"El saber de mis hijos hará mi grandeza"

UNIVERSIDAD DE SONORA

División de Ciencias Exactas y Naturales

Programa de Posgrado en Matemáticas

Pontryagin Maximum Principle and some applications in Biology

T E S I S

Que para obtener el título de:

Maestro en Ciencias

Presenta:

Nohemy Palafox Lacarra

Directores: Dr. Saúl Díaz Infante Velasco y Dr. David Gonzalez Sanchez.

Hermosillo, Sonora, México, January 17, 2020

SINODALES

Dr. David González Sánchez
CONACYT-Universidad de Sonora, Hermosillo, México

Dr. Francisco Ramón Peñuñuri Anguiano
Universidad Autónoma de Yucatán

Dr. Yofre Hernán Garcia Gómez
Universidad Autónoma de Chiapas

Dr. Saúl Díaz Infante Velasco
CONACYT-Universidad de Sonora, Hermosillo, México

A Saúl, por guiarme y enseñarme tantas cosas en estos años.

"Las Matemáticas no son un recorrido prudente por una autopista despejada, sino un viaje a un terreno salvaje y extraño, en el cual los exploradores se pierden a menudo."

Acknowledgements

A lo largo de este proyecto conocí muchas personas que me han traído hasta este día. Por ello dedico un espacio de mi trabajo a estas.

Primero que nada quiero agradecer a mis dos directores Saúl y David por su gran apoyo, por todo el tiempo que me dedicaron y por todas sus enseñanzas. Todo lo que aprendí de ustedes es inigualable.

Muchas gracias a Yofre, Francisco y Daniel por tomarse el tiempo de revisar este trabajo y aportar sus valiosas ideas.

A mis amigos y familia gracias por todo el amor y motivación que me brindaron durante este proceso. Sobre todo muchas gracias Irasema, Alfonso, Adrián y Mayra por todos los ratos divertidos que vivimos juntos y por hacer de este proceso más ligero.

Gracias Claudio por estar conmigo siempre, por estar ahí cada vez que quería tirar la toalla, por acompañarme incluso los fines de semana a la universidad cuando tenía que trabajar en este proyecto.

Gracias a todos los que creyeron en mi,

Contents

List of Figures

Abstract

The main objective of this thesis is to explore the applications of the deterministic optimal control theory in biology. In particular we are interested in open-loop policies. That is, a policy that only depends on time. Thus, we expose the theory of existence of this kind of policies and a way to obtain the optimal policy by the Pontryagin Maximum Principle.

Chapter 1

Introduction

In this thesis we review a complete and self-contained proof of the Pontryagin's Maximum Principle [22]. To proof this result we apply the so-called Ekeland's ε-Variational Principle [11]. We also reproduce the simulation of some literature examples which follow the optimal control framework of Lenhart [13]. That is, to control a given dynamics with linear terms and approximate the optimally policies with the forward-backward sweep method.

In the literature of optimal control theory applied to biological models, contingent policies such as vaccination, quarantine, isolation, treatment, among others, are naturally described by control terms [1, 5, 6, 12, 13, 20]. A common practice is to describe these contingent policies with linear terms. This linear form simplifies the characterization of optimal policies. In short, a policy is a function that prescribes which actions apply according to information. If the control policy only depends on time, then this policy is of open-loop. In the other way, if it depends on the current state, then it is of closed-loop. Thus, if a policy optimizes a given cost functional —a function from $\mathbb{R}^n$ to $\mathbb{R}$ that describes the resource consumption and the profit generation —then this policy is optimal. For example, consider a vaccination campaign as a control policy and the cost functional describes the balance between the necessary money to run the campaign and the number of infected individuals to be minimized.

First, we have to ensure the existence of an optimal policy. To this end, we appeal to the theorems Arzela-Ascoli and Banach-Sacks, to the Filippov lemma and some other results. Next, we apply the Pontryagin's Maximum Principle to characterize optimal control. That is, we get the necessary conditions to approximate the optimal policy. However, some problems lacks of unique optimal policies, which still is, an open problem.

The aim of this thesis is to review the existence and characterization of the underlying solution to optimal control problems with applications to biology and approximate the concerning optimal policies.

