riccati differential equation possesses a bounded solution $F\colon [t_0, t_1] \to \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ subject to the final condition $F(t_1) = 0_{n\times n}$. Then, for $x_0 \in \mathbb{R}^n$, the unique element $(\xi_*, \mu_*) \in \mathscr{P}(A, B, Q, R, x_0, t_0, t_1)$ satisfies the initial value problem*

$$\dot{\xi}_*(t) = (A - BR^{-1}B^TF(t))\xi_*(t), \quad \xi_*(t_0) = x_0.$$

Note that the optimal trajectories are thus simply solutions to a linear differential equation in the state since the optimal control μ_* is given as a linear function of the state: $\mu_*(t) = -R^{-1}B^TF(t)\xi_*(t)$. Thus we say that the linear quadratic optimal control problem gives a "linear state feedback" as a solution. This is somewhat remarkable.

There is much one can say about the Riccati differential equation and its relationship to linear quadratic optimal control. For example, it can be shown that if Q is positive-semidefinite, then the solutions of the Riccati differential equation exist on intervals of arbitrary length. However, we shall terminate our discussion with what we have since our aim was merely establish a connection with the Maximum Principle.

8.4. The infinite horizon problem

In this section we extend the terminal time for the linear quadratic optimal control problem to infinity. In doing so, we must make an additional assumption about our system to ensure that the limiting process is well defined. However, upon doing so we improve the character of the final result in that the resulting linear state feedback is time-independent.

The problem we study is thus the following.

8.7 Problem: (Infinite horizon linear quadratic optimal control problem) Let $\Sigma = (A, B, \mathbb{R}^m)$ be a linear control system, and let $Q \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ and $R \in \mathrm{L}(\mathbb{R}^m; \mathbb{R}^m)$ be symmetric with $R > 0$. Let $\mathscr{U}_\infty$ be the set of controls $\mu \in L^2([0, \infty); \mathbb{R}^m)$ for which all controlled trajectories (ξ, μ) satisfy $\xi \in L^2([0, \infty); \mathbb{R}^n)$. For $x_0 \in \mathbb{R}^n$, a solution to the ***infinite horizon linear quadratic optimal control problem*** from x_0 is a pair $(\xi_*, \mu_*) \in \mathrm{Ctraj}([0, \infty))$ with $\mu_* \in \mathscr{U}_\infty$ and with $\xi_*(0) = x_0$ such that, for any other pair $(\xi, \mu) \in \mathrm{Ctraj}([0, \infty))$ with $\mu \in \mathscr{U}_\infty$ and with $\xi(0) = x_0$,

$$\int_0^\infty (\tfrac{1}{2}\xi_*(t)^TQ\xi_*(t) + \tfrac{1}{2}\mu_*(t)^TR\mu_*(t))\,\mathrm{d}t \leq \int_0^\infty (\tfrac{1}{2}\xi(t)^TQ\xi(t) + \tfrac{1}{2}\mu(t)^TR\mu(t))\,\mathrm{d}t.$$

We denote by $\mathscr{P}_\infty(A, B, Q, R, x_0)$ the set of solutions to the infinite horizon linear quadratic optimal control problem from x_0. •

We now wish to state the analogue to Theorem 8.3 for this infinite horizon problem. To do so we need what will turn out to be the analogue of the Riccati differential equation. This turns out to be, interestingly, an algebraic equation called the ***algebraic Riccati equation***:

$$A^TF + FA - FSF + Q = 0,$$

where, as usual, $S = BR^{-1}B^T$. The main theorem is then the following. We omit the proof since it would take us a little far afield from the Maximum Principle. The reader can refer instead to [Dullerud and Paganini 1999].

8.8 Theorem: (Characterisation of solutions of the infinite horizon linear quadratic optimal control problem) *Let $\Sigma = (A, B, \mathbb{R}^m)$ be a linear control system and let $Q \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ and $R \in \mathrm{L}(\mathbb{R}^m; \mathbb{R}^m)$ be symmetric with $R > 0$. Then the following statements are equivalent:*

(i) for each $x_0 \in \mathbb{R}^n$, $\mathscr{P}_\infty(A, B, Q, R, x_0) \neq \emptyset$;

(ii) for each $x_0 \in \mathbb{R}^n$, $\mathscr{P}_\infty(A, B, Q, R, x_0)$ is a singleton;

(iii) there exists a solution F to the algebraic Riccati equation such that the matrix

$$A - BR^{-1}B^TF$$

is Hurwitz;

(iv) the matrix

$$\begin{bmatrix} A & -S \\ Q & -A^T \end{bmatrix}$$

has no eigenvalues on the imaginary axis.

8.9 Remark: (The rôle of controllability and stabilisability) It is clear that the linear system $(A, B, \mathbb{R}^m)$ must be stabilisable in order for any one of the equivalent conditions of the theorem to be satisfied. This condition, however, is not sufficient. A sufficient condition, one that is often met in practice, is that Q be positive-semidefinite and that $(A, B, \mathbb{R}^m)$ be controllable, by which it is meant that the Kalman controllability matrix

$$\left[\begin{array}{c|c|c|c|c} B & AB & A^2B & \cdots & A^{n-1}B \end{array}\right]$$

has maximal rank. As for the discussion in the finite horizon case, we refer to [Brockett 1970] for details of this sort. •

The theorem has the following corollary which, as with the corresponding corollary to Theorem 8.3, is often the point of most interest.

8.10 Corollary: (Solution to infinite horizon linear quadratic optimal control problems as state feedback) *Let $\Sigma = (A, B, \mathbb{R}^m)$ be a linear control system and let $Q \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ and $R \in \mathrm{L}(\mathbb{R}^m; \mathbb{R}^m)$ be symmetric with $R > 0$. Suppose that the algebraic Riccati equation possesses a solution F such that $A - BR^{-1}B^TF$ is Hurwitz. Then, for $x_0 \in \mathbb{R}^n$, the unique element $(\xi_*, \mu_*) \in \mathscr{P}_\infty(A, B, Q, R, x_0)$ satisfies the initial value problem*

$$\dot{\xi}_*(t) = (A - BR^{-1}B^TF)\xi_*(t), \quad \xi_*(0) = x_0.$$

The point is exactly that seen in the finite horizon case: the optimal trajectories are solutions of a linear differential equation. Now we even have the additional feature that the differential equation is time-independent. Note that the linear feedback $\mu_*(t) = -R^{-1}B^TF\xi(t)$ makes the system a stable linear system, even when A itself is not Hurwitz. Thus the optimal control problem leads to a stabilising linear state feedback. Moreover, this linear feedback can be shown to have many remarkable properties. We do not touch on this, but refer the reader to, for example, [Zhou 1996].

8.5. Linear quadratic optimal control as a case study of abnormality

In this section we show that the linear quadratic optimal control problem of steering between two specified points possesses no abnormal extremals. We do this for two reasons: (1) the discussion reveals something interesting about situations where possibly abnormal, but not abnormal, extremals arise; (2) the details of this are often not carried out carefully in the literature.

We consider a linear control system $\Sigma = (A, B, \mathbb{R}^m)$, we let $x_0, x_1 \in \mathbb{R}^n$, we let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and we let $Q \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ and $R \in \mathrm{L}(\mathbb{R}^m; \mathbb{R}^m)$ be symmetric and satisfy $R > 0$. We let L be the usual Lagrangian defined by Q and R. The optimal control problem we consider is that with solutions $\mathscr{P}(\Sigma, L, \{x_0\}, \{x_1\}, [t_0, t_1])$ which steer from x_0 at time t_0 to x_1 at time t_1 with minimum cost. To study the rôle of abnormality in this problem the Kalman controllability matrix

$$C(A, B) = \left[\begin{array}{c|c|c|c|c} B & AB & A^2B & \cdots & A^{n-1}B \end{array} \right].$$

plays a key part. As we discussed in Section 7.3, the system Σ is controllable (meaning one state can be steered to another) if and only if $C(A, B)$ has maximal rank.

The first result we prove shows that problems where $C(A, B)$ does not have maximal rank are, in some sense, degenerate.

8.11 Proposition: (All trajectories for uncontrollable systems are possibly abnormal extremals) *Let $\Sigma = (A, B, \mathbb{R}^m)$ be a linear control system and let (ξ, μ) be a controlled trajectory for Σ defined on $I \subset \mathbb{R}$. Let $t_0 \in I$ and let $\lambda_0 \in \mathbb{R}^n$ be orthogonal to the columnspace of the matrix $C(A, B)$. Then the adjoint response $\lambda\colon I \to \mathbb{R}^n$ for Σ along (ξ, μ) which satisfies $\lambda(t_0) = \lambda_0$ has the property that $\lambda(t)$ is orthogonal to the columnspace of the matrix $C(A, B)$ for every $t \in I$.*

In particular, if $C(A, B)$ is not of maximal rank and if $(\xi, \mu) \in \mathrm{Carc}(\Sigma, [t_0, t_1])$, then (ξ, μ) is a possibly abnormal extremal for $\mathscr{P}(\Sigma, L, \{x_0\}, \{x_1\}, [t_0, t_1])$.

Proof: We recall from basic linear systems theory (see, for example, [Brockett 1970]) that if we make an orthogonal change of basis to $\{v_1, \ldots, v_k, v_{k+1}, \ldots, v_n\}$ so that the first k basis vectors form a basis for the column space of the matrix $C(A, B)$, then the system in the partitioned basis has the form

$$\begin{bmatrix} \dot{\xi}_1(t) \\ \dot{\xi}_2(t) \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ 0_{(n-k)\times k} & A_{22} \end{bmatrix} \begin{bmatrix} \xi_1(t) \\ \xi_2(t) \end{bmatrix} + \begin{bmatrix} B_1 \\ 0_{(n-k)\times m} \end{bmatrix} \mu,$$

for appropriate matrices A_{11}, A_{12}, A_{22}, and B_1. The corresponding equation for an adjoint response for Σ along (ξ, μ) is then

$$\begin{bmatrix}\dot\lambda_1(t)\\ \dot\lambda_2(t)\end{bmatrix} = -\begin{bmatrix}A_{11}^T & 0_{k\times(n-k)}\\ A_{12}^T & A_{22}^T\end{bmatrix}\begin{bmatrix}\lambda_1(t)\\ \lambda_2(t)\end{bmatrix}.$$

The first assertion of the proposition follows from the observation that the subspace orthogonal to the columnspace of the matrix $C(A, B)$ is spanned by $\{v_{k+1}, \dots, v_n\}$, and the fact that this subspace is invariant under the adjoint equation.

The final assertion of the proposition follows since an adjoint response for Σ along (ξ, μ) is an adjoint response for $(\Sigma, \lambda^0 L)$ along (ξ, μ) with $\lambda^0 = 0$. ■

The result states, rather alarmingly, that *all* controlled trajectories satisfy the conclusions of the Maximum Principle in the case when $C(A, B)$ does not have full rank. Therefore, taken at face value, the Maximum Principle possibly tells us *nothing* about solutions of the problem. However, the next result indicates that it is possible to still get useful information from the Maximum Principle.

8.12 Proposition: (Reduction of linear quadratic optimal control problems to the normal case) *Let $\Sigma = (A, B, \mathbb{R}^m)$ be a linear control system, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, let $Q \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^n)$ and $R \in \mathrm{L}(\mathbb{R}^m; \mathbb{R}^m)$ be symmetric with $R > 0$, let L be the quadratic Lagrangian defined by Q and R, and let $S_0, S_1 \subset \mathbb{R}^n$ be smooth constraint sets. If $(\xi_*, \mu_*) \in \mathscr{P}(\Sigma, L, S_0, S_1, [t_0, t_1])$ then there exists an nonzero adjoint response $\lambda_* \colon [t_0, t_1] \to \mathbb{R}^n$ for $(\Sigma, -L)$ such that*

$$H_{\Sigma,\lambda_*^0 L}(\xi_*(t), \lambda_*(t), \mu_*(t)) = H^{\max}_{\Sigma,\lambda_*^0 L}(\xi_*(t), \lambda_*(t))$$

for almost every $t \in [t_0, t_1]$.

Proof: We suppose that we make an orthogonal change of basis to $\{v_1, \dots, v_k, v_{k+1}, \dots, v_n\}$ where the first k basis vectors form a basis for the columnspace of the matrix $C(A, B)$. The corresponding adjoint equation for the extended system is then

$$\begin{bmatrix}\dot\lambda^0(t)\\ \dot\lambda_1(t)\\ \dot\lambda_2(t)\end{bmatrix} = \begin{bmatrix}0 & 0 & 0\\ -Q_{11}\xi_1(t) - Q_{12}\xi_2(t) & -A_{11}^T & 0_{k\times(n-k)}\\ -Q_{21}\xi_1(t) - Q_{22}\xi_2(t) & -A_{12}^T & -A_{22}^T\end{bmatrix}\begin{bmatrix}\lambda^0(t)\\ \lambda_1(t)\\ \lambda_2(t)\end{bmatrix}.$$

Note that if the initial equation for this system at some time $\tau \in [t_0, t_1]$ satisfies $\lambda^0(\tau) = 0$ and $\lambda_1(\tau) = 0$ then it follows that $\lambda^0(t) = 0$ and $\lambda_1(t) = 0$ for all $t \in [t_0, t_1]$. This is because the subspace defined by $p^0 = 0$ and $p_1 = 0$ is invariant under the matrix for the adjoint response. For this reason, if $\lambda^0 \in \{0, -1\}$ and if $t \mapsto \lambda(t)$ is an adjoint response for $(\Sigma, \lambda^0 L)$, then $t \mapsto \lambda(t) + \tilde\lambda(t)$ is also an adjoint response for $(\Sigma, \lambda^0 L)$ where $\tilde\lambda$ is any adjoint response for Σ satisfying $\tilde\lambda_1(t_0) = 0$.

Now suppose that λ_*^0 and $\tilde\lambda_*$ satisfy the conclusions of the Maximum Principle for the optimal controlled trajectory (ξ_*, μ_*). Let us write the decomposition of $\tilde\lambda_*$ in the basis $\{v_1, \dots, v_n\}$ as $\tilde\lambda_* = \tilde\lambda_{*1} + \tilde\lambda_{*2}$. Now define $\lambda_* = \tilde\lambda_{*1}$ so that $\lambda_*(t)$ lies in the columnspace of $C(A, B)$ for each $t \in [t_0, t_1]$. Our computations above ensure that λ_* is an adjoint response for $(\Sigma, \lambda_*^0 L)$ along (ξ_*, μ_*).

Let us show that the Hamiltonian maximisation condition is satisfied for λ_*^0 and λ_*. If we write the Hamiltonian using the decomposition of the coordinates corresponding to the basis $\{v_1, \dots, v_n\}$ we have

$$H_{\Sigma,\lambda_*^0 L}((x_1, x_2), (p_1, p_2), u) = \langle p_1, A_{11}x_1 + A_{12}x_2 + B_1 u\rangle + \langle p_2, A_{22}x_2\rangle + \lambda_*^0(\tfrac{1}{2}x^T Q x + \tfrac{1}{2}u^T R u).$$

The control which minimises this is

$$u = -R^{-1}B^T p = -R^{-1}B_1^T p_1.$$

Since this expression is independent of p_2, the Hamiltonian maximisation condition holds for λ_*^0 and $\tilde{\lambda}_*$ if and only if it holds for λ_*^0 and λ_*.

Now let us show that if (ξ_*, μ_*) is optimal, then $\lambda_*^0 = -1$. Suppose otherwise. Then the adjoint response λ_* defined above must satisfy the adjoint differential equation for the extended system with $\lambda_*^0 = 0$:

$$\dot{\lambda}(t) = -A^T(\lambda(t)).$$

Moreover, the Hamiltonian must satisfy the Hamiltonian maximisation condition with $\lambda_*^0 = 0$. One readily sees that this is equivalent to the condition that

$$\langle \lambda_*(t), Bu\rangle = 0, \qquad u \in \mathbb{R}^m, \tag{8.3}$$

for almost every $t \in [t_0, t_1]$. Thus $\lambda_*(t)$ must lie in the subspace orthogonal to image(B) for almost every $t \in [t_0, t_1]$. But this implies, since λ_* is absolutely continuous (indeed, it is analytic), that $\lambda_*(t)$ is orthogonal to image(B) for *all* $t \in [t_0, t_1]$. One may now differentiate the relation (8.3) with respect to t:

$$\langle \dot{\lambda}(t), Bu\rangle = -\langle A^T\lambda(t), Bu\rangle, \qquad u \in \mathbb{R}^m.$$

Differentiating thusly $n-1$ times gives

$$\langle (A^T)^j\lambda(t), Bu\rangle = \langle \lambda(t), A^j Bu\rangle = u, \qquad j \in \{0, 1, \dots, n-1\},\ u \in \mathbb{R}^m.$$

This shows that if $\lambda_*^0 = 0$ then the adjoint response λ_* is orthogonal to the columnspace of the matrix $C(A, B)$ on all of $[t_0, t_1]$. Combining this with the fact that $\lambda_*(t)$ lies in the column space of the matrix $C(A, B)$ for all $t \in [t_0, t_1]$, we arrive at the conclusion that $\lambda_*(t) = 0$ for all $t \in [t_0, t_1]$. But this is in contradiction with the Maximum Principle. ■

Let us make a few observations and some fairly vague inferences based on these observations that we do not bother to prove. We leave for the motivated reader the task of making the inferences less vague, and to understanding where they come from.

1. In Proposition 8.11 we show that for uncontrollable systems *all* controlled trajectories are possibly abnormal extremals. This is entirely related to the fact that, for uncontrollable systems, the fixed interval tangent cone has an empty interior (cf. the discussion in Section 7.3). This points out the care that must be taken in interpreting the Maximum Principle in these cases. We have also discussed this matter following Lemma 5.10 and Theorem 5.16, and give an instance of this in terms of the character of the reachable set in Exercise E5.2. The problem is that, when the fixed interval tangent cone has no interior, it is possible to choose a support hyperplane for the cone that contains the cone. If the reader thinks for a moment about the proof of Proposition 8.11, they will see that this is exactly what is happening.

2. There are two reasons why the fixed interval tangent cone might have an empty interior.
   (a) One reason is that the state space for the system is "too big." This is essentially what is happening in Proposition 8.11 when $C(A, B)$ does not have full rank. Since the system is not controllable, this indicates that there are some restrictions of the character of the controlled trajectories. It turns out that these restrictions manifest themselves by the fixed interval tangent cones possessing support hyperplanes that contain the cones.
   (b) Even if the state space is not "too big," the interior of the fixed interval tangent cone may be empty if the system is not linearly controllable along the reference trajectory. In such cases one must go to higher-order Maximum Principles to get refined necessary conditions for optimality. This is very much related, then, to the connection between optimal control theory and controllability alluded to in Remark 7.7.
3. In cases when possibly abnormal extremals arise from the state space being "too big," it is often possible to infer from the Maximum Principle the information required to get useful necessary conditions. This is seen, for example, in Proposition 8.12. Even though Proposition 8.11 says that all trajectories are possibly abnormal extremals, in the proof of Proposition 8.12 we show that one can still prove that all optimal trajectories are, in fact, normal. Note, however, that this requires some work, and that in more complicated situations this work might be hard.

Exercises

E8.1 Let $\Sigma = (A, B, \mathbb{R}^m)$ be a linear control system and consider the quadratic optimal control problem defined using symmetric matrices Q and R, but with R not being positive-semidefinite. Show that there exists $x_0 \in \mathbb{R}^n$ such that $\mathscr{P}(A, B, Q, R, x_0, t_0, t_1) = \emptyset$. (Remember Assumption 8.1.)

E8.2 Consider the linear quadratic optimal control problem with

1. $n = 2$ and $m = 1$,
2. $A = \left[\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right]$ and $B = \left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]$, and
3. $Q = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ and $R = [1]$.

Let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$. Answer the following questions.

(a) Solve the associated Riccati differential equation on $[t_0, t_1]$ subject to the final condition $F(t_1) = 0_{2\times 2}$.

(b) If $t \mapsto F(t)$ is the solution to the Riccati differential equation from part (a), show that the limit $F_\infty = \lim_{t\to-\infty} F(t)$ exists and is independent of the final condition t_1.

(c) Verify that the matrix $A - BR^{-1}B^T F_\infty$ is Hurwitz.

Chapter 9

Linear time-optimal control

The problem of time-optimal control is one of great importance in control theory. It can be of importance in applications, although it is very often the case that time-optimal trajectories are not the ones you really want, even though you think they might be. The reason for this is that time-optimal control can be quite "violent." Often perfectly satisfactory time-suboptimal control laws are what is best. However, from the point of view of structural properties of control systems, time-optimal control is nice because it does not bring along much baggage from its cost function. That is to say, the structure of time-optimal extremals says a lot about the system itself, whereas other optimal problems often say just as much about the particular Lagrangian as anything else. This is as it should be, of course. But the point is that, from the point of view of fundamental control theory, time-optimal control is useful to study. In this chapter we focus on linear systems, since in this case one can actually say something interesting about the time-optimal extremals.

9.1. Some general comments about time-optimal control

For time-optimal control, the Lagrangian is obviously defined by $L(x, u) = 1$. For a control system $\Sigma = (\mathcal{X}, f, U)$ and for $S_0, S_1 \subset \mathcal{X}$, let $\mathscr{P}_{\text{time}}(\Sigma, S_0, S_1)$ denote the set of solutions for the time-optimal problem.

The following result is one reason why the time-optimal control problem is of such interest.

9.1 Proposition: (Time-optimal trajectories lie on the boundary of the reachable set) *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system and let $S_0, S_1 \subset \mathcal{X}$. If $(\xi_*, \mu_*) \in \mathscr{P}_{\text{time}}(\Sigma, S_0, S_1)$ is defined on $[t_0, t_1]$, then $\xi_*(t_1) \in \text{bd}(\mathcal{R}(\xi_*(t_0), t_0, t_1))$.*

Proof: Note that

$$\hat{\mathcal{R}}(\hat{\xi}_*(t_0), t_0, t) = \{t\} \times \mathcal{R}(\xi_*(t_0), t_0, t), \qquad t \in [t_0, t_1].$$

If $\xi_*(t_1) \in \text{int}(\mathcal{R}(\xi_*(t_0), t_0, t_1))$ then $(\xi_*(t_1), t_1) \in \text{int}(\hat{\mathcal{R}}(\hat{\xi}_*(t_0), t_0))$. Now, since every neighbourhood of $(\xi_*(t_1), t_1) \in \hat{\mathcal{X}}$ contains a point in

$$\{(\xi_*(t_1), t) \in \hat{\mathcal{X}} \mid t < t_1\},$$

this means that there are points in $\mathcal{R}(\xi_*(t_0), t_0)$ that are also in this set. But this contradicts Lemma 6.2. ■

One of the consequences of the preceding result is that if the reachable sets do not have boundaries, then the time-optimal control problem will generally not have a solution. It is very often the case for systems with unbounded controls that the reachable sets do not, in fact, have boundaries. For this reason, one typically needs to constrain the control to lie in a compact set to ensure solutions to time-optimal control problems. This is not a theorem.

A useful property of extremals for time-optimal problems is the following.

9.2 Proposition: *Let $\Sigma = (\mathcal{X}, f, U)$ be a control system and let $S_0, S_1 \subset \mathcal{X}$. If $(\xi_*, \mu_*) \in \mathscr{P}_{\text{time}}(\Sigma, S_0, S_1)$ is defined on $[t_0, t_1]$ and if $\lambda_* \colon [t_0, t_1] \to \mathbb{R}^n$ is the adjoint response guaranteed by Theorem 3.4, then λ_* is nowhere zero.*

Proof: We consider two cases. First, when $\lambda_*^0 = 0$ it immediately follows from Theorem 3.4 that $\lambda_*(t_0) \neq 0$. Since the adjoint equation is linear it follows that $\lambda_*(t) \neq 0$ for all $t \in [t_0, t_1]$. Second, when $\lambda_*^0 = -1$ then the Hamiltonian along the optimal trajectory is specified by

$$t \mapsto \langle \lambda_*(t), f(\xi_*(t), \mu_*(t)) \rangle - 1.$$

If $\lambda_*(t) = 0$ for some $t \in [t_0, t_1]$ then λ_* is identically zero, since the component of adjoint equation for λ_* is linear. The Hamiltonian must, therefore, be identically -1 along the trajectory, and this contradicts Theorem 3.4. ■

9.2. The Maximum Principle for linear time-optimal control

In this section we consider a linear control system $\Sigma = (A, B, U)$ where U is a convex polytope in $\mathbb{R}^m$. In practice one often takes

$$U = [a_1, b_1] \times \cdots \times [a_m, b_m]$$

for some $a_1, b_1, \ldots, a_m, b_m \in \mathbb{R}^m$. This reflects the fact that in practice one typically has actuator constraints on each control. Most often these constraints are also symmetric about zero. However, the geometry of the situation is best revealed if one uses a general convex polytope instead. For simplicity let us take $t_0 = 0$, $S_0 = \{x_0\}$, and $S_1 = \{x_1\}$.

The adjoint equations for the system are

$$\dot{\xi}(t) = A(\xi(t)) + B(\mu(t)), \quad \dot{\lambda}(t) = -A^T(\lambda(t)).$$

Note that the equation for the adjoint response decouples so that we simply have

$$\lambda(t) = \exp(-A^T t)\lambda(0). \tag{9.1}$$

The Hamiltonian is

$$H_{\Sigma,\lambda^0 L}(x, p, u) = \langle p, Ax + Bu \rangle + \lambda^0.$$

The maximisation of the Hamiltonian as a function of u is equivalent to the maximisation of the function $u \mapsto \langle p, Bu \rangle$. For fixed p this is a linear function of $u \in U$. The Maximum Principle tells us that this function must be maximised almost everywhere along an optimal trajectory. From Theorem B.25 we know that either

1. $u \mapsto \langle p, Bu \rangle$ is constant or

2. the maximum is achieved on the boundary of the convex polytope U.

In any case, the situation is completely determined by the fact that the adjoint response is given by (9.1). We would like to use this fact to say as much as possible about the character of the time-optimal extremals. The following result is an example of the sort of thing one can say.

9.3 Proposition: (Property of time-optimal extremals) *Let $\Sigma = (A, B, U)$ be a linear control system with $U \subset \mathbb{R}^m$ a convex polytope. Suppose that, if $w \in \mathbb{R}^m \setminus \{0\}$ is collinear with a rib of U, then the matrix*

$$\left[\begin{array}{c|c|c|c} Bw & ABw & \cdots & A^{n-1}Bw \end{array} \right] \tag{9.2}$$

has full rank, i.e., the vector Bw is contained in no proper A-invariant subspace of $\mathbb{R}^n$. If $\lambda\colon [0,T] \to \mathbb{R}^n$ satisfies (9.1) with $\lambda(0) \neq 0$, then there exists a unique $\mu\colon [0,T] \to U$ with the following properties:

(i) $\langle \lambda(t), B\mu(t) \rangle = \max\{\langle \lambda(t), Bu \rangle \mid u \in U\}$;

(ii) μ is continuous at 0 and T and continuous from the left for all $t \in (0,T)$.

Moreover, the control μ is piecewise constant.

Proof: From Theorem B.25 we know that the function

$$u \mapsto \langle \lambda(t), Bu \rangle$$

is maximised either at a vertex of U or on a face of dimension $j \geq 1$ of U (we consider U itself to be a face of dimension n). We will first show that the latter circumstance can arise only at a finite number of points in $[0,T]$. Assume otherwise. Then, since there are finitely many faces of all dimensions, there must exist some face F of dimension $j \geq 1$ and an infinite set $S \subset [0,T]$ such that, for each $t \in S$, the function $u \mapsto \langle \lambda(t), Bu \rangle$ is constant on F. Choose two distinct adjacent vertices v_1 and v_2 for U lying in F, supposing that the vector $w = v_2 - v_1$ is collinear with the rib connecting v_1 and v_2. By our assumption that $u \mapsto \langle \lambda(t), Bu \rangle$ is constant on F, we have

$$\langle \lambda(t), Bw \rangle = \langle \lambda(t), Bv_1 \rangle - \langle \lambda(t), Bv_2 \rangle = 0, \qquad t \in S.$$

Since $[0,T]$ is compact and since S is infinite, there exists a convergent sequence $\{t_j\}_{j \in \mathbb{Z}_{>0}} \subset S$. Thus $\langle \lambda(t_j), Bw \rangle = 0$ for $j \in \mathbb{Z}_{>0}$. Since the function $t \mapsto \langle \lambda(t), Bw \rangle$ is analytic, we can deduce that this function is identically zero, it being zero on the convergent subsequence $\{t_j\}_{j \in \mathbb{Z}_{>0}}$. Now differentiate the equality $\langle \lambda(t), Bw \rangle = 0$ successively $n-1$ times to get

$$\langle (A^T)^j \lambda(t), Bw \rangle = \langle \lambda(t), A^j Bw \rangle = 0, \qquad j \in \{0, 1, \ldots, n-1\},\ t \in [0,T].$$

Since we assumed that $\lambda(0) \neq 0$, this contradicts (9.2). Thus we conclude that the function $u \mapsto \langle \lambda(t), Bu \rangle$ is constant on a face of positive dimension only for finitely many times $t \in [0,T]$. That is, for all but finitely many times $t \in [0,T]$ the control $\mu(t)$ takes values in a unique vertex of U. Uniqueness of the control on $[0,T]$ follows by imposing condition (ii).

Now let us show that μ is piecewise constant. To do this we partition $[0,T]$ into a finite number of disjoint intervals such that, for t lying in the interior of any of these intervals,

$\mu(t)$ takes values in a unique vertex. Let I be one of these intervals, let $v_1, \dots, v_N$ be the vertices of U, and let S_j, $j \in \{1, \dots, N\}$, be the set of points in I for which μ takes values in v_j. We then have

$$S_{j_1} \cap S_{j_2} = \emptyset, \ j_1, j_2 \in \{1, \dots, N\} \text{ distinct}, \qquad I = \cup_{j=1}^{N} S_j.$$

We claim that S_j is open for each $j \in \{1, \dots, N\}$. The only pertinent case is when S_j is nonempty, so suppose this is the case. Let $t' \in S_j$. Then

$$\langle \lambda(t'), Bv_l \rangle < \langle \lambda(t'), Bv_j \rangle, \quad l \neq j.$$

By continuity of $t \mapsto \langle \lambda(t), Bw \rangle$ for $w \in \mathbb{R}^m$, there exists a neighbourhood $\mathcal{I}$ of t' such that

$$\langle \lambda(t), Bv_l \rangle < \langle \lambda(t), Bv_j \rangle, \quad l \neq j, \ t \in \mathcal{I}.$$

This shows that S_j is open.

Finally, since I is a disjoint union of the open sets $S_1, \dots, S_N$ and since I is connected, we must have $I = S_j$ for some fixed $j \in \{1, \dots, N\}$. This gives μ as piecewise constant as claimed. ■

9.4 Remarks: 1. One might ask whether the condition (9.2) can be expected to be satisfied. First of all, note that this condition implies the controllability of Σ. If a system is not controllable, one can restrict attention to the controllable subspace (i.e., the columnspace of the Kalman controllability matrix), so one loses nothing by assuming controllability. Moreover, if a system *is* controllable, then the set of vectors $w \in \mathbb{R}^m$ for which the matrix in (9.2) does *not* have maximal rank is small. To be somewhat more precise, the set of such vectors lies in the intersection of the set of zeros of a finite number of polynomial equations. These things are discussed by Wonham [1985]. For us, the point is that a generic convex polytope will satisfy the hypotheses of Proposition 9.3.

2. The second condition in the statement of the result is merely a technical condition that ensures uniqueness. The fact of the matter is that at points of discontinuity, the exact value of the control is of no consequence as far as the trajectory is concerned. •

9.3. An example

We take the system $\Sigma = (M, f, U)$ where

1. $M = \mathbb{R}^2$,
2. $f((x^1, x^2), u) = (x^2, -x^1 + u)$,
3. $U = [-1, 1] \subset \mathbb{R}$.

The cost function we choose is that associated with time-optimisation; thus we take $L(x, u) = 1$. We consider the problem $\mathscr{P}_{\text{time}}(\Sigma, \{(x_0^1, x_0^2)\}, \{x_1^1, x_1^2\})$. The Hamiltonian for this system is

$$H_{\Sigma, \lambda^0 L}(x, p, u) = p^1 x^2 - p^2 x^1 + p^2 u + \lambda^0.$$

This gives the equations governing extremals as

$$\dot{x}^1 = x^2, \quad \dot{x}^2 = -x^1 + u, \quad \dot{p}^1 = p^2, \quad \dot{p}^2 = -p^1.$$

We may solve the equations for the adjoint variables p^1 and p^2 directly:

$$p^1(t) = A\sin(t-\phi), \quad p^2 = A\cos(t-\phi) \tag{9.3}$$

for some $A, \phi \in \mathbb{R}$.

The control $u(t)$ for an extremal satisfies

$$p^2(t)u(t) = \max\{p^2(t)\tilde{u} \mid \tilde{u} \in U\},$$

meaning that when $p^2(t) < 0$ we have $u(t) = -1$, and when $p^2(t) > 0$ we have $u(t) = +1$. Thus $u(t)$ alternates between $+1$ and -1, depending on the sign of $p^2(t)$. However, given the form of $p^2(t)$, this means that $u(t)$ switches every π seconds.

This shows that extremals will be concatenations of solutions of the two differential equations

1. $\dot{x}^1 = x^2$, $\dot{x}^2 = -x^1 + 1$ and
2. $\dot{x}^1 = x^2$, $\dot{x}^2 = -x^1 - 1$.

The solutions to the first equation are

$$x^1(t) = B_1\sin(t-\psi_1) + 1, \quad x^2(t) = B_1\cos(t-\psi_1) \tag{9.4}$$

for constants $B_1, \psi_1 \in \mathbb{R}$. These are simply circles in the (x^1, x^2)-plane centred at $(1, 0)$. In like manner, the solutions for the other class of optimal arcs are determined by

$$x^1(t) = B_2\sin(t-\psi_2) - 1, \quad x^2(t) = B_2\cos(t-\psi_2) \tag{9.5}$$

for constants $B_2, \psi_2 \in \mathbb{R}$. These are simply circles in the (x^1, x^2)-plane centred at $(-1, 0)$. Thus, to steer from (x_0^1, x_0^2) to (x_1^1, x_1^2) in a time-optimal manner, one would go from (x_0^1, x_0^2) to (x_1^1, x_1^2) along a curve consisting of a concatenation of circles centred at $(1, 0)$ and at $(-1, 0)$ (see Figure 9.1).

Next we look at the abnormal extremals. In this case constancy (in fact, equality with zero) of the Hamiltonian as guaranteed by the Maximum Principle tells us we must have

$$H_{\Sigma,0}(u, x, p) = p^1x^2 - p^2x^1 + p^2u = 0.$$

A straightforward calculation, using (9.3), (9.4), and (9.5), and the fact that $u(t) = \operatorname{sign}(p^2(t))$, gives

$$p^1x^2 - p^2x^1 + p^2u = AB\sin(\psi - \phi).$$

Thus an extremal is possibly abnormal if and only if $\psi - \phi = n\pi$, $n \in \mathbb{Z}$. Note that to verify abnormality one must also verify that there are no extremals with $\lambda^0 = -1$ that give the same extremal trajectory.

For this problem, there exist time-optimal trajectories that are abnormal extremal trajectories. For example, suppose that one wishes to go from $(x_0^1, x_0^2) = (0, 0)$ to $(x_1^1, x_1^2) = (2, 0)$. In this case the time-optimal control is given by $u(t) = 1$ that is applied for π seconds. The corresponding trajectory in state space is

$$x^1(t) = -\cos t + 1, \quad x^2(t) = \sin t.$$

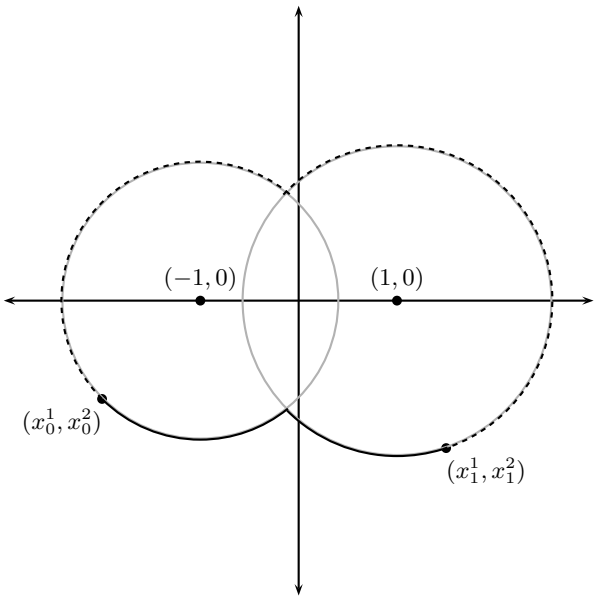

Figure 9.1. Two concatenations of circles to form an extremal trajectory. The solid line is the solution to optimal control problem, and the dashed line is another extremal trajectory.