After of this introduction, in Chapter 2 we introduce and prove the Ekeland's ε-Variational Principle, which will give the existence of an approximate control. Chapter 3 presents the necessary theory to ensure the existence of an optimal control. In Chapter 4 we enunciate and prove the Pontryagin's maximum principle. Chapter 5 discusses the forward-backward sweep method, which approximate the optimal policies. Chapter 6 describes the multidimensional control problems with one control and two controls on their dynamics. We closed this work with the conclusion and perspectives in Chapter 7.

Chapter 2

The Ekeland's Variational Principle

The purpose of this Chapter is to enunciate and prove the Ekeland's ε-Variational principle [8]. First section introduces the theory of lower and upper semicontinuous functions. Thus, in second section we enunciate the Ekeland's principle, and prove it following the ideas presented in [11].

2.1 Lower and Upper Semicontinuous Functions

Here we discuss about preliminary results from semicontinuous functions. The objective is to characterize the concept of semicontinuity and assure the existence of global minimum. Those concepts are defined over a metric space (U, d).

Define the set of extended real numbers as $\overline{\mathbb{R}} := \{-\infty\} \cup \mathbb{R} \cup \{+\infty\}$. This set is ordered and we can define the following operations

$$\begin{aligned}
x \in \mathbb{R} \cup \{+\infty\} &\implies x + (+\infty) = +\infty, \\
x \in \mathbb{R} \cup \{-\infty\} &\implies x + (-\infty) = -\infty, \\
x > 0 &\implies x(+\infty) = +\infty, \\
x < 0 &\implies x(+\infty) = -\infty, \\
(-\infty)(+\infty) &= -\infty, \\
(-\infty)(-\infty) = (+\infty)(+\infty) &= +\infty, \\
0(+\infty) = 0(-\infty) &= 0.
\end{aligned}$$

The semicontinuity of a function is defined by the limit inferior and superior of the function. So, we introduce the definition of limit inferior and superior for a given sequence.

Definition 2.1. *Let $\{x_n\}$ be a sequence of extended real numbers, that is $x_n \in \overline{\mathbb{R}}$. The limit inferior of $\{x_n\}$ is*

$$\underline{\lim_{n\to\infty}}\, x_n := \lim_{n\to\infty} \inf_{k\geq n} x_k = \sup_n \inf_{k\geq n} x_k,$$

where the second equality follows since $\{\inf_{k\geq n} x_k\}$ is an increasing sequence in n. Similarly, the limit superior of $\{x_n\}$ is

$$\overline{\lim_{n\to\infty}}\, x_n := \lim_{n\to\infty} \sup_{k\geq n} x_k = \inf_n \sup_{k\geq n} x_k.$$

Now, the limit inferior and superior of a function is defined as follows.

Definition 2.2. *Let $f : U \to \overline{\mathbb{R}}$ be an extended real-valued function. The limit inferior of f as $x \in U$ converges to $x_0 \in E$ is defined by*

$$\underline{\lim_{x\to x_0}}\, f(x) := \lim_{\delta\to 0} \inf_{d(x,x_0)<\delta} f(x) = \sup_{\delta>0} \inf_{d(x,x_0)<\delta} f(x),$$

and its limit superior by

$$\overline{\lim_{x\to x_0}}\, f(x) := \lim_{\delta\to 0} \sup_{d(x,x_0)<\delta} f(x) = \inf_{\delta>0} \sup_{d(x,x_0)<\delta} f(x).$$

Combining the above definition we establish the following lemma.

Lemma 2.1. *Let $f : U \to \overline{\mathbb{R}}$. We have*

$$\underline{\lim_{x\to x_0}}\, f(x) = \inf_{\{x_n\}} \underline{\lim_{n\to\infty}}\, f(x_n),$$

where the infimum on the right-hand side is taken over all sequences $x_n \to x_0$. Similarly,

$$\overline{\lim_{x\to x_0}}\, f(x) = \sup_{\{x_n\}} \overline{\lim_{n\to\infty}}\, f(x_n).$$

Proof. Define

$$M := \underline{\lim}_{x\to x_0} f(x),$$
$$L := \inf_{x_n} \underline{\lim}_{n\to\infty} f(x_n),$$
$$N_\delta := \{x \in U : d(x, x_0) < \delta\}.$$

We have to consider the cases $M = -\infty$, $M = \infty$ and $-\infty < M < \infty$.