That this is the time-optimal trajectory is intuitively clear: one pushes as hard as one can in the direction one wants to go until one gets there.[1] However, this extremal is abnormal. Let's see how this works. Since the controlled trajectory (ξ, u) just described is minimising, it must satisfy the conditions of the Maximum Principle. In particular, the maximisation condition on the Hamiltonian must obtain. This means that $p^2(t)$ must be positive for $0 \leq t \leq \pi$, except possibly at the endpoints. If $p^2(t)$ changes sign in the interval $[0, \pi]$, then u must also change sign, but this cannot happen since $u(t) = 1$. This implies that $p^2(t) = A \sin t$, and so this immediately gives $p^1(t) = -A \cos t$. We see then that we may take $\phi = \frac{\pi}{2}$ and $\psi = \frac{\pi}{2}$. Given our characterisation of abnormal extremals, this shows that the time-optimal control we have found is only realisable as an abnormal extremal.

Let's see if we can provide a geometric interpretation of what is going on here. In Figure 9.2 we show a collection of concatenated extremals that emanate from the origin. From this picture it is believable that the set of points reachable from $(0, 0)$ in time π is precisely the circle of radius 2 in the (x^1, x^2)-plane. Why are the points $(\pm 2, 0)$ distinguished? (We have only looked at the point $(2, 0)$, but the same arguments hold for $(-2, 0)$.) Well, look at how the extremal curves touch the boundary of the reachable set. Only at $(\pm 2, 0)$ do the extremal curves approach the boundary so that they are tangent to the supporting hyperplane at the boundary.

[1] It is also easy to make this rigorous. For if $u(t) < 1$ for t in some set A of positive measure, it is fairly evident that $x^1(\pi) < 2$. Thus any such trajectory will remain in the half-plane $x^1 < 2$ for the first π seconds, and so cannot be time-optimal for the problem of steering from $(0, 0)$ to $(2, 0)$.

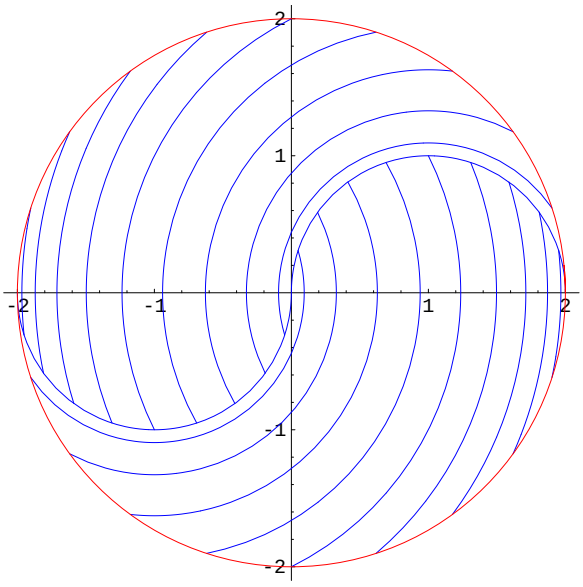

Figure 9.2. The set of points reachable from $(0,0)$ in time π

Exercises

E9.1 Consider a linear control system $\Sigma = (A, B, U)$ with $U = \mathbb{R}^m$. For any $x_0, x_1 \in \mathbb{R}^n$, show that the time-optimal control problem of steering from x_0 to x_1 in minimum time has no solution.

Hint: *Define*

$$\mathrm{P(A,B)(t)} = \int_0^{\mathrm{t}} \mathrm{e}^{\mathrm{A}\tau}\mathrm{B}\mathrm{B}^{\mathrm{T}}\mathrm{e}^{\mathrm{A}^{\mathrm{T}}\tau}\,\mathrm{d}\tau$$

and show that the control

$$\mu(\mathrm{t}) = -\mathrm{B}^{\mathrm{T}}\mathrm{e}^{\mathrm{A}^{\mathrm{T}}(\mathrm{T}-\mathrm{t})}\mathrm{P(A,B)(T)}\big(\mathrm{e}^{\mathrm{AT}}\mathrm{x}_0 - \mathrm{x}_1\big)$$

steers the system from x_0 *to* x_1 *in time* T.

Appendices

Appendix A

Some results from the theory of ordinary differential equations with measurable time-dependence

In this appendix we provide a quick and definitely not self-contained overview of the theory of differential equations as needed in control theory. The theory here is a little more involved than the standard theory in that the time dependence we must allow is rather general; normally one considers continuous dependence on time, and we shall need to allow more generality than this. Thus we begin with some measure theory.

A.1. Concepts from measure theory

In this section we shall give a brief discussion of the ideas from measure theory that we shall use. For readers with no background in measure theory, this will be a woefully inadequate introduction; we refer to books like [Cohn 1980, Halmos 1974] for more background. Fortunately, we only need the Lebesgue measure on $\mathbb{R}$, so we can keep the discussion more focussed.

A.1.1. Lebesgue measure. The first step in defining the Lebesgue measure on $\mathbb{R}$ is to define the so-called outer measure. This can be applied to a general set, and generalises the notion of length for intervals.

A.1 Definition: The ***Lebesgue outer measure*** of a subset $S \subset \mathbb{R}$ is the element $\lambda^*(S) \in \overline{\mathbb{R}}_{\geq 0}$ defined by

$$\lambda^*(S) = \inf\Big\{ \sum_{j=1}^{\infty} |b_j - a_j| \;\Big|\; S \subset \bigcup_{j \in \mathbb{Z}_{>0}} (a_j, b_j) \Big\}. \qquad \bullet$$

It is possible to verify that the Lebesgue outer measure has the following properties:

1. $\lambda^*(\emptyset) = 0$;
2. if $S \subset T \subset \mathbb{R}$, then $\lambda^*(S) \leq \lambda^*(T)$;
3. $\lambda^*\Big(\bigcup_{n \in \mathbb{N}} S_n\Big) \leq \sum_{n=1}^{\infty} \mu^*(S_n)$ for every collection $\{S_n\}_{n \in \mathbb{N}}$ of subsets of $\mathbb{R}$;

4. if $I \subset \mathbb{R}$ is an interval, then $\lambda^*(I)$ is the length of I.

Unfortunately, the Lebesgue outer measure is not a measure (whatever that is) on the subsets of $\mathbb{R}$. To obtain a measure, we must restrict it to certain subsets of $\mathbb{R}$.

A.2 Definition: (Lebesgue measure) Denote by $\mathscr{L}(\mathbb{R})$ the collection of subsets A of $\mathbb{R}$ for which

$$\lambda^*(S) = \lambda^*(S \cap A) + \lambda^*(S \cap (\mathbb{R} \setminus A)), \qquad S \subset \mathbb{R}.$$

A set in $\mathscr{L}(\mathbb{R})$ is ***Lebesgue measurable***. The map $\lambda \colon \mathscr{L}(\mathbb{R}) \to \overline{\mathbb{R}}_{\geq 0}$ defined by $\lambda(A) = \lambda^*(A)$ is the ***Lebesgue measure***. A subset $Z \in \mathscr{L}(\mathbb{R})$ is a ***set of measure zero*** if $\lambda(Z) = 0$. •

A.3 Remarks: 1. Most subsets one dreams up are measurable. Indeed, it may be shown that any definition of a non-measurable set must rely of the Axiom of Choice. Thus such sets will not be able to be given an "explicit" characterisation.

2. We will frequently be interested in subsets, not of all of $\mathbb{R}$, but of an interval $I \subset \mathbb{R}$. One can define measurable subsets of I by $\mathscr{L}(I) = \{A \cap I \mid A \in \mathscr{L}(\mathbb{R})\}$.

3. A property P holds ***almost everywhere*** (***a.e.***) on I, or for ***almost every $t \in I$*** (***a.e. $t \in I$***) if there exists a subset $N \subset I$ of zero measure such that P holds for all $t \in I \setminus N$. •

A.1.2. Integration. One of the principal ideas in measure theory is the notion of integration using measure. For the Lebesgue measure on $\mathbb{R}$, this leads to a notion of integration that strictly generalises Riemann integration, and which has some extremely important properties not possessed by Riemann integration. We do not get into this here, but refer to the references. Instead, we simply plough ahead with the definitions.

A.4 Definition: (Measurable function) Let $I \subset \mathbb{R}$ be an interval. A function $f \colon I \to \overline{\mathbb{R}}$ is ***measurable*** if, for every $a \in \mathbb{R}$, we have $f^{-1}([a, \infty]) \in \mathscr{L}(I)$. •

Since it is not easy to find sets that are not measurable, it is also not so easy to define functions that are not measurable. Thus the class of measurable functions will include nearly any sort of function one is likely to encounter in practice.

Now we indicate how to integrate a certain sort of function. To do so, if S is a set and if $A \subset S$, then we define the ***characteristic function*** of A to be the function

$$\chi_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases}$$

We now make a definition.

A.5 Definition: (Simple function) Let $I \subset \mathbb{R}$ be an interval. A ***simple function*** is a function $f \colon I \to \mathbb{R}$ of the form

$$f = \sum_{j=1}^{k} c_j \chi_{A_j} \tag{A.1}$$

for $A_j \in \mathscr{L}(I)$ and $c_j \in \mathbb{R}$, $j \in \{1, \ldots, k\}$, $k \in \mathbb{N}$. The ***integral*** of a simple function of the form (A.1) is defined by

$$\int_I f \, \mathrm{d}\lambda = \sum_{j=1}^{k} c_j \lambda(A_j). \qquad \bullet$$

It turns out that it is possible to approximate a measurable function taking values in $\overline{\mathbb{R}}_{\geq 0}$ with an increasing sequence of simple functions. With this in mind, if $f\colon I \to \overline{\mathbb{R}}_{\geq 0}$ is measurable, then we define

$$\int_I f\,\mathrm{d}\lambda = \sup\Big\{ \int_I g\,\mathrm{d}\lambda \;\Big|\; g \text{ is a positive simple function with } g(t) \leq f(t) \text{ for } t \in I \Big\}.$$

If $f\colon I \to \overline{\mathbb{R}}$, then we define

$$f^+(x) = \max\{0, f(x)\}, \quad f^-(x) = -\min\{0, f(x)\}.$$

and then define

$$\int_I f\,\mathrm{d}\lambda = \int f^+\,\mathrm{d}\lambda - \int f^-\,\mathrm{d}\lambda.$$

This is the ***Lebesgue integral*** of f.

A.1.3. Classes of integrable functions. Let us define some language surrounding the Lebesgue integral of functions.

A.6 Definition: (Integrable, essentially bounded, absolutely continuous) Let $I \subset \mathbb{R}$ be an interval and let $f\colon I \to \overline{\mathbb{R}}$ be measurable.

(i) If at least one of $\int_I f^+\,\mathrm{d}\lambda$ or $\int_I f^-\,\mathrm{d}\lambda$ are finite, then the Lebesgue integral of f ***exists*** (and may be infinite).

(ii) If both $\int_I f^+\,\mathrm{d}\lambda$ and $\int_I f^-\,\mathrm{d}\lambda$ are infinite, then the Lebesgue integral of f ***does not exist***.

(iii) If $\int_I f^+\,\mathrm{d}\lambda < \infty$ and $\int_I f^-\,\mathrm{d}\lambda < \infty$, then f is ***Lebesgue integrable***, or simply ***integrable***.

(iv) If, for each compact subinterval $J \subset I$, the function $f|J$ is integrable, then f is ***locally integrable***.

(v) If there exists $M > 0$ such that $\lambda(\{t \in I \mid |f(t)| > M\}) = 0$, then f is ***essentially bounded***, and we write

$$\operatorname{ess\,sup}_{t\in I}|f(t)| = \inf\{M \in \mathbb{R} \mid \lambda(\{t \in I \mid |f(t)| > M\}) = 0\}.$$

(vi) If there exist a locally integrable function $g\colon I \to \overline{\mathbb{R}}$ and some $t_0 \in I$ such that

$$f(t) = \int_{[t_0,t]} (g|[t_0,t])\,\mathrm{d}\lambda,$$

then f is ***locally absolutely continuous***. If I is compact, then locally absolutely continuous will be used interchangeably with ***absolutely continuous***. •

An absolutely continuous function has a derivative almost everywhere. Indeed, the Fundamental Theorem of Calculus holds, and if

$$f(t) = \int_{t_0}^{t} g(\tau)\,\mathrm{d}\tau$$

for some locally integrable function g, then $\dot{f}(t) = g(t)$ for almost every t. Moreover, if an absolutely continuous function has an almost everywhere zero derivative, it can be shown to be constant. We will make several uses of this fact.

Related to these matters, although not in a straightforward manner, is the following notion.

A.7 Definition: (Lebesgue point) Let $I \subset \mathbb{R}$ be an interval and let $f\colon I \to \mathbb{R}$ be locally integrable. A point $t_0 \in I$ is a ***Lebesgue point*** for f if

$$\lim_{\epsilon \to 0} \frac{1}{2\epsilon} \int_{t_0-\epsilon}^{t_0+\epsilon} |f(t) - f(t_0)| \, \mathrm{d}t = 0. \qquad \bullet$$

It can be shown that the complement of the set of Lebesgue points has zero measure.

All of the above notions for $\mathbb{R}$-valued functions–measurability, integrability, absolute continuity–can be applied to $\mathbb{R}^n$-valued functions by applying the above definitions component-wise.

A.2. Ordinary differential equations with measurable time-dependence

In this section we give the standard results concerning the character of solutions to the differential equations governing control systems. These will be used throughout the text, sometimes without explicit reference.

A.8 Theorem: (Carathéodory's existence and uniqueness theorem) *Let $\mathcal{X} \subset \mathbb{R}^n$ be open, let $I \subset \mathbb{R}$ be an interval, and suppose that $f\colon I \times \mathcal{X} \to \mathbb{R}^n$ has the property that $t \mapsto f(t, x)$ is locally integrable for each $x \in \mathcal{X}$ and that $x \mapsto f(t, x)$ is of class C^1 for each $t \in I$. Let $t_0 \in I$ and let $x_0 \in \mathcal{X}$. Then there exists an interval $J \subset I$ and a locally absolutely continuous curve $\xi\colon J \to \mathcal{X}$ such that*

(i) $\mathrm{int}(J) \neq \emptyset$,

(ii) $t_0 \in J$,

(iii) $\xi(t_0) = x_0$, *and*

(iv) $\dot{\xi}(t) = f(t, \xi(t))$ *for almost every* $t \in J$.

Moreover, if $\tilde{J}$ and $\tilde{\xi}\colon \tilde{J} \to \mathcal{X}$ also have the properties (i)–(iv), then $\xi(t) = \tilde{\xi}(t)$ for each $t \in J \cap \tilde{J}$.

For a control system subject to an admissible control, the bottom line is that controlled trajectories exist and are unique on sufficiently small time intervals around the initial time. Of course, just as is the case in the standard theory of differential equations, it is possible that, for a given admissible control $t \mapsto \mu(t)$, the largest time interval on which a controlled trajectory can exist might be bounded above, below, or both above and below.

We will many times use the fact that solutions to differential equations depend continuously on initial condition and on the differential equation itself. The following result encodes this.

A.9 Theorem: (Continuous dependence on initial conditions and parameters) *Let $\mathcal{X} \subset \mathbb{R}^n$ be an open set, let $t_0, t_1 \in \mathbb{R}$ satisfy $t_0 < t_1$, and let $\delta > 0$. Suppose that $f, h\colon [t_0, t_1] \times \mathcal{X} \to \mathbb{R}^n$ satisfy*

(i) $t \mapsto f(t, x)$ *and* $t \mapsto h(t, x)$ *are integrable for each* $x \in \mathcal{X}$ *and*

(ii) $x \mapsto f(t, x)$ *and* $x \mapsto h(t, x)$ *are of class* C^1 *for each* $t \in [t_0, t_1]$.

Let $\xi\colon [t_0, t_1] \to \mathcal{X}$ be a solution of the differential equation

$$\dot{\xi}(t) = f(t, \xi(t)),$$

and suppose that

$$\{x \in \mathbb{R}^n \mid \ \|x - \xi(t)\| \leq \delta, \ t \in [t_0, t_1]\} \subset \mathcal{X}.$$

Define

$$H(t) = \int_{t_0}^{t} h(\tau, \xi(\tau))\,\mathrm{d}\tau, \quad H_{\max} = \sup\{H(t) \mid \ t \in [t_0, t_1]\}.$$

Suppose that $z_0 \in \mathcal{X}$ *and* $\alpha\colon [t_0, t_1] \to \mathbb{R}_{\geq 0}$ *are such that*

(iii) α *is integrable,*

(iv) $\max\{H_{\max}, \|\xi(t_0) - z_0\|\} \leq \frac{\delta}{2}\mathrm{e}^{-\int_{t_0}^{t_1} \alpha(t)\,\mathrm{d}t}$, *and*

(v) $\|f(t, x_1) + h(t, x_1) - f(t, x_2) - h(t, x_2)\| \leq \alpha(t)\|x_1 - x_2\|$ *for* $x_1, x_2 \in \mathcal{X}$ *and* $t \in [t_0, t_1]$.

Then the solution of the initial value problem

$$\dot{\zeta}(t) = f(t, \zeta(t)) + h(t, \zeta(t)), \quad \zeta(t_0) = z_0,$$

exists on $[t_0, t_1]$ *and satisfies*

$$\|\xi(t) - \zeta(t)\| \leq (\|\xi(t_0) - z_0\| + H_{\max})\mathrm{e}^{\int_{t_0}^{t_1} \alpha(s)\,\mathrm{d}s}$$

for all $t \in [t_0, t_1]$.

Appendix B

Convex sets, affine subspaces, and cones

An important part of the proof of the Maximum Principle is the use of cones and convex cones to approximate the reachable set. In this appendix we give the basic definitions and properties that we shall use.

B.1. Definitions

B.1 Definition: (Convex set, cone, convex cone, affine subspace)

(i) A subset $C \subset \mathbb{R}^n$ is ***convex*** if, for each $x_1, x_2 \in C$, we have

$$\{sx_1 + (1-s)x_2 \mid \ s \in [0,1]\} \subset C.$$

(ii) A subset $K \subset \mathbb{R}^n$ is a ***cone*** if, for each $x \in K$, we have

$$\{\lambda x \mid \ \lambda \in \mathbb{R}_{\geq 0}\} \subset K.$$

(iii) A subset $K \subset \mathbb{R}^n$ is a ***convex cone*** if it is both convex and a cone.

(iv) A subset $A \subset \mathbb{R}^n$ is an ***affine subspace*** if, for each $x_1, x_2 \in A$, we have

$$\{sx_1 + (1-s)x_2 \mid \ s \in \mathbb{R}\} \subset A. \qquad \bullet$$

Note that the set $\{sx_1 + (1-s)x_2 \mid \ s \in [0,1]\}$ is the line segment in $\mathbb{R}^n$ between x_1 and x_2. Thus a set is convex when the line segment connecting any two points in the set remains in the set. In a similar manner, $\{\lambda x \mid \ \lambda \in \mathbb{R}_{\geq 0}\}$ is the ray emanating from $0 \in \mathbb{R}^n$ through the point x. A set is thus a cone when the rays emanating from 0 through all points remain in the set. One usually considers cones whose rays emanate from a general point in $\mathbb{R}^n$, but we will not employ this degree of generality. An affine subspace is a set where the (bi-infinite) line through any two points in the set remains in the set. We illustrate some of the intuition concerning these various sorts of sets in Figure B.1.

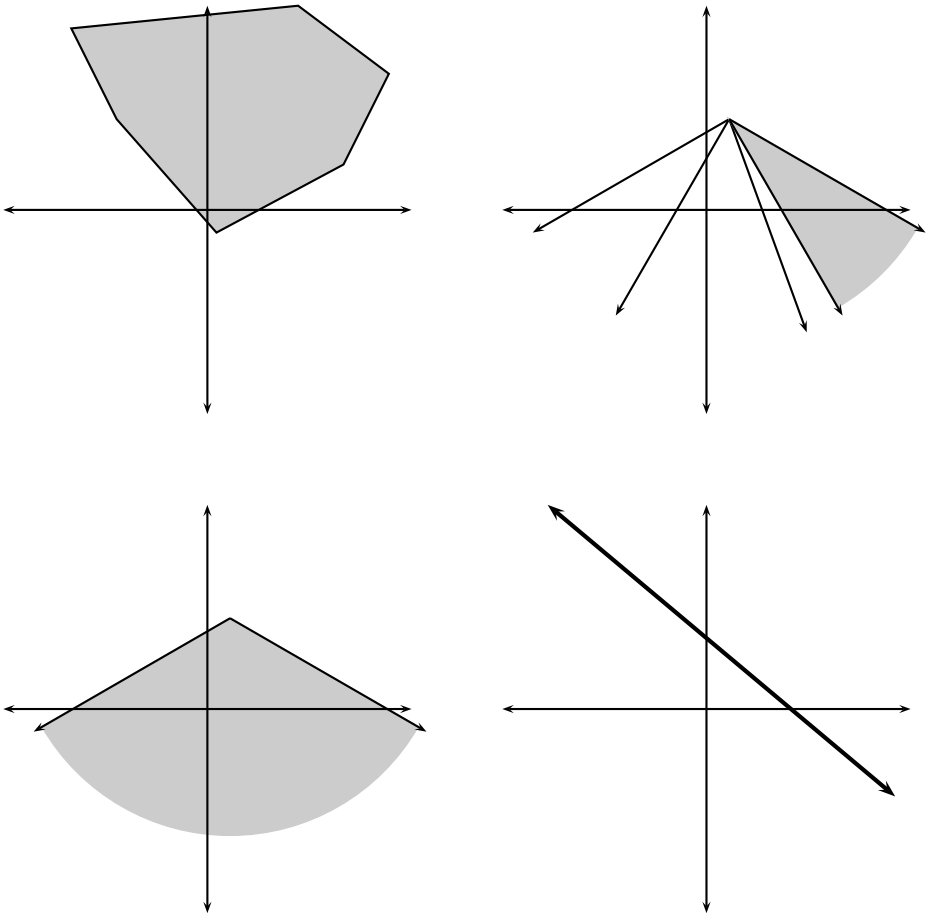

Figure B.1. An illustration of a convex set (top left), a cone (top right), a convex cone (bottom left), and an affine subspace (bottom right)

B.2. Combinations and hulls

We shall be interested in generating convex sets, cones, and affine subspaces containing given sets.

B.2 Definition: (Convex hull, coned hull, coned convex hull, affine hull) Let $S \subset \mathbb{R}^n$ be nonempty.

(i) A ***convex combination*** from S is a linear combination in $\mathbb{R}^n$ of the form

$$\sum_{j=1}^{k} \lambda_j v_j, \qquad k \in \mathbb{Z}_{>0},\ \lambda_1, \ldots, \lambda_k \in \mathbb{R}_{\geq 0},\ \sum_{j=1}^{k} \lambda_j = 1,\ v_1, \ldots, v_k \in S.$$

(ii) The ***convex hull*** of S, denoted by $\operatorname{conv}(S)$, is the smallest convex subset of $\mathbb{R}^n$ containing S.

(iii) The ***coned hull*** of S, denoted by $\operatorname{cone}(S)$, is the smallest cone in $\mathbb{R}^n$ containing S.

(iv) A ***coned convex combination*** from S is a linear combination in $\mathbb{R}^n$ of the form

$$\sum_{j=1}^{k} \lambda_j v_j, \qquad k \in \mathbb{Z}_{>0},\ \lambda_1, \ldots, \lambda_k \in \mathbb{R}_{\geq 0},\ v_1, \ldots, v_k \in S.$$

(v) The ***coned convex hull*** of S, denoted by $\operatorname{conv}\operatorname{cone}(S)$, is the smallest convex cone in $\mathbb{R}^n$ containing S.

(vi) An ***affine combination*** from S is a linear combination in $\mathbb{R}^n$ of the form

$$\sum_{j=1}^{k} \lambda_j v_j, \qquad k \in \mathbb{Z}_{>0},\ \lambda_1, \ldots, \lambda_k \in \mathbb{R},\ \sum_{j=1}^{k} \lambda_j = 1,\ v_1, \ldots, v_k \in S.$$

(vii) The ***affine hull*** of S, denoted by $\operatorname{aff}(S)$, is the smallest affine subspace of $\mathbb{R}^n$ containing S. •

B.3 Remark: (Sensibility of hull definitions) The definitions of $\operatorname{conv}(S)$, $\operatorname{cone}(S)$, $\operatorname{conv}\operatorname{cone}(S)$, and $\operatorname{aff}(S)$ make sense because intersections of convex sets are convex, intersections of cones are cones, and intersections of affine subspaces are affine subspaces. •

The terms "coned hull" and "coned convex hull" are not standard. In the literature these will often be called the "cone generated by S" and the "convex cone generated by S," respectively.

Convex combinations have the following useful property which also describes the convex hull.

B.4 Proposition: (The convex hull is the set of convex combinations) *Let $S \subset \mathbb{R}^n$ be nonempty and denote by $C(S)$ the set of convex combinations from S. Then $C(S) = \operatorname{conv}(S)$.*

Proof: First we show that $C(S)$ is convex. Consider two elements of $C(S)$ given by

$$x = \sum_{j=1}^{k} \lambda_j u_j, \quad y = \sum_{l=1}^{m} \mu_l v_l.$$

Then, for $s \in [0, 1]$ we have

$$sx + (1 - s)y = \sum_{j=1}^{k} s\lambda_j u_j + \sum_{l=1}^{m} (1 - s)\mu_j v_j.$$

For $r \in \{1, \ldots, k + m\}$ define

$$w_r = \begin{cases} u_r, & r \in \{1, \ldots, k\}, \\ v_{r-k}, & r \in \{k + 1, \ldots, k + m\} \end{cases}$$

and

$$\rho_r = \begin{cases} s\lambda_r, & r \in \{1, \ldots, k\}, \\ (1 - s)\mu_{r-k}, & r \in \{k + 1, \ldots, k + m\}. \end{cases}$$

Clearly $w_r \in S$ and $\rho_r \geq 0$ for $r \in \{1, \ldots, k+m\}$. Also,

$$\sum_{r=1}^{k+m} \rho_r = \sum_{j=1}^{k} s\lambda_j + \sum_{l=1}^{m}(1-s)\mu_l = s + (1-s) = 1.$$

Thus $sx + (1-s)y \in C(S)$, and so $C(S)$ is convex.

This necessarily implies that $\operatorname{conv}(S) \subset C(S)$ since $\operatorname{conv}(S)$ is the smallest convex set containing S. To show that $C(S) \subset \operatorname{conv}(S)$ we will show by induction on the number of elements in the linear combination that all convex combinations are contained in the convex hull. This is obvious for the convex combination of one vector. So suppose that every convex combination of the form

$$\sum_{j=1}^{k} \lambda_j u_j, \qquad k \in \{1, \ldots, m\},$$

is in $\operatorname{conv}(S)$, and consider a convex combination from S of the form

$$y = \sum_{l=1}^{m+1} \mu_l v_l = \sum_{l=1}^{m} \mu_l v_l + \mu_{m+1} v_{m+1}.$$

If $\sum_{l=1}^{m} \mu_l = 0$ then $\mu_l = 0$ for each $l \in \{1, \ldots, m\}$. Thus $y \in \operatorname{conv}(S)$ by the induction hypothesis. So we may suppose that $\sum_{l=1}^{m} \mu_l \neq 0$ which means that $\mu_{m+1} \neq 1$. Let us define $\mu'_l = \mu_l(1-\mu_{m+1})^{-1}$ for $l \in \{1, \ldots, m\}$. Since

$$1 - \mu_{m+1} = \sum_{l=1}^{m} \mu_l$$

it follows that

$$\sum_{l=1}^{m} \mu'_l = 1.$$

Therefore,

$$\sum_{l=1}^{m} \mu'_l v_l \in \operatorname{conv}(S)$$

by the induction hypothesis. But we also have

$$y = (1-\mu_{m+1}) \sum_{l=1}^{m} \mu'_l v_l + \mu_{m+1} v_{m+1}$$

by direct computation. Therefore, y is a convex combination of two elements of $\operatorname{conv}(S)$. Since $\operatorname{conv}(S)$ is convex, this means that $y \in \operatorname{conv}(S)$, giving the result. ■

For cones one has a similar result.

B.5 Proposition: (The set of positive multiples is the coned hull) *Let $S \subset \mathbb{R}^n$ be nonempty and denote*

$$K(S) = \{\lambda x \mid \ x \in S, \ \lambda \in \mathbb{R}_{\geq 0}\}.$$

Then $K(S) = \operatorname{cone}(S)$.

Proof: Note that $K(S)$ is clearly a cone which contains S. Thus $\text{cone}(S) \subset K(S)$. Now suppose that $y \in K(S)$. Thus $y = \lambda x$ for $x \in S$ and $\lambda \in \mathbb{R}_{\geq 0}$. Since $\text{cone}(S)$ is a cone containing x, we must have $y \in \text{cone}(S)$, giving $K(S) \subset \text{cone}(S)$. ■

Finally, one has an interpretation along these lines for convex cones.

B.6 Proposition: (The coned convex hull is the set of coned convex combinations) *Let $S \subset \mathbb{R}^n$ be nonempty and denote by $K'(S)$ the set of coned convex combinations from S. Then $K'(S) = \text{conv}\,\text{cone}(S)$.*

Proof: We first show that if $x, y \in K'(S)$ then $x + y \in K'(S)$ and that if $x \in K'(S)$ then $\lambda x \in K'(S)$ for $\lambda \in \mathbb{R}_{\geq 0}$. The second of these assertions is obvious. For the first, let $z = \frac{1}{2}x + \frac{1}{2}y$. Then $z \in K'(S)$ and so $2z = x + y \in K'(S)$. Thus $K'(S)$ is closed under addition and positive scalar multiplication. From this it immediately follows that $(1 - s)x + sy \in K'(S)$ for any $x, y \in K'(S)$ and $s \in [0, 1]$. Thus $K'(S)$ is convex. It is evident that $K'(S)$ is also a cone, and so we must have $\text{conv}\,\text{cone}(S) \subset K'(S)$.

Now let

$$y = \sum_{j=1}^{k} \lambda_j v_j \in K'(S).$$

By the fact that $\text{conv}\,\text{cone}(S)$ is a cone containing S we must have $k\lambda_j v_j \in \text{conv}\,\text{cone}(S)$ for $j \in \{1, \ldots, k\}$. Since $\text{conv}\,\text{cone}(S)$ is convex and contains $k\lambda_j v_j$ for $j \in \{1, \ldots, k\}$ we must have

$$\sum_{j=1}^{k} \frac{1}{k}(k\lambda_j v_j) = y \in \text{conv}\,\text{cone}(S),$$

giving the result. ■

Finally, we prove the expected result for affine subspaces, namely that the affine hull is the set of affine combinations. In order to do this we first give a useful characterisation of affine subspaces.

B.7 Proposition: (Characterisation of an affine subspace) *A nonempty subset $A \subset \mathbb{R}^n$ is an affine subspace if and only if there exists $x_0 \in \mathbb{R}^n$ and a subspace $U \subset \mathbb{R}^n$ such that*

$$A = \{x_0 + u \mid u \in U\}.$$

Proof: Let $x_0 \in A$ and define $U = \{x - x_0 \mid x \in A\}$. The result will be proved if we prove that U is a subspace. Let $x - x_0 \in U$ for some $x \in A$ and $a \in \mathbb{R}$. Then

$$a(x - x_0) = ax + (1 - a)x_0 - x_0,$$

and so $a(x - x_0) \in U$ since $ax + (1 - a)x_0 \in A$. For $x_1 - x_0, x_2 - x_0 \in U$ with $x_1, x_2 \in A$ we have

$$(x_1 - x_0) + (x_2 - x_0) = (x_1 + x_2 - x_0) - x_0.$$

Thus we will have $(x_1 - x_0) + (x_2 - x_0) \in U$ if we can show that $x_1 + x_2 - x_0 \in A$. However, we have

$$\begin{aligned}
& x_1 - x_0, x_2 - x_0 \in U, \\
\implies \quad & 2(x_1 - x_0), 2(x_2 - x_0) \in U, \\
\implies \quad & 2(x_1 - x_0) + x_0, 2(x_2 - x_0) + x_0 \in A, \\
\implies \quad & \tfrac{1}{2}(2(x_1 - x_0) + x_0) + \tfrac{1}{2}(2(x_2 - x_0) + x_0) \in A,
\end{aligned}$$

which gives the result after we notice that

$$\tfrac{1}{2}(2(x_1 - x_0) + x_0) + \tfrac{1}{2}(2(x_2 - x_0) + x_0) = x_1 + x_2 - x_0. \qquad \blacksquare$$

Now we can characterise the affine hull as the set of affine combinations.

B.8 Proposition: (The affine hull is the set of affine combinations) *Let $S \subset \mathbb{R}^n$ be nonempty and denote by $A(S)$ the set of affine combinations from S. Then $A(S) = \operatorname{aff}(S)$.*

Proof: We first show that the set of affine combinations is an affine subspace. Choose $x_0 \in S$ and define

$$U(S) = \{v - x_0 \mid \ v \in A(S)\}.$$

We first claim that $U(S)$ is the set of linear combinations of the form

$$\sum_{j=1}^{k} \lambda_j v_j, \qquad k \in \mathbb{Z}_{>0},\ \lambda_1, \dots, \lambda_k \in \mathbb{R},\ \sum_{j=1}^{k} \lambda_j = 0,\ v_1, \dots, v_k \in S. \tag{B.1}$$

To see this, note that if

$$u = \sum_{j=1}^{k} \lambda_j u_j - x_0 \in U(S)$$

then we can write

$$u = \sum_{j=1}^{k+1} \lambda_j u_j, \qquad \lambda_1, \dots, \lambda_{k+1} \in \mathbb{R},\ \sum_{j=1}^{k+1} \lambda_j = 0,\ u_1, \dots, u_{k+1} \in S,$$

by taking $\lambda_{k+1} = -1$ and $u_{k+1} = x_0$. Similarly, consider a linear combination of the form (B.1). We can without loss of generality suppose that $x_0 \in \{v_1, \dots, v_k\}$, since if this is not true we can simply add $0x_0$ to the sum. Thus we suppose, without loss of generality, that $v_k = x_0$. We then have

$$u = \Big(\sum_{j=1}^{k-1} \lambda_j v_j + (\lambda_k + 1)x_0\Big) - x_0.$$

Since the term in the parenthesis is clearly an element of $A(S)$ it follows that $u \in U(S)$.

With this characterisation of $U(S)$ it is then easy to show that $U(S)$ is a subspace of $\mathbb{R}^n$. Moreover, it is immediate from Proposition B.8 that $A(S)$ is then an affine subspace. Since $\operatorname{aff}(S)$ is the smallest affine subspace containing S it follows that $\operatorname{aff}(S) \subset A(S)$. To show that $A(S) \subset \operatorname{aff}(S)$ we use induction on the number of elements in an affine combination in $A(S)$. For an affine combination with one term this is obvious. So suppose that every affine combination of the form

$$\sum_{j=1}^{k} \lambda_j v_j, \qquad k \in \{1, \dots, m\},$$

is in $\operatorname{aff}(S)$ and consider an affine combination of the form

$$x = \sum_{j=1}^{m+1} \lambda_j v_j = \sum_{j=1}^{m} \lambda_j v_j + \lambda_{m+1} v_{m+1}.$$

It must be the case that at least one of the numbers $\lambda_1, \dots, \lambda_{m+1}$ is not equal to 1. So, without loss of generality suppose that $\lambda_{m+1} \neq 1$ and then define $\lambda'_j = (1 - \lambda_{m+1}^{-1})\lambda_j$, $j \in \{1, \dots, m\}$. We then have

$$\sum_{j=1}^{m} \lambda'_j = 1,$$

so that

$$\sum_{j=1}^{m} \lambda'_j v_j \in \operatorname{aff}(S)$$

by the induction hypothesis. It then holds that

$$x = (1 - \lambda_{m+1}) \sum_{j=1}^{m} \lambda'_j v_j + \lambda_{m+1} v_{m+1}.$$

This is then in $\operatorname{aff}(S)$. ■

B.3. Topology of convex sets and cones

Let us now say a few words about the topology of convex sets. Note that every convex set is a subset of its affine hull. Moreover, as a subset of its affine hull, a convex set has an interior.