Case: $M = -\infty$.

Note that it is enough to prove the existence of a sequence $x_n \to x_0$ such that $f(x_n) \to -\infty$, because

$$\underline{\lim}_{n\to\infty} f(x_n) = \lim_{n\to\infty} \inf\{f(x_n), f(x_{n+1}), \ldots\} = -\infty.$$

Since, $M = -\infty$ we have that

$$M = \underline{\lim}_{x\to x_0} f(x) = \lim_{\delta\to 0} \inf_{x\in N_\delta} f(x) = -\infty.$$

Let $\delta = \frac{1}{n}$, for all $n \in \mathbb{N}$, then

$$\inf_{x\in N_{1/n}} f(x) = -\infty, \forall n \in \mathbb{N}.$$

Thus, there is $y_n \in N_{1/n}$ such that $f(y_n) < -n$. Define the sequence by

$$x_n := y_n, \quad y_n \in N_{1/n} \text{ for each } n \in \mathbb{N}.$$

By the above, there is a sequence $x_n \to x_0$ such that $f(x_n) \to -\infty$. Hence, $M = L$.

Case: $M = \infty$.

Since,

$$\lim_{\delta\to 0} \inf_{x\in N_\delta} f(x) = \infty,$$

for a given $\varepsilon > 0$ there exists $\delta > 0$ such that $\inf_{x\in N_\delta} f(x) > \varepsilon$. By the convergence of x_n, there is $N \in \mathbb{N}$ such that $x_n \in N_\delta$ for all $n \geq N$. Then,

$$f(x_n) \geq \inf_{x_n\in N_\delta} f(x_n) > \varepsilon, \quad \forall n \geq N.$$

That is, $f(x_n) \to \infty$ for any sequence $x_n \to x_0$. Therefore $L = M$.

Case: $-\infty < M < \infty$.

First, we prove that $M \leq L$. By Definition 2.2, given $\varepsilon > 0$ there is $\delta > 0$ such that

$\inf_{x\in N_\delta} f(x) > M-\varepsilon$, implying that $f(x) > M-\varepsilon$ for all $x \in N_\delta$. Let $\{x_n\}$ be a sequence that converges to x_0. Thus, there is $N \in \mathbb{N}$ such that $x_n \in N_\delta$ for all $n \in \mathbb{N}$ and $f(x_n) > M-\varepsilon$ for all $x_n \in N_\delta$, $n \geq N$. Then

$$\underline{\lim}_{n\to\infty} f(x_n) = \lim_{n\to\infty} \inf_{x_n\in N_\delta} f(x_n) \geq M-\varepsilon.$$

The above holds for any sequence $x_n \to x_0$, then $\inf_{\{x_n\}} \underline{\lim}_{n\to\infty} f(x_n) \geq M-\varepsilon$, for any $\varepsilon > 0$. Hence, $M \leq L$.
For the reverse inequality, we have that $\inf_{x\in N_\delta} f(x) \to M$, as $\delta \to 0$. Let $\delta = \dfrac{1}{n}$. By definition of infimum there is $k_\delta \in \mathbb{N}$ such that $x_{k_\delta} \in N_\delta$ and

$$f(x_{k_\delta}) \leq \inf_{x_{k_\delta}\in N_\delta} f(x_{k_\delta}) + \frac{1}{n}.$$

Thus, we choose $x_n \in N_{\frac{1}{n}}$ satisfying the above for each $n \in \mathbb{N}$. Then

$$\begin{aligned} L &= \inf_{x_n} \underline{\lim}_{n\to\infty} f(x_n) \\ &\leq \underline{\lim}_{n\to\infty} f(x_n) \\ &\leq \underline{\lim}_{n\to\infty} \left(\inf_{x_n\in N_{\frac{1}{n}}} f(x_n) + \frac{1}{n} \right) \\ &= M. \end{aligned}$$

Hence $L \leq M$. Finally, notice that $\overline{\lim}_{n\to\infty} y_n = -\underline{\lim}_{n\to\infty}(-y_n)$, and $\sup_{x\in A} f(x) = -\inf_{x\in A}(-f(x))$, implies the second assertion. ∎

To fix ideas, in Figure (2.1) we present the function $f(x) := \sin(1/x)$, where we see that $\underline{\lim}_{x\to 0} f(x) = -1$ and $\overline{\lim}_{x\to 0} f(x) = 1$.