B.9 Definition: (Relative interior and relative boundary) If $C \subset \mathbb{R}^n$ is a convex set, the set

$$\operatorname{rel\,int}(C) = \{x \in C \mid x \in \operatorname{int}_{\operatorname{aff}(C)}(C)\}$$

is the ***relative interior*** of C and the set $\operatorname{rel\,bd}(C) = \operatorname{cl}(C) \setminus \operatorname{rel\,int}(C)$ is the ***relative boundary*** of C. •

The point is that, while a convex set may have an empty interior, its interior can still be defined in a weaker, but still useful, sense. The notion of relative interior leads to the following useful concept.

B.10 Definition: (Dimension of a convex set) Let $C \subset \mathbb{R}^n$ be convex and let $U \subset \mathbb{R}^n$ be the subspace for which $\operatorname{aff}(C) = \{x_0 + u \mid u \in U\}$ for some $x_0 \in \mathbb{R}^n$. The ***dimension*** of C, denoted by $\dim(C)$, is the dimension of the subspace U. •

The following result will be used in our development.

B.11 Proposition: (Closures and relative interiors of convex sets and cones are convex sets and cones) *Let $C \subset \mathbb{R}^n$ be convex and let $K \subset \mathbb{R}^n$ be a convex cone. Then*

(i) $\operatorname{cl}(C)$ *is convex and* $\operatorname{cl}(K)$ *is a convex cone and*

(ii) $\operatorname{rel\,int}(C)$ *is convex and* $\operatorname{rel\,int}(K)$ *is a convex cone.*

Moreover, $\operatorname{aff}(C) = \operatorname{aff}(\operatorname{cl}(C))$ *and* $\operatorname{aff}(K) = \operatorname{aff}(\operatorname{cl}(K))$.

Proof: (i) Let $x, y \in \operatorname{cl}(C)$ and let $s \in [0, 1]$. Suppose that $\{x_j\}_{j \in \mathbb{Z}_{>0}}$ and $\{y_j\}_{j \in \mathbb{Z}_{>0}}$ are sequences in C converging to x and y, respectively. Note that $sx_j + (1 - s)y_j \in C$ for each $j \in \mathbb{Z}_{>0}$. Moreover, if $\epsilon > 0$ then

$$\|sx + (1 - s)y - sx_j - (1 - s)y_j\| \leq s\|x - x_j\| + (1 - s)\|y - y_j\| < \epsilon,$$

provided that j is sufficiently large that $s\|x - x_j\| < \frac{\epsilon}{2}$ and $(1-s)\|y - y_j\| < \frac{\epsilon}{2}$. Thus the sequence $\{sx_j + (1-s)y_j\}_{j\in\mathbb{Z}_{>0}}$ converges to $sx + (1-s)y$ and so $sx + (1-s)y \in \mathrm{cl}(C)$. This shows that $\mathrm{cl}(C)$ is convex. Since $C \subset \mathrm{cl}(C)$ it follows that $\mathrm{aff}(C) \subset \mathrm{aff}(\mathrm{cl}(C))$. Moreover, since $C \subset \mathrm{aff}(C)$ and since $\mathrm{aff}(C)$ is closed we have

$$\mathrm{cl}(C) \subset \mathrm{cl}(\mathrm{aff}(C)) = \mathrm{aff}(C),$$

so giving $\mathrm{aff}(C) = \mathrm{aff}(\mathrm{cl}(C))$ as desired.

An entirely similar argument shows that $\mathrm{cl}(K)$ is convex and that $\mathrm{aff}(K) = \mathrm{aff}(\mathrm{cl}(K))$.

(ii) Let us first consider the convex set C. To simplify matters, since the relative interior is the interior relative to the affine subspace containing C, and since the topology of an affine subspace is "the same as" Euclidean space, we shall assume that $\dim(C) = n$ and show that $\mathrm{int}(C)$ is convex.

We first prove a lemma.

1 Lemma: *If C is a convex set, if $x \in \mathrm{rel\,int}(C)$, and if $y \in \mathrm{cl}(C)$ then*

$$[x, y) \triangleq \{sx + (1-s)y \mid s \in [0, 1)\}$$

is contained in $\mathrm{rel\,int}(C)$.

Proof: As in the proof of (ii), let us assume, without loss of generality, that $\dim(C) = n$. Since $x \in \mathrm{int}(C)$ there exists $r > 0$ such that $B(x, r) \subset C$. Since $y \in \mathrm{cl}(C)$, for every $\epsilon > 0$ there exists $y_\epsilon \in C \cap B(y, \epsilon)$. Let $z = \alpha x + (1-\alpha y) \in [x, y)$ for $\alpha \in [0, 1)$, and define $\delta = \alpha r - (1-\alpha)\epsilon$. If ϵ is sufficiently small we can ensure that $\delta \in \mathbb{R}_{>0}$, and we assume that ϵ is so chosen. For $z' \in B(z, \delta)$ we have

$$\begin{aligned}
& \|z' - z\| < \delta \\
\Longrightarrow \quad & \|z' - (\alpha x + (1-\alpha)y_\epsilon + (1-\alpha)(y - y_\epsilon))\| < \delta \\
\Longrightarrow \quad & \|z' - (\alpha x + (1-\alpha)y_\epsilon)\| \leq \delta + (1-\alpha)\epsilon = \alpha r \\
\Longrightarrow \quad & z' \in \{\alpha x' + (1-\alpha)y_\epsilon \mid x' \in B(x, r)\}.
\end{aligned}$$

Since $y_\epsilon \in C$ and $B(x, r) \subset C$ it follows that $z' \in C$ and so $B(z, \delta) \subset C$. This gives our claim that $[x, y) \subset \mathrm{int}(C)$. ▼

That $\mathrm{int}(C)$ is convex follows immediately since, if $x, y \in \mathrm{int}(C)$, Lemma 1 ensures that the line segment connecting x and y is contained in $\mathrm{int}(C)$.

Now consider the convex cone K. We know now that $\mathrm{rel\,int}(K)$ is convex so we need only show that it is a cone. This, however, is obvious. Indeed, if $x \in \mathrm{rel\,int}(K)$ suppose that $\lambda x \notin \mathrm{rel\,int}(K)$ for some $\lambda \in \mathbb{R}_{>0}$. Since $\lambda x \in K$ we must then have $\lambda x \in \mathrm{bd}(K)$. By Lemma 1 this means that $(\lambda + \epsilon)x \notin K$ for all $\epsilon \in \mathbb{R}_{>0}$. This contradicts the fact that K is a cone. ■

The following result will also come up in our constructions.

B.12 Proposition: (The closure of the relative interior) *If $C \subset \mathbb{R}^n$ is a convex set then* $\mathrm{cl}(\mathrm{rel\,int}(C)) = \mathrm{cl}(C)$.

Proof: It is clear that $\mathrm{cl}(\mathrm{rel\,int}(C)) \subset \mathrm{cl}(C)$. Let $x \in \mathrm{cl}(C)$ and let $y \in \mathrm{rel\,int}(C)$. By Lemma 1 in the proof of Proposition B.11 it follows that the half-open line segment $[y, x)$ is contained in $\mathrm{rel\,int}(C)$. Therefore, there exists a sequence $\{x_j\}_{j\in\mathbb{Z}_{>0}}$ in this line segment, and so in $\mathrm{rel\,int}(C)$, converging to x. Thus $x \in \mathrm{cl}(\mathrm{rel\,int}(C))$. ■

B.4. Simplices and simplex cones

We now concern ourselves with special examples of convex sets and convex cones, and show that these special objects can always be found as neighbourhoods in general convex sets and cones.

We begin with the definitions.

B.13 Definition: (Affine independence, simplex, simplex cone) Let $n \in \mathbb{Z}_{>0}$.

(i) A set $\{x_0, x_1, \dots, x_k\} \subset \mathbb{R}^n$ is ***affinely independent*** if the set $\{x_1 - x_0, \dots, x_k - x_0\}$ is linearly independent.

(ii) A ***k-simplex*** is the convex hull of a set $\{x_0, x_1, \dots, x_k\}$ of affinely independent points.

(iii) A ***k-simplex cone*** is the coned convex hull of a set $\{x_1, \dots, x_k\}$ which is linearly independent. •

Let us give the standard examples of such objects.

B.14 Examples: (Standard n-simplex, standard n-simplex cone)

1. The ***standard n-simplex*** is the subset of $\mathbb{R}^n$ given by

$$\Delta_n = \Big\{ x \in \mathbb{R}^n \;\Big|\; x_1, \dots, x_n \geq 0, \; \sum_{j=1}^{n} x_j \leq 1 \Big\}.$$

Thus Δ_n is the convex hull of the n standard basis vectors along with the origin.

2. The ***standard n-simplex cone*** is the subset of $\mathbb{R}^n$ given by

$$K_n = \{ x \in \mathbb{R}^n \mid x_1, \dots, x_n \geq 0 \}.$$

Note that K_n is the coned convex hull of the n standard basis vectors.

In Figure B.2 we depict the standard n-simplex and the standard n-simplex cone when $n = 2$. •

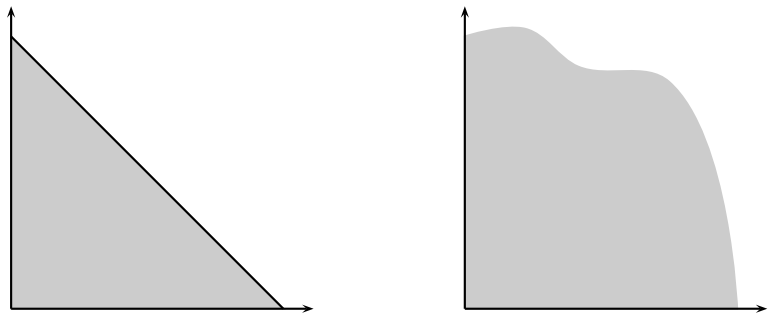

Figure B.2. The standard 2-simplex (left) and the standard 2-simplex cone (right)

The following result about the dimension of simplices and simplex cones is intuitively clear.

B.15 Proposition: (Dimension of simplices and simplex cones) *If $C, K \subset \mathbb{R}^n$ are a k-simplex and a k-simplex cone, respectively, then* $\dim(C) = \dim(K) = k$.

Proof: Let us first consider the k-simplex C defined by points $\{x_0, x_1, \ldots, x_k\}$. Clearly $\operatorname{aff}(\{x_0, x_1, \ldots, x_k\}) \subset \operatorname{aff}(C)$ since $\{x_0, x_1, \ldots, x_k\} \subset C$. Let $x \in \operatorname{aff}(C)$ so that

$$x = \sum_{l=1}^{m} \mu_l y_l$$

for $\mu_1, \ldots, \mu_m \in \mathbb{R}$ summing to 1 and for $y_1, \ldots, y_m \in C$. For each $l \in \{1, \ldots, m\}$ we have

$$y_l = \sum_{j=0}^{k} \lambda_{lj} x_j$$

for $\lambda_{l0}, \lambda_{l1}, \ldots, \lambda_{lk} \in \mathbb{R}_{\geq 0}$ summing to 1. Therefore,

$$x = \sum_{l=1}^{m} \sum_{j=0}^{k} \mu_j \lambda_{lj} x_j = \sum_{j=0}^{k} \Big(\sum_{l=1}^{m} \lambda_{lj} \mu_l \Big) x_j,$$

and so $x \in \operatorname{aff}(\{x_0, x_1, \ldots, x_k\})$ since

$$\sum_{j=0}^{k} \Big(\sum_{l=1}^{m} \lambda_{lj} \mu_l \Big) = 1.$$

Thus $\operatorname{aff}(C) = \operatorname{aff}(\{x_0, x_1, \ldots, x_k\})$. That $\dim(C) = k$ follows since the subspace corresponding to the affine subspace $\operatorname{aff}(\{x_0, x_1, \ldots, x_k\})$ is generated by $\{x_1 - x_0, \ldots, x_k - x_0\}$, and this subspace has dimension k.

The proof for the k-simplex cone K follows in an entirely similar manner, merely with convex combinations being replaced by coned convex combinations. ■

One of the things we will need to be able to do is find neighbourhoods of points in convex sets and convex cones that are simplices and simplex cones, respectively. For convex sets we have the following result.

B.16 Proposition: (Existence of simplicial neighbourhoods) *Let $C \subset \mathbb{R}^n$ be convex and of dimension k, let $x_0 \in \operatorname{rel\,int}(C)$, and let $\mathcal{U}$ be a neighbourhood of x_0 in $\mathbb{R}^n$. Then there exists a k-simplex $C_0 \subset C$ such that $C_0 \subset \mathcal{U}$ and $x_0 \in \operatorname{rel\,int}(C_0)$.*

Proof: Let $r \in \mathbb{R}_{>0}$ be such that $B(x_0, r) \subset \mathcal{U}$ and such that $B(x_0, r) \cap \operatorname{aff}(C) \subset C$. The existence of such an r follows since $x_0 \in \operatorname{rel\,int}(C)$. Let $\{v_1, \ldots, v_k\}$ be an orthogonal basis for the subspace $U(C)$ corresponding to $\operatorname{aff}(C)$ and suppose that $v_1, \ldots, v_k \in B(0, r)$. Then $y_j \triangleq x_0 + v_j \in B(x_0, r)$, $j \in \{1, \ldots, k\}$.

We now use a linear algebra lemma.

1 Lemma: *If V is a finite-dimensional $\mathbb{R}$-inner product space and if $\{v_1, \ldots, v_n\}$ is a basis for V, then there exists $v_0 \in \mathsf{V}$ such that $\langle v_0, v_j \rangle < 0$ for every $j \in \{1, \ldots, n\}$.*

Proof: Let $L\colon \mathsf{V} \to \mathbb{R}^n$ be the unique linear map defined by asking that $L(v_j)$ be equal to e_j, the jth standard basis vector for $\mathbb{R}^n$. Note that if we take $e_0 = (-1, \dots, -1) \in \mathbb{R}^n$ then, with respect to the standard inner product, $\langle e_0, e_j \rangle = -1 < 0$, $j \in \{1, \dots, n\}$. Let $\alpha \in (\mathbb{R}^n)^*$ correspond to e_0 under the identification of $\mathbb{R}^n$ with $(\mathbb{R}^n)^*$ induced by the standard inner product and take $\beta = L^*(\alpha)$. Then

$$\beta(v_j) = L^*(\alpha) \cdot v_j = \alpha \cdot L(v_j) = \alpha \cdot e_j = -1$$

for $j \in \{1, \dots, n\}$. Then take v_0 to correspond to β under the identification of V^* with V using the inner product on V. We then have $\langle v_0, v_j \rangle = -1$, $j \in \{1, \dots, n\}$. ▼

We now apply the lemma to the subspace $U(C)$ to assert the existence of $v_0 \in U(C)$ such that $\langle v_0, v_j \rangle < 0$ for $j \in \{1, \dots, k\}$. We may assume that $\|v_0\| < r$. We claim that the set $\{v_0, v_1, \dots, v_k\}$ is affinely independent. Indeed, suppose that

$$c_1(v_1 - v_0) + \cdots + c_k(v_k - v_0) = 0.$$

Then $c_j(\langle v_j, v_j \rangle - \langle v_j, v_0 \rangle) = 0$ for $j \in \{1, \dots, k\}$. Since $\langle v_j, v_j \rangle - \langle v_j, v_0 \rangle > 0$ it follows that $c_j = 0$ for $j \in \{1, \dots, k\}$, so giving affine independence of $\{v_0, v_1, \dots, v_k\}$. Define $y_0 = x_0 + v_0 \in B(x_0, r)$ and take $C_0 = \operatorname{conv}(\{y_0, y_1, \dots, y_k\})$.

We claim that $C_0 \in B(x_0, r) \subset \mathcal{U}$. Indeed, if $x \in C_0$ then we can write x as a convex combination:

$$x = \sum_{j=0}^{k} \lambda_j y_j \quad \Longrightarrow \quad x - x_0 = \sum_{j=0}^{k} \lambda_j (y_j - x_0) = \sum_{j=0}^{k} v_j.$$

Applying the triangle inequality a bunch of times gives

$$\|x - x_0\| \leq \sum_{j=0}^{k} \lambda_j \|v_j\| < r,$$

as desired.

Finally, we claim that $x_0 \in \operatorname{rel\,int}(C_0)$. This will follow if we can show that $0 \in \operatorname{rel\,int}(\operatorname{conv}(\{v_0, v_1, \dots, v_k\}))$. By the lemma above and since we chose the basis $\{v_1, \dots, v_j\}$ to be orthogonal,

$$v_0 = \sum_{j=1}^{k} \frac{\langle v_j, v_0 \rangle}{\|v_j\|^2} v_j \quad \Longrightarrow \quad \|v_1\|^2 \cdots \|v_k\|^2 v_0 - \sum_{j=1}^{k} \langle v_j, v_0 \rangle v_j = 0,$$

showing that 0 is a linear combination of the vectors $\{v_0, v_1, \dots, v_k\}$ with the coefficients being strictly positive. By scaling the coefficients this linear combination can be made convex with all coefficients positive. Therefore, $0 \in \operatorname{rel\,int}(\operatorname{conv}(\{v_0, v_1, \dots, v_k\}))$, as desired. ■

For cones we have a similarly styled result.

B.17 Proposition: (Existence of simplex cone neighbourhoods) *Let $K \subset \mathbb{R}^n$ be a convex cone of dimension k, let $x_0 \in \operatorname{rel\,int}(K) \setminus \{0\}$, and let $\mathcal{U}$ be a neighbourhood of $x_0 \in \mathbb{R}^n$. Then there exists a k-simplex cone $K_0 \subset K$ such that $K_0 \subset \operatorname{cone}(\mathcal{U})$ and $x_0 \in \operatorname{rel\,int}(K_0)$.*

Proof: Denote by P_{x_0} the orthogonal complement to x_0 and let

$$\mathcal{U}_{x_0} = \{v \in P_{x_0} \mid x_0 + v \in \mathcal{U}\}.$$

Note that $\mathcal{U}_{x_0}$ is a neighbourhood of 0 in P_{x_0}. By Proposition B.16 let $C_0 \subset P_{x_0}$ be a $(k-1)$-simplex contained in $\mathcal{U}_{x_0}$ and having 0 in its relative interior. Then define K_0 to be the coned convex hull of $x_0 + C_0 \triangleq \{x_0 + v \mid v \in C_0\}$, noting that K_0 is then the coned convex hull of the points $x_j \triangleq x_0 + v_j$, $j = \{1, \ldots, k\}$, where the points $v_1, \ldots, v_k$ are defined so that their convex hull is C_0.

We claim that $K_0 \subset \text{cone}(\mathcal{U})$. This follows since

$$x_0 + C_0 \subset \{x_0 + v \mid v \in \mathcal{U}_{x_0}\} \subset \mathcal{U},$$

and so $K_0 = \text{cone}(x_0 + C_0) \subset \text{cone}(\mathcal{U})$.

We also claim that $x_0 \in \text{rel int}(K_0)$. Since $0 \in \text{rel int}(C_0)$ we can write

$$x_0 = x_0 + 0 = x_0 + \sum_{j=1}^{k} \lambda_j v_j$$

for appropriate $\lambda_1, \ldots, \lambda_k \in \mathbb{R}_{>0}$ summing to 1. Therefore

$$x_0 = \sum_{j=1}^{k} \lambda_j (v_j + x_0),$$

and so x_0 is a linear combination of the points $x_1, \ldots, x_k$ with strictly positive coefficients. Thus $x_0 \in \text{rel int}(K_0)$. ■

If C is the simplex obtained by taking the convex hull of the points $\{x_0, x_1, \ldots, x_k\}$, then every point $x \in C$ is uniquely written as

$$x = \sum_{j=0}^{k} \lambda_j v_j$$

for $\lambda_0, \ldots, \lambda_k \in \mathbb{R}_{\geq 0}$ summing to 1. Note that the set of λ's appearing in such a linear combination have the property that the point

$$\sum_{j=0}^{k} \lambda_j e_j$$

lies in the standard k-simplex if we take the convention that $e_0 = 0$. Indeed the map

$$\sum_{j=0}^{k} \lambda_j e_j \mapsto \sum_{j=0}^{k} \lambda_j v_j$$

defines a homeomorphism of Δ_k with C. This parameterisation of a simplex C by $(\lambda_0, \lambda_1, \ldots, \lambda_k)$ defines ***barycentric coordinates*** for C.

A similar construction can be made for a k-simplex cone $K = \operatorname{conv}\operatorname{cone}(\{x_1, \dots, x_k\})$. We fix some nonzero vector $v_0 \in \operatorname{rel}\operatorname{int}(K) \setminus \{0\}$ and let P_{v_0} be the orthogonal complement to v_0. We may suppose, without loss of generality (by scaling if necessary), that

$$x_1, \dots, x_k \in \{v_0 + x \mid x \in P_{v_0}\},$$

i.e., that the points $x_1, \dots, x_k$ lie in a plane parallel to P_{v_0} passing through v_0. We then define a $(k-1)$-simplex $C_{v_0} \subset P_{v_0}$ by asking that

$$C_{v_0} = \{x \in P_{v_0} \mid x + v_0 \in K\}$$

(we leave it to the reader to check that C_{v_0} is indeed a $(k-1)$-simplex). We then let $(\lambda_1, \dots, \lambda_k)$ be barycentric coordinates for C_{v_0}. A point in $x \in K$ is then uniquely specified by $(l(x), \lambda(x))$ where $l(x) = \frac{\langle x, v_0 \rangle}{\|v_0\|}$ and where $\lambda(x)$ are the barycentric coordinates in C_{v_0} for the point $l(x)x - v_0$. We call the coordinates (l, λ) for K ***barycentric coordinates***. For the reader for whom this definition of coordinates (l, λ) for K is not immediately clear, we give an illustration of their meaning in Figure B.3. One can verify that

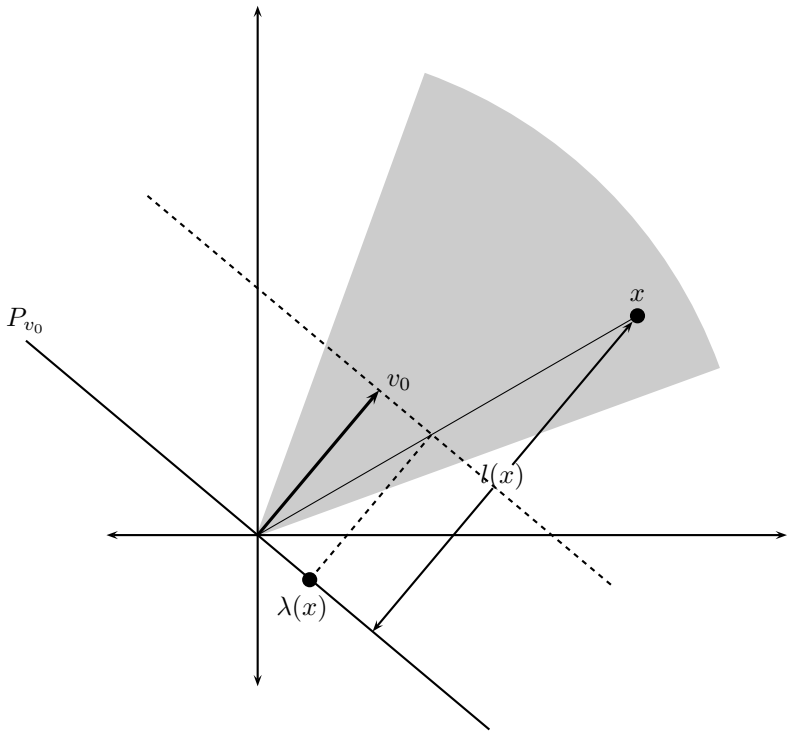

Figure B.3. Barycentric coordinates for a simplex cone

$$x = l(x)(\lambda_1(x)x_1 + \cdots + \lambda_k(x)x_k).$$

B.5. Separation theorems for convex sets

One of the most important properties of convex sets in convex analysis, and indeed for us in our proof of the Maximum Principle, is the notion of certain types of convex sets

being separated by hyperplanes. We shall only examine those parts of the theory that we will use; we refer to [Rockafellar 1970] for further discussion.

In order to make things clear, let us define all of our terminology precisely.

B.18 Definition: (Hyperplane, half-space, support hyperplane)

(i) A ***hyperplane*** in $\mathbb{R}^n$ is a subset of the form

$$\{x \in \mathbb{R}^n \mid \langle \lambda, x \rangle = a\}$$

for some $\lambda \in \mathbb{R}^n \setminus \{0\}$ and $a \in \mathbb{R}$. Such a hyperplane is denoted by $P_{\lambda,a}$.

(ii) A ***half-space*** in $\mathbb{R}^n$ is a subset of the form

$$\{x \in \mathbb{R}^n \mid \langle \lambda, x \rangle > a\}$$

for some $\lambda \in \mathbb{R}^n \setminus \{0\}$ and $a \in \mathbb{R}$. We shall denote

$$H^-_{\lambda,a} = \{x \in \mathbb{R}^n \mid \langle \lambda, x \rangle < a\}, \quad H^+_{\lambda,a} = \{x \in \mathbb{R}^n \mid \langle \lambda, x \rangle > a\}.$$

(iii) If $A \subset \mathbb{R}^n$, a ***support hyperplane*** for A is a hyperplane $P_{\lambda,a}$ such that $A \subset H^+_{\lambda,a} \cup P_{\lambda,a}$.

(iv) For subsets $A, B \subset \mathbb{R}^n$, a ***separating hyperplane*** is a hyperplane $P_{\lambda,a}$ for which

$$A \subset H^+_{\lambda,a} \cup P_{\lambda,a}, \ B \subset H^-_{\lambda,a} \cup P_{\lambda,a}. \qquad \bullet$$

The following result is a basis for many separation theorems for convex sets.

B.19 Theorem: (Convex sets possess supporting hyperplanes) *If $C \subset \mathbb{R}^n$ is a convex set not equal to $\mathbb{R}^n$, then C possesses a supporting hyperplane.*

Proof: Let $x_0 \notin \text{cl}(C)$, let $z \in C$, and define $r = \|x_0 - z\|$. Define $A = \text{cl}(C) \cap B(x_0, r)$ noting that A is a nonempty compact set. Define $f\colon A \to \mathbb{R}_{>0}$ by $f(y) = \|x_0 - y\|$. The map f is continuous and so there exists $y_0 \in A \subset \text{cl}(C)$ such that $f(y_0)$ is the minimum value of f. Let $\lambda = y_0 - x_0$ and $a = \langle y_0, y_0 - x_0 \rangle$. We will show that $P_{\lambda,a}$ is a support hyperplane for C.

First let us show that $P_{\lambda,a}$ separates $\{x_0\}$ and $\text{cl}(C)$. A direct computation shows that $\langle \lambda, x_0 \rangle = -\|x_0 - y_0\|^2 + a < a$. To show that $\langle \lambda, x \rangle \geq a$ for all $x \in \text{cl}(C)$, suppose otherwise. Thus let $x \in C$ be such that $\langle \lambda, x \rangle < a$. By Lemma 1 in the proof of Proposition B.11 the line segment from y to y_0 is contained in $\text{cl}(C)$. Define $g\colon [0,1] \to \mathbb{R}$ by $g(s) = \|(1-s)y_0 + sy - x_0\|^2$. Thus g is the square of the distance from x_0 to points on the line segment from y to y_0. Note that $g(s) \geq g(0)$ for all $s \in (0,1]$ since y_0 is the closest point in $\text{cl}(C)$ to x_0. A computation gives

$$g(s) = (1-s)^2\|y_0 - x_0\|^2 + 2s(1-s)\langle y - x_0, y_0 - x_0 \rangle + s^2\|y - x_0\|^2$$

and another computation gives $g'(0) = 2(\langle \lambda, y \rangle - a)$ which is strictly negative by our assumption about y. This means that g strictly decreases near zero, which contradicts the definition of y_0. Thus we must have $\langle \lambda, y \rangle \geq a$ for all $y \in \text{cl}(C)$. ■

During the course of the proof of the theorem we almost proved the following result.

B.20 Corollary: (Separation of convex sets and points) *If $C \subset \mathbb{R}^n$ is convex and if $x_0 \notin \operatorname{int}(C)$ then there exists a separating hyperplane for $\{x_0\}$ and C.*

Proof: If $x_0 \notin \operatorname{cl}(C)$ then the result follows immediately from the proof of Theorem B.19. If $x_0 \in \operatorname{bd}(C)$ then let $\{x_j\}_{j \in \mathbb{Z}_{>0}}$ be a sequence in $\mathbb{R}^n \setminus \operatorname{cl}(C)$ converging to x_0. For each $j \in \mathbb{Z}_{>0}$ let $\lambda_j \in \mathbb{R}^n \setminus \{0\}$ and $a_j \in \mathbb{R}$ have the property that

$$\begin{aligned} \langle \lambda_j, x_j \rangle &\leq a_j, \qquad j \in \mathbb{Z}_{>0}, \\ \langle \lambda_j, y \rangle &> a_j, \qquad y \in C,\ j \in \mathbb{Z}_{>0}. \end{aligned}$$

Let us without loss of generality take $a_j = \langle \lambda_j, x_j \rangle$; this corresponds to choosing the hyperplane separating C from x_j to pass through x_j. Let $\alpha_j = \frac{\lambda_j}{\|\lambda_j\|}$, $j \in \mathbb{Z}_{>0}$. The sequence $\{\alpha_j\}_{j \in \mathbb{Z}_{>0}}$ is a sequence in the $(n-1)$-sphere which is compact. Thus we can choose a convergent subsequence which we also denote, by an abuse of notation, by $\{\alpha_j\}_{j \in \mathbb{Z}_{>0}}$. Let $\alpha \in \mathbb{R}^n$ denote the limit of this sequence. Defining $c_j = \langle \alpha_j, x_j \rangle$ we then have

$$\begin{aligned} \langle \alpha_j, x_j \rangle &= c_j, \qquad j \in \mathbb{Z}_{>0}, \\ \langle \alpha_j, y \rangle &> c_j, \qquad y \in C,\ j \in \mathbb{Z}_{>0}. \end{aligned}$$

Let $c = \lim_{j \to \infty} c_j$. For $y \in C$ this gives

$$\begin{aligned} \langle \alpha, x_0 \rangle &= \lim_{j \to \infty} \langle \alpha_j, x_j \rangle = c, \\ \langle \alpha, y \rangle &= \lim_{j \to \infty} \langle \alpha_j, y \rangle \geq c, \end{aligned}$$

as desired. ■

The following consequence of Theorem B.19 is also of independent interest.

B.21 Corollary: (Disjoint convex sets are separated) *If $C_1, C_2 \subset \mathbb{R}^n$ are disjoint convex sets, then there exists a hyperplane separating C_1 and C_2.*

Proof: Define

$$C_1 - C_2 = \{x_1 - x_2 \mid x_1 \in C_1,\ x_2 \in C_2\}.$$

One checks directly that $C_1 - C_2$ is convex. Since C_1 and C_2 are disjoint it follows that $0 \notin C_1 - C_2$. By Theorem B.19 there exists a hyperplane P, passing through 0, separating $C_1 - C_2$ from 0. We claim that this implies that the same hyperplane P, appropriately translated, separates C_1 and C_2. To see this note that P gives rise to $\lambda \in \mathbb{R}^n \setminus \{0\}$ such that

$$\langle \lambda, x_1 - x_2 \rangle \geq 0, \qquad x_1 \in C_1,\ x_2 \in C_2.$$

Let

$$a_1 = \inf\{\langle \lambda, x_1 \rangle \mid x_1 \in C_1\}, \quad a_2 = \sup\{\langle \lambda, x_2 \rangle \mid x_2 \in C_2\}$$

so that $a_1 - a_2 \geq 0$. For any $a \in [a_2, a_1]$ we have

$$\begin{aligned} \langle \lambda, x_1 \rangle &\geq a, \qquad x_1 \in C_1, \\ \langle \lambda, x_2 \rangle &\leq a, \qquad x_2 \in C_2, \end{aligned}$$

giving the separation of C_1 and C_2, as desired. ■

We shall require the following quite general result concerning separation of convex sets by hyperplanes.

B.22 Theorem: (A general separation theorem) *If $C_1, C_2 \subset \mathbb{R}^n$ are convex sets, then they possess a separating hyperplane if and only if either of the following two conditions holds:*

(i) there exists a hyperplane P such that $C_1, C_2 \subset P$;

(ii) $\operatorname{rel\,int}(C_1) \cap \operatorname{rel\,int}(C_2) = \emptyset$.

Proof: Suppose that C_1 and C_2 possess a separating hyperplane P. Therefore, there exists $\lambda \in \mathbb{R}^n \setminus \{0\}$ and $a \in \mathbb{R}$ such that

$$\begin{aligned} \langle \lambda, x_1 \rangle &\geq a, \qquad x_1 \in C_1, \\ \langle \lambda, x_2 \rangle &\leq a, \qquad x_2 \in C_2. \end{aligned}$$

If $\langle \lambda, x \rangle = a$ for all $x \in C_1 \cup C_2$ then (i) holds. Now suppose that $\langle \lambda, x_1 \rangle > a$ for some $x_1 \in C_1$ (a similar argument will obviously apply if this holds for some $x_2 \in C_2$) and let $x_0 \in \operatorname{rel\,int}(C_1)$. Since P is a support hyperplane for C_1 and since $C_1 \not\subset P$, it follows that the relative interior, and so x_0, lies in the appropriate half-space defined by P. Since P separates C_1 and C_2 this precludes x_0 from being in C_2. Thus (ii) holds.

Now suppose that (i) holds. It is then clear that P is a separating hyperplane for C_1 and C_2.

Finally, suppose that (ii) holds. From Proposition B.11 and Corollary B.21 it holds that $\operatorname{rel\,int}(C_1)$ and $\operatorname{rel\,int}(C_2)$ possess a separating hyperplane. Thus there exists $\lambda \in \mathbb{R}^n \setminus \{0\}$ and $a \in \mathbb{R}$ such that

$$\begin{aligned} \langle \lambda, x_1 \rangle &\leq a, \qquad x_1 \in \operatorname{rel\,int}(C_1), \\ \langle \lambda, x_2 \rangle &\geq a, \qquad x_2 \in \operatorname{rel\,int}(C_2). \end{aligned}$$

Therefore, by Proposition B.12 we also have

$$\begin{aligned} \langle \lambda, x_1 \rangle &\leq a, \qquad x_1 \in \operatorname{cl}(C_1), \\ \langle \lambda, x_2 \rangle &\geq a, \qquad x_2 \in \operatorname{cl}(C_2), \end{aligned}$$

which implies this part of the theorem. ■

B.6. Linear functions on convex polytopes

In our study of linear time-optimal control we will ask that controls take their values in a convex polytope (to be defined shortly). It turns out that we will also seek to maximise a linear function on a convex polytope. This is a well studied problem, going under the general name of ***linear programming***. In this section we shall define all the terminology needed in this problem, and give the main result in linear programming that we shall make use of.

First let us talk about convex polytopes. Notationally it will be convenient, for $x, y \in \mathbb{R}^n$, to write $x \leq y$ if $x^i \leq y^i$, $i \in \{1, \dots, n\}$.

B.23 Definition: (Convex polyhedron, convex polytope) A nonempty subset $C \subset \mathbb{R}^n$ is

(i) a ***convex polyhedron*** if there exists $A \in \mathrm{L}(\mathbb{R}^n; \mathbb{R}^k)$ and $b \in \mathbb{R}^k$ such that

$$C = \{x \in \mathbb{R}^n \mid Ax \leq b\}$$

and is

(ii) a ***convex polytope*** if it is a compact convex polyhedron. •

Thus a convex polyhedron is the intersection of the solutions to a finite number of linear inequalities, i.e., an intersection of a finite number of half-spaces. Let us denote the half-spaces by $H_1, \ldots, H_k$ and the boundary hyperplanes by $P_1, \ldots, P_k$. Thus

$$C = \mathrm{cl}(H_1) \cap \cdots \cap \mathrm{cl}(H_k).$$

We can (and do) assume without loss of generality that the normals of the boundary hyperplanes of the defining half-spaces are not collinear. This amounts to saying that no two rows of the matrix A are collinear. We can also assume, by restricting to the affine hull of C if necessary, that $\mathrm{int}(C) \neq \emptyset$. This simplifies the discussion. The intersection F_j of a convex polyhedron C with the hyperplane P_j, $j \in \{1, \ldots, k\}$, is a ***face of dimension $n-1$***. Thus there are as many faces of dimension $n-1$ as there are rows in the matrix A. Let $F_1, \ldots, F_k$ denote the faces of dimension $n-1$. For fixed $j_1, j_2 \in \{1, \ldots, k\}$ the set $F_{(j_1,j_2)} = C \cap P_{j_1} \cap P_{j_2}$ is a ***face of dimension $n-2$***. Thus we can write the set of faces of dimension $n-2$ as $F_{(j_{11},j_{12})}, \ldots, F_{(j_{l1},j_{l2})}$ for suitable pairs $(j_{11}, j_{12}), \ldots, (j_{l1}, j_{l2}) \in \{1, \ldots, k\}^2$. One can proceed in this way, defining faces of dimensions $n-1, n-2, \ldots, 1, 0$. A face of dimension 0 is called a ***vertex*** and a face of dimension 1 is sometimes called a ***rib***.

Next let us introduce the fundamental problem of linear programming.

B.24 Problem: (Linear programming) The ***linear programming problem*** is: For a convex polyhedron

$$C = \{x \in \mathbb{R}^n \mid Ax \leq b\}$$

and for $c \in \mathbb{R}^n$, minimise the function $x \mapsto \langle c, x \rangle$ over C. A ***solution*** to the linear programming problem is thus a point $x_0 \in C$ such that $\langle c, x_0 \rangle \leq \langle c, x \rangle$ for every $x \in C$. •

The following result describes the solutions to the linear programming problem.

B.25 Theorem: (Solutions to linear programming problem) *Let $c \in \mathbb{R}^n$, let*

$$C = \{x \in \mathbb{R}^n \mid Ax \leq b\}$$

be a convex polyhedron, and consider the linear programming problem for the function $x \mapsto \langle c, x \rangle$. Then the following statements hold:

(i) the linear programming problem has a solution if and only if $x \mapsto \langle c, x \rangle$ is bounded below on C;

(ii) if C is a convex polytope then the linear programming problem has a solution;

(iii) if $x \mapsto \langle c, x \rangle$ is not constant on C then any solution of the linear programming problem lies in $\mathrm{rel\,bd}(C)$.

Proof: (i) Certainly if the linear programming problem has a solution, then $x \mapsto \langle c, x \rangle$ is bounded below on C. So suppose that $x \mapsto \langle c, x \rangle$ is bounded below on C. Specifically, suppose that $\langle c, x \rangle \geq -M$ for some $M \in \mathbb{R}_{>0}$. If $c = 0$ then the linear programming

problem obviously has a solution, indeed many of them. So we suppose that $c \neq 0$. Let $x_0 \in C$ and define

$$A = \{x \in \mathbb{R}^n \mid \langle c, x\rangle \in [-M, \langle c, x_0\rangle]\}.$$

Note that A is nonempty, closed (since $x \mapsto \langle c, x\rangle$ is continuous), and bounded (since linear functions are proper). Thus $A \cap C$ is nonempty and compact. The function $x \mapsto \langle c, x\rangle$, restricted to $A \cap C$, therefore achieves its minimum on $A \cap C$ at some point, say $\bar{x}$. It holds that $\langle c, \bar{x}\rangle \leq \langle c, x\rangle$ for every $x \in C$ since, it clearly holds for $x \in A \cap C$, and if $x \notin A \cap C$ then $\langle c, x\rangle \geq \langle c, x_0\rangle$. Thus the point $\bar{x}$ solves the linear programming problem.

(ii) This follows immediately since $x \mapsto \langle c, x\rangle$ is bounded on C if C is a convex polytope.

(iii) That $x \mapsto \langle c, x\rangle$ is not constant on C is equivalent to $x \mapsto \langle c, x\rangle$ not being constant on $\operatorname{aff}(C)$. This is in turn equivalent to the subspace $U(C)$ of $\mathbb{R}^n$ associated to $\operatorname{aff}(C)$ not being contained in $\ker(c)$. Let $x \in \operatorname{rel\,int}(C)$. Let $u_c \in U(C)$ be the unit vector such that

$$\langle u_c, c\rangle = \inf\{\langle u, c\rangle \mid u \in U(C),\ \|u\| = 1\}.$$

That such a u_c exists since $u \mapsto \langle u, c\rangle$ is a continuous function on the compact set $U(C) \cap \mathbb{S}^{n-1}$. Moreover, since $\langle -u, c\rangle = -\langle u, c\rangle$, it follows that $\langle u_c, c\rangle < 0$. Since $x \in \operatorname{rel\,int}(C)$ there exists $r \in \mathbb{R}_{>0}$ such that $x + ru_c \in C$. Then

$$\langle c, x + ru_c\rangle = \langle c, x\rangle + r\langle c, u_c\rangle > \langle c, x\rangle,$$

showing that $c \mapsto \langle c, x\rangle$ cannot achieve its minimum at $x \in \operatorname{rel\,int}(C)$. Thus it must achieve its minimum on $\operatorname{rel\,bd}(C)$. ■

Exercises

EB.1 Show that $\mathbb{D}^n = \{x \in \mathbb{R}^{n+1} \mid \|x\| \leq 1\}$ is convex.
Hint: *Use the triangle inequality.*

EB.2 If the following statements are true, prove them true. If they are false, give a counterexample to demonstrate this.

(a) The intersection of two convex sets is convex.
(b) The intersection of two cones is a cone.
(c) The union of two convex sets is a convex set.
(d) The union of two cones is a cone.
(e) The intersection of two affine subspaces is an affine subspace.
(f) The union of two affine subspaces is an affine subspace.

EB.3 Show that the image of a convex set (resp. cone) under a linear map is a convex set (resp. cone).

Appendix C

Two topological lemmata

In this appendix we present two topological results which will be useful in our approximations of the reachable set using cones generated by needle variations and in our establishing of the transversality conditions. The results, or at least our proofs of them, rely on the Brouwer Fixed Point Theorem which we first present and prove. Our presentation follows that of Milnor [1978], and so relies first on a rather interesting proof of the so-called Hairy Ball Theorem.

C.1. The Hairy Ball Theorem

First some notation. For $n \in \mathbb{Z}_{>0}$ we denote

$$\mathbb{S}^n = \{x \in \mathbb{R}^{n+1} \mid \|x\| = 1\}, \quad \mathbb{D}^n = \{x \in \mathbb{R}^n \mid \|x\| \leq 1\}.$$

With this notation we have the following preliminary result which is of independent interest.

C.1 Theorem: (Hairy Ball Theorem) *Let $n \in \mathbb{Z}_{>0}$ be even. If $f \colon \mathbb{S}^n \to \mathbb{R}^{n+1}$ is continuous and has the property that $\langle f(x), x\rangle = 0$ for each $x \in \mathbb{S}^n$, then there exists $x_0 \in \mathbb{S}^n$ such that $f(x_0) = 0$.*

Proof: We first prove the result supposing that f is not only continuous but of class C^1.

We use a series of lemmata to prove the theorem in this case.

1 Lemma: *Let $A \subset \mathbb{R}^n$ be compact, let $\mathcal{U}$ be a neighbourhood of A, and let $g \colon \mathcal{U} \to \mathbb{R}^k$ be of class C^1. Then there exists $M \in \mathbb{R}_{>0}$ such that*

$$\|g(y) - g(x)\| \leq M\|y - x\|, \qquad x, y \in A.$$

Proof: Let $B \subset \mathcal{U}$ be an open ball and let $x, y \in B$. Define $\gamma \colon [0,1] \to B$ by $\gamma(t) = (1-t)x + ty$, i.e., γ is the line connecting x and y. Then define $\alpha = g \circ \gamma$. Then

$$\begin{aligned} g(y) - g(x) = \alpha(1) - \alpha(0) &= \int_0^1 \dot{\alpha}(t)\,\mathrm{d}t = \int_0^1 \boldsymbol{D}g(\gamma(t)) \cdot \dot{\gamma}(t)\,\mathrm{d}t \\ &= \int_0^1 \boldsymbol{D}g((1-t)x + ty) \cdot (y - x)\,\mathrm{d}t. \end{aligned}$$

Using the fact that $\boldsymbol{D}g$ is continuous this gives

$$\|g(y) - g(x)\| \leq M_B \|y - x\|, \qquad x, y \in B,$$

where

$$M_B = \sup\{\|\boldsymbol{D}g(x)\| \mid x \in B\}.$$

Now, since A is compact, we can cover it with a finite number of balls $B_1, \ldots, B_N$, each contained in $\mathcal{U}$. Let us denote

$$C = (A \times A) - \cup_{j=1}^{N} B_j \times B_j$$

and note that C is compact. Moreover, the function $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ defined by $d(x, y) = \|x - y\|$ is strictly positive when restricted to C. Therefore, there exists $m \in \mathbb{R}_{>0}$ such that $d(x, y) \geq m$ for all $(x, y) \in C$. Let

$$M_0 = \sup\{\|g(y) - g(x)\| \mid x, y \in A\},$$

noting that this number is finite since g is continuous and A is compact. Now define

$$M = \max\{\tfrac{M_0}{m}, M_{B_1}, \ldots, M_{B_N}\}.$$

Now let $x, y \in A$. If $x, y \in B_j$ for some $j \in \{1, \ldots, N\}$ then

$$\|g(y) - g(x)\| \leq M_{B_j} \|y - x\| \leq M \|y - x\|.$$

If x and y are not together contained in any of the balls $B_1, \ldots, B_N$ then $(x, y) \in C$. Thus

$$\|g(y) - g(x)\| \leq M_0 = \tfrac{M_0}{m} m \leq M \|y - x\|.$$

Thus we have

$$\|g(y) - g(x)\| \leq M \|y - x\|, \qquad x, y \in A,$$

i.e., g is uniformly Lipschitz in A. ▼

2 Lemma: *Let $A \subset \mathbb{R}^{n+1}$ be compact, let $\mathcal{U}$ be a neighbourhood of A, let $g\colon \mathcal{U} \to \mathbb{R}^{n+1}$ be of class C^1, and for $s \in \mathbb{R}$ define $h_s\colon A \to \mathbb{R}^{n+1}$ by $h_s(x) = x + sg(x)$. Then there exists $\epsilon \in \mathbb{R}_{>0}$ such that*

(i) for each $s \in [-\epsilon, \epsilon]$, h_s is injective and

(ii) the function $s \mapsto \text{vol}(h_s(A))$ is a polynomial.

Proof: By Lemma 1 let $M \in \mathbb{R}_{>0}$ be such that

$$\|g(y) - g(x)\| \leq M \|y - x\|, \qquad x, y \in A,$$

and let $\epsilon \in (0, M^{-1})$. Then for $|s| < \epsilon$ we claim that h_s is injective. Indeed, if $h_s(x) = h_s(y)$ then

$$x - y = s(g(y) - g(x)) \quad \Longrightarrow \quad \|y - x\| = |s| \|g(y) - g(x)\| \leq |s| M \|x - y\|.$$

Since $|s|M < 1$ this implies that $x = y$. This gives the first assertion in the lemma.

For the second assertion we observe that $\det \boldsymbol{D}h_s(x)$ is a polynomial function of s, it being the determinant of a matrix whose entries are linear in s. Thus we can write

$$\det \boldsymbol{D}h_s(x) = 1 + a_1(x)s + \cdots + a_{n+1}(x)s^{n+1}$$

for continuous functions $a_1, \ldots, a_{n+1}$. Therefore

$$\operatorname{vol}(h_s(A)) = \int_{h_s(A)} \mathrm{d}x_1 \cdots \mathrm{d}x_{n+1} = \int_A \det \boldsymbol{D}h_s(x)\, \mathrm{d}x_1 \cdots \mathrm{d}x_{n+1},$$

which is clearly a polynomial function in s. ▼

3 Lemma: *Let $f\colon \mathbb{S}^n \to \mathbb{R}^{n+1}$ have the following properties:*

(i) $\langle f(x), x\rangle = 0$ for each $x \in \mathbb{S}^n$;

(ii) $\|f(x)\| = 1$ for each $x \in \mathbb{S}^n$.

Let $\mathcal{U}$ be a neighbourhood of $\mathbb{S}^n$ with $\bar{f}\colon \mathcal{U} \to \mathbb{R}^{n+1}$ a continuously differentiable extension of f, and for $s \in \mathbb{R}$ define $h_s\colon \mathcal{U} \to \mathbb{R}^{n+1}$ by $h_s(x) = x + s\bar{f}(x)$. Then, for $|s|$ sufficiently small, $\bar{f}$ maps $\mathbb{S}^n$ onto the sphere

$$\mathbb{S}^n(\sqrt{1+s^1}) = \{x \in \mathbb{R}^{n+1} \mid \|x\| = \sqrt{1+s^2}\}$$

of radius $\sqrt{1+s^2}$.

Proof: First note that $h_s(\mathbb{S}^n) \subset \mathbb{S}^n(\sqrt{1+s^2})$ for any $s \in \mathbb{R}$ by direct computation. As we saw in the proof of Lemma 2, for s sufficiently small $\boldsymbol{D}h_s(x)$ is nonsingular for each $x \in \mathbb{S}^n$. By the Inverse Function Theorem this means that, for s sufficiently small, h_s is a local diffeomorphism about every point in $\mathbb{S}^n$. Thus $h_s|\mathbb{S}^n$ maps every sufficiently small open set to an open set, provided that s is sufficiently small. This in turn means that $h_s|\mathbb{S}^n$ is an open mapping for s sufficiently small. In particular, $h_s(\mathbb{S}^n)$ is an open subset of $\mathbb{S}^n(\sqrt{1+s^2})$ for s sufficiently small. However, $h_s(\mathbb{S}^n)$ is also compact, the image of compact sets under continuous maps being compact. The only subset of $\mathbb{S}^n(\sqrt{1+s^2})$ that is open and closed is $\mathbb{S}^n(\sqrt{1+s^2})$ since $\mathbb{S}^n(\sqrt{1+s^2})$ is connected. ▼

Now suppose that $f\colon \mathbb{S}^n \to \mathbb{R}^{n+1}$ is such that

1. f is of class C^1,
2. $\langle f(x), x\rangle = 0$ for $x \in \mathbb{S}^n$, and
3. $f(x) \neq 0$ for every $x \in \mathbb{S}^n$.

We may assume without loss of generality (by dividing f by the function $x \mapsto \|f(x)\|$) that $\|f(x)\| = 1$ for each $x \in \mathbb{S}^n$. For $a, b \in \mathbb{R}_{>0}$ satisfying $a < 1 < b$ define

$$A = \{x \in \mathbb{R}^{n+1} \mid a \leq \|x\| \leq b\},$$

and note that A is compact. For the function f as in the theorem statement (but now of class C^1) extend f to A by $f(rx) = rf(x)$, $x \in \mathbb{S}^n$. Then, with $h_s(x) = x + sf(x)$ for $s \in \mathbb{R}$, $h_s(rx) = rh_s(x)$ for $x \in \mathbb{S}^n$. Therefore, h_s maps the sphere of radius $r \in [a, b]$ into

the sphere of radius $r\sqrt{1+s^2}$. By Lemma 3, for s sufficiently small h_s maps the sphere of radius r onto the sphere of radius $r\sqrt{1+s^2}$ for each $r \in [a, b]$. Therefore,

$$\text{vol}(h_s(A)) = (\sqrt{1+s^2})^{n+1}\text{vol}(A)$$

for s sufficiently small. For n even this is not a polynomial in s, and this contradicts Lemma 2. This proves the theorem for f of class C^1.

Finally, we prove the theorem for f continuous. Thus we let $f\colon \mathbb{S}^n \to \mathbb{R}^{n+1}$ have the following properties:

1. f is continuous;
2. $\langle f(x), x\rangle = 0$ for each $x \in \mathbb{S}^n$;
3. $\|f(x)\| \neq 0$ for $x \in \mathbb{S}^n$.

Let

$$m = \inf\{\|f(x)\| \mid x \in \mathbb{S}^n\}$$

and let $p\colon \mathbb{S}^n \to \mathbb{R}^{n+1}$ be a polynomial function such that

$$\sup\{\|p(x) - f(x)\| \mid x \in \mathbb{S}^n\} < \tfrac{m}{2},$$

this being possible by the Weierstrass Approximation Theorem. Now define a continuously differentiable function $g\colon \mathbb{S}^n \to \mathbb{R}^{n+1}$ by

$$g(x) = p(x) - \langle p(x), x\rangle x,$$

and note that $\langle g(x), x\rangle = 0$ by direct computation. We have

$$\begin{aligned} &\langle p(x) - f(x), x\rangle = \langle p(x), x\rangle \\ \implies \quad &|\langle p(x), x\rangle| = |\langle p(x) - f(x), x\rangle| \leq \|p(x) - f(x)\| < \tfrac{m}{2} \end{aligned}$$

for each $x \in \mathbb{S}^n$. This gives

$$\|g(x) - p(x)\| = |\langle p(x), x\rangle| < \tfrac{m}{2},$$

and so

$$|\|g(x)\| - \|f(x)\|| \leq \|g(x) - f(x)\| \leq \|g(x) - p(x)\| + \|p(x) - f(x)\| < m$$

for all $x \in \mathbb{S}^n$. This implies that $\|g(x)\| > 0$ for all $x \in \mathbb{S}^n$, which is in contradiction to what we proved in the first part of the proof since g is continuously differentiable. ■

The intuition of the Hairy Ball Theorem is this. Note that the function f in the theorem statement, by virtue of the fact that $\langle f(x), x\rangle = 0$ for all $x \in \mathbb{S}^n$, can be thought of as assigning a tangent vector to $\mathbb{S}^n$ at each point, i.e., as defining a vector field. The result then says that, when n is even, any such vector field must vanish somewhere. Try to picture this to yourself when $n = 2$. Note that the result requires that n be even, since otherwise the function

$$f(x_1, \ldots, x_{n+1}) = (x_2, -x_1, x_4, -x_3, \ldots, x_n, -x_{n+1})$$

defines a unit vector field that is everywhere tangent to $\mathbb{S}^n$.

C.2. The Brouwer Fixed Point Theorem

Now we state and prove the Brouwer Fixed Point Theorem.

C.2 Theorem: (Brouwer Fixed Point Theorem) *If* $f\colon \mathbb{D}^n \to \mathbb{D}^n$ *is continuous then there exists* $x_0 \in \mathbb{D}^n$ *such that* $f(x_0) = x_0$.

Proof: First let us suppose that n is even. Suppose that $f(x) \neq x$ for every $x \in \mathbb{D}^n$. Then define $g\colon \mathbb{D}^n \to \mathbb{R}^n$ by

$$g(x) = x - f(x)\frac{1 - \langle x, x\rangle}{1 - \langle f(x), x\rangle}.$$

Note that $g(x) = x$ for $\langle x, x\rangle = 1$, and so g points "outward" on $\mathbb{S}^{n-1} = \mathrm{bd}(\mathbb{D}^n)$. Since

$$|\langle f(x), x\rangle| < |f(x)| \leq 1$$

it follows that g is continuous. We also claim that g is nowhere zero. If $\{f(x), x\}$ is linearly independent then $g(x)$ is clearly nonzero. If $\{f(x), x\}$ is linearly dependent then $\langle x, x\rangle f(x) = \langle f(x), x\rangle x$ and so

$$g(x) = \frac{x - f(x)}{1 - \langle f(x), x\rangle} \neq 0.$$

Now consider $\mathbb{S}^n \subset \mathbb{R}^{n+1}$ and denote by

$$\mathbb{S}^n_- = \{x \in \mathbb{S}^n \mid x_{n+1} \leq 0\}, \quad \mathbb{S}^n_+ = \{x \in \mathbb{S}^n \mid x_{n+1} \geq 0\}$$

the southern and northern hemispheres, respectively. We also denote by $E = \mathbb{S}^n_- \cap \mathbb{S}^n_+$ the equator. Now define a map ϕ_- from $\mathbb{D}^n$ to $\mathbb{S}^n_-$ by

$$\phi_-(x) = \frac{(2x_1, \ldots, 2x_n, 1 - \langle x, x\rangle)}{1 + \langle x, x\rangle}.$$

(One may verify that this map is stereographic projection from the north pole, thinking of $\mathbb{D}^n$ as being the disk whose boundary is E.) Now define a vector field h on $\mathbb{S}^n_-$ by

$$h(z) = \boldsymbol{D}\phi_-(\phi_-^{-1}(z)) \cdot g(\phi_-^{-1}(z)).$$

This is a nowhere zero vector field on $\mathbb{S}^n_-$. Moreover, a direct computation shows that for $z \in E$ we have $h(z) = (0, \ldots, 0, 1)$. Define a map ϕ_+ from $\mathbb{D}^n$ to $\mathbb{S}^n_+$ by

$$\phi_+(x) = \frac{(2x_1, \ldots, 2x_n, -1 + \langle x, x\rangle)}{1 + \langle x, x\rangle}.$$

Then define a vector field h on $\mathbb{S}^n_+$ by

$$h(z) = -\boldsymbol{D}\phi_+(\phi_+^{-1}(z)) \cdot g(\phi_+^{-1}(z)).$$

This vector field does not vanish on $\mathbb{S}^n_+$ and a computation gives $h(z) = (0, \ldots, 0, 1)$ for $z \in E$, so h is consistently defined. Moreover, h is continuous and nowhere zero. This contradicts the Hairy Ball Theorem since we are assuming that n is even.

If n is odd, suppose again that $f(x) \neq x$ for every $x \in \mathbb{D}^n$. Then define $F\colon \mathbb{D}^{n+1} \to \mathbb{D}^{n+1}$ by

$$F(z_1, \ldots, z_{n+1}) = (f(z_1, \ldots, z_n), 0),$$

and note that F is continuous and has the property that $F(z) \neq z$ for every $z \in \mathbb{D}^{n+1}$. But we have just showed that this is a contradiction since $n + 1$ is even. ■

C.3. The desired results

Finally, we prove two consequences of the Brouwer Fixed Point Theorem that we shall use in the Sections 5.4 and 6.4 in our approximation of reachable sets with cones of needle variations and in establishing the transversality conditions.

C.3 Lemma: (A property of certain maps from compact convex sets) *Let $K \subset \mathbb{R}^n$ be compact and convex with $\text{int}(K) \neq \emptyset$, and let $f\colon K \to \mathbb{R}^n$ be continuous. If $x_0 \in \text{int}(K)$ has the property that*

$$\|f(x) - x\| < \|x - x_0\|$$

for each $x \in \text{bd}(K)$, then $x_0 \in \text{image}(f)$.

Proof: Without loss of generality suppose that $x_0 = 0$. Define a map $\phi\colon K \to \mathbb{D}^n$ as follows. For $x \in K$ let $\lambda_x \in \mathbb{R}_{>0}$ have the property that $\lambda_x x \in \text{bd}(K)$. Since K is compact such a λ_x always exists and since K is convex with $0 \in \text{int}(K)$ it follows that λ_x is uniquely defined (cf. Lemma 1 in the proof of Proposition B.11). We then define $\phi(x) = \frac{x}{\lambda_x \|x\|}$. We leave it to the reader to verify that ϕ is a homeomorphism.

Now suppose that $f(x) \neq 0$ for every $x \in K$. Define $g\colon \mathbb{D}^n \to \mathbb{R}^n$ by $g(z) = f \circ \phi^{-1}(z)$. For $x \in \text{bd}(K)$ we have

$$\begin{aligned} & \|f(x) - x\| < \|x\|, \\ \implies \quad & 0 \leq \|f(x)\| < 2\langle f(x), x\rangle, \\ \implies \quad & \langle f(x), x\rangle > 0. \end{aligned}$$

This means that $f(x)$ has a strictly positive component in the direction of x for each $x \in \text{bd}(K)$. This then implies that $g(z)$ has a strictly positive component in the direction of z for each $z \in \text{bd}(\mathbb{D}^n)$. We then define $h\colon \mathbb{D}^n \to \mathbb{R}^n$ by

$$h(z) = z - g(z)\frac{1 - \langle z, z\rangle}{1 - \langle g(z), z\rangle},$$

and we verify, just as in our proof of the Brouwer Fixed Point Theorem, that $h(z)$ points strictly outwards on $\text{bd}(\mathbb{D}^n)$. However, as we saw in the proof of the Brouwer Fixed Point Theorem, this leads to a contradiction of the Hairy Ball Theorem. Therefore, we must have $f(x) = 0$ for some $x \in K$. ■

In Figure C.1 we depict the idea behind the lemma. The gist of the matter is that if the boundary does not deform too much under the continuous map f–specifically, it is deformed sufficiently little that the region containing the image of the boundary does not contain x_0–then the image covers x_0.

The next result deals with the intersection of transverse planes under the image of a continuous map.

C.4 Lemma: (Intersections of continuous images of transverse planes) *Let $n, k \in \mathbb{Z}_{>0}$ with $k < n$. Define*

$$\begin{aligned} C^n &= \{(x^1, \dots, x^n) \mid \ \max\{|x^1|, \dots, |x^n|\} \leq 1\}, \\ P_1 &= \{(x^1, \dots, x^n) \in C^n \mid \ x^{k+1} = \cdots = x^n = 0\}, \\ P_2 &= \{(x^1, \dots, x^n) \in C^n \mid \ x^1 = \cdots = x^k = 0\}. \end{aligned}$$

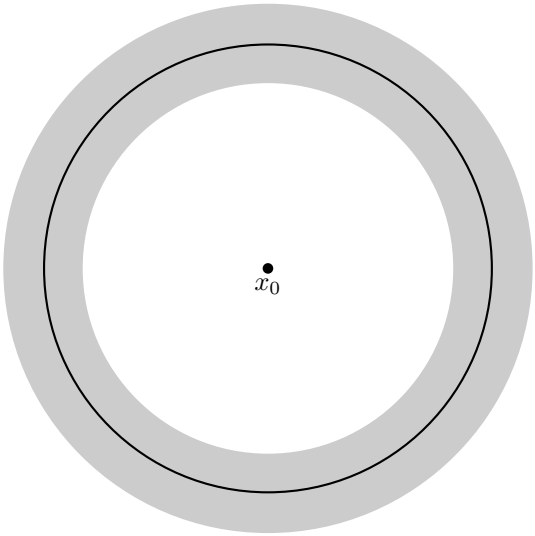

Figure C.1. The idea behind Lemma C.3. The solid circle represents the boundary of C and the shaded region is where the boundary gets mapped to.

Suppose that $f_a\colon P_a \to \mathbb{R}^n$, $a \in \{1,2\}$ *are continuous maps such that*

$$\|f_a(x_a) - x_a\| < \tfrac{1}{4}, \qquad x_a \in P_a,\ a \in \{1,2\}.$$

Then $f_1(P_1) \cap f_2(P_2) \neq \emptyset$.

Proof: Denote a point in C^n by $(x_1, x_2) \in \mathbb{R}^k \times \mathbb{R}^{n-k}$ and define a map $g\colon C^n \to \mathbb{R}^n$ by $g(x_1, x_2) = f_1(x_1) - f_2(-x_2)$. Then, for every $(x_1, x_2) \in C^n$, we have

$$\begin{aligned}\|g(x_1,x_2) - (x_1,x_2)\| &= \|(f_1(x_1) - x_1) - (f(-x_2) - (-x_2))\| \\ &\leq \|f_1(x_1) - x_1\| + \|f(-x_2) - (-x_2)\| < \tfrac{1}{2}.\end{aligned}$$

This implies that for every $(x_1, x_2) \in \mathrm{bd}(C^n)$ we have

$$\|g(x_1,x_2) - (x_1,x_2)\| < \|(x_1,x_2) - (0,0)\|,$$

and so by Lemma C.3 we have $(0,0) \in \mathrm{image}(g)$. Thus there exists $(x_1, x_2) \in C^n$ such that $f(x_1) = f(-x_2)$, and the lemma thus follows. ■

In Figure C.2 we depict the idea behind Lemma C.4. The idea is that, provided the planes P_1 and P_2 do not get deformed too much by f_1 and f_2, respectively, then they will intersect after the maps are applied provided they intersect before in a sufficiently "robust" manner.

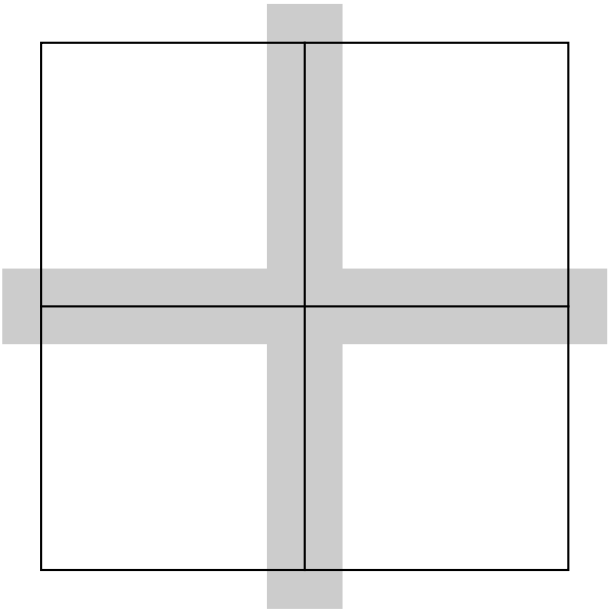

Figure C.2. The idea behind Lemma C.4. The vertical line represents P_1, the horizontal line represents P_2, and the shaded regions represent where these sets get mapped to.

Bibliography

Agrachev, A. A. and Sachkov, Y. [2004] *Control Theory from the Geometric Viewpoint*, volume 87 of *Encyclopedia of Mathematical Sciences*, Springer-Verlag, New York–Heidelberg–Berlin, ISBN 3-540-21019-9.

Berkovitz, L. D. [1974] *Optimal Control Theory*, number 12 in Applied Mathematical Sciences, Springer-Verlag, New York–Heidelberg–Berlin, ISBN 0-387-90106-X.

Bianchini, R. M. and Stefani, G. [1993] *Controllability along a trajectory: A variational approach*, SIAM Journal on Control and Optimization, **31**(4), 900–927.

Bliss, G. A. [1946] *Lectures on the Calculus of Variations*, Phoenix Science Series, University of Chicago Press, Chicago, IL.

Bolza, O. [1961] *Lectures on the Calculus of Variations*, second edition, Chelsea, New York, ISBN 0-8218-2144-X.

Bonnard, B. and Chyba, M. [2003] *Singular Trajectories and their Role in Control Theory*, number 40 in Mathematics & Applications, Springer-Verlag, New York–Heidelberg–Berlin, ISBN 3-540-00838-1.

Brockett, R. W. [1970] *Finite Dimensional Linear Systems*, John Wiley and Sons, New York, New York.

Bullo, F. and Lewis, A. D. [2004] *Geometric Control of Mechanical Systems: Modeling, Analysis, and Design for Simple Mechanical Systems*, number 49 in Texts in Applied Mathematics, Springer-Verlag, New York–Heidelberg–Berlin, ISBN 0-387-22195-6.

Carathéodory, C. [1935] *Variationsrechnung und partielle Differentialgleichungen erster Ordnung*, Teubner, Leipzig, translation: [Carathéodory 1982].

— [1982] *Calculus of Variations*, Chelsea, New York, ISBN 0-8284-0318-X, translation of *Variationsrechnung und partielle Differentialgleichungen erster Ordnung*.

Cohn, D. L. [1980] *Measure Theory*, Birkhäuser, Boston/Basel/Stuttgart, ISBN 0-8176-3003-1.

Dullerud, G. E. and Paganini, F. [1999] *A Course in Robust Control Theory*, number 36 in Texts in Applied Mathematics, Springer-Verlag, New York–Heidelberg–Berlin, ISBN 0-387-98945-5.

Fuller, A. T. [1960] *Relay control systems optimized for various performance criteria*, in *Proceedings of the First IFAC World Congress*, pages 510–519, IFAC, Butterworth & Co., Ltd. London, Moscow.

Gelfand, I. M. and Fomin, S. V. [2000] *Calculus of Variations*, Dover Publications, Inc., New York, ISBN 0-486-41448-5, reprint of 1963 translation from Russian by Richard A. Silverman.

Giaquinta, M. and Hildebrandt, S. [1996] *Calculus of Variations*, number 310 and 311 in Grundlehren der mathematischen Wissenschaften, Springer-Verlag, New York–Heidelberg–Berlin, ISBN 3-540-50625-X and 3-540-57961-3.

Goldstine, H. H. [1980] *A History of the Calculus of Variations From the 17th Through the 19th Century*, number 5 in Studies in the History of Mathematics and Physical Sciences, Springer-Verlag, New York–Heidelberg–Berlin, ISBN 0-387-90521-9.

Halmos, P. R. [1974] *Measure Theory*, number 18 in Graduate Texts in Mathematics, Springer-Verlag, New York–Heidelberg–Berlin, ISBN 0-387-90088-8, reprint of 1950 edition by Litton Educational Publishing, Inc.

Hansen, E. K. [2005] *Coordinate-free derivation of the Euler–Lagrange equations and identification of global solutions via local behavior*, Master's thesis, Queen's University, Kingston, Department of Mathematics & Statistics, Kingston, ON K7L 3N6, Canada.

Hestenes, M. R. [1966] *Calculus of Variations and Optimal Control Theory*, Applied Mathematics Series, John Wiley and Sons, New York, New York.

Jurdjevic, V. [1997] *Geometric Control Theory*, number 51 in Cambridge Studies in Advanced Mathematics, Cambridge University Press, New York/Port Chester/Melbourne/Sydney, ISBN 0-521-49502-4.

Kalman, R. E. [1960] *Contributions to the theory of optimal control*, Boletín de la Sociedad Matemática Mexicana. Segunda Serie, **5**, 102–119.

Krener, A. J. [1977] *The higher order maximum principle and its applications to singular extremals*, SIAM Journal on Control and Optimization, **15**(2), 256–293.

Lanczos, C. [1949] *The Variational Principles of Mechanics*, University of Toronto Press, Toronto, reprint of revised edition: [Lanczos 1986].

— [1986] *The Variational Principles of Mechanics*, Dover Publications, Inc., New York, ISBN 0-486-65067-7, reprint of fourth edition.

Lee, E. B. and Markus, L. [1967] *Foundations of Optimal Control Theory*, John Wiley and Sons, New York, New York.

Lewis, A. D. and Tyner, D. R. [2003] *Jacobian linearisation in a geometric setting*, in *Proceedings of the 42nd IEEE Conference on Decision and Control*, pages 6084–6089, Maui, HI.

Milnor, J. W. [1978] *Analytic proofs of the "Hairy Ball Theorem" and the Brouwer Fixed Point Theorem*, The American Mathematical Monthly, **85**(7), 521–524.

Pontryagin, L. S., Boltyanskii, V. G., Gamkrelidze, R. V., and Mishchenko, E. F. [1961] *Matematicheskaya teoriya optimal′ nykh protsessov*, Gosudarstvennoe izdatelstvo fiziko-matematicheskoi literatury, Moscow, reprint of translation: [Pontryagin, Boltyanskii, Gamkrelidze, and Mishchenko 1986].

— [1986] *The Mathematical Theory of Optimal Processes*, Classics of Soviet Mathematics, Gordon & Breach Science Publishers, New York, ISBN 2-88124-134-4, reprint of 1962 translation from the Russian by K. N. Trirogoff.

Rockafellar, R. T. [1970] *Convex Analysis*, Princeton Mathematical Series, Princeton University Press, Princeton, New Jersey.

Rudin, W. [1976] *Principles of Mathematical Analysis*, third edition, International Series in Pure & Applied Mathematics, McGraw-Hill, New York, ISBN 0-07-054235-X.

Skinner, R. [1983] *First-order equations of motion for classical mechanics*, Journal of Mathematical Physics, **24**(11), 2581–2588.

Skinner, R. and Rusk, R. [1983a] *Generalized Hamiltonian dynamics. I. Formulation on* $T^*Q \oplus TQ$, Journal of Mathematical Physics, **24**(11), 2589–2594.

— [1983b] *Generalized Hamiltonian dynamics. II. Guage transformations*, Journal of Mathematical Physics, **24**(11), 2595–2601.

Sussmann, H. J. [1997] *An introduction to the coordinate-free maximum principle*, in *Geometry of Feedback and Optimal Control*, B. Jakubczyk and W. Respondek, editors, pages 463–557, Dekker Marcel Dekker, New York.

Sussmann, H. J. and Willems, J. C. [1997] *300 years of optimal control: From the brachystochrone to the maximum principle*, IEEE Control Systems Magazine.

Troutman, J. L. [1996] *Variational Calculus and Optimal Control, Optimization with Elementary Convexity*, second edition, Undergraduate Texts in Mathematics, Springer-Verlag, New York–Heidelberg–Berlin, ISBN 0-387-94511-3.

Wonham, W. M. [1985] *Linear Multivariable Control, A Geometric Approach*, third edition, number 10 in Applications of Mathematics, Springer-Verlag, New York–Heidelberg–Berlin, ISBN 0-387-96071-6.

Yan, F. Y. M. [1995] *Introduction to the Calculus of Variations and its Applications*, Chapman & Hall, New York/London, ISBN 0-534-09984-X.

Zhou, K. [1996] *Robust and Optimal Control*, Prentice-Hall, Englewood Cliffs, NJ, ISBN 0-13-456567-3, with John C. Doyle and Keith Glover.