The following definitions describe the lower and upper semicontinuity of a function.

Definition 2.3. *Let $f : U \to \overline{\mathbb{R}}$. The function f is lower semicontinuous (*l. s. c.*) at a point $x_0 \in E$ if*

$$f(x_0) \leq \underline{\lim}_{x\to x_0} f(x).$$

Equivalently, by Lemma (2.1), f is l. s. c. *at x_0 if $f(x_0) \leq \underline{\lim}_{n\to\infty} f(x_n)$, for every sequence $x_n \to x_0$.*

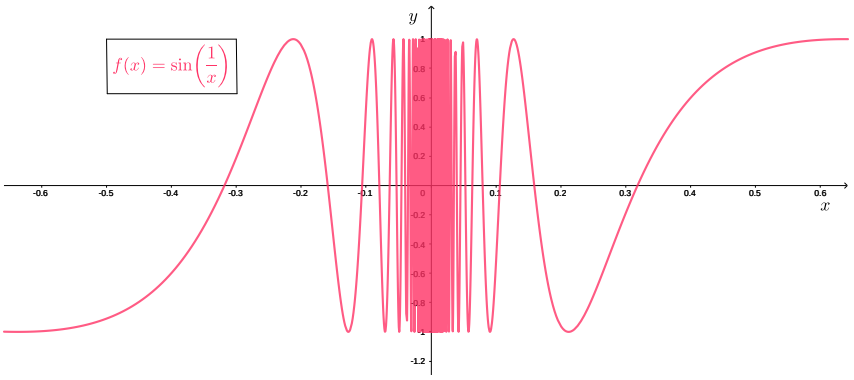


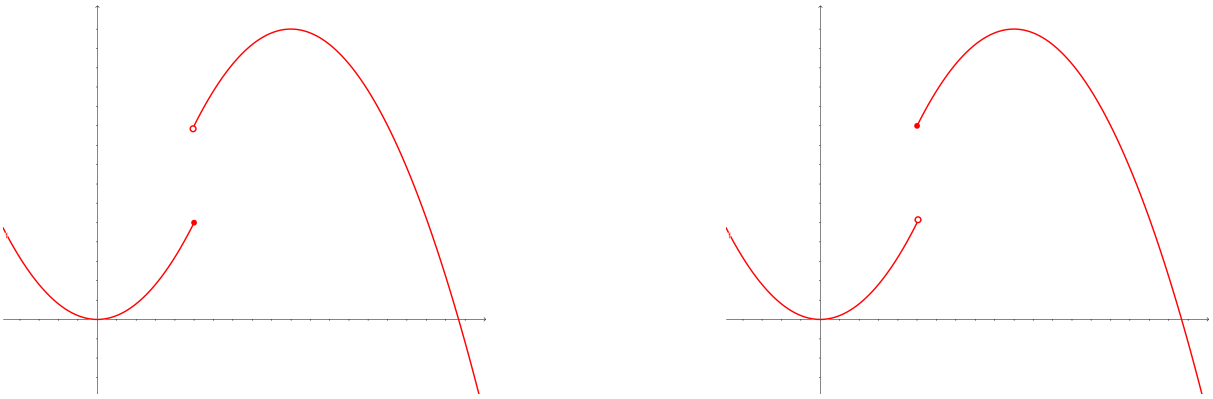
FIGURE 2.1: Example of a semicontinuous function.

Definition 2.4. *The function f is upper semicontinuous* (u. s. c.) *at a point x_0 if*

$$f(x_0) \geq \overline{\lim_{x \to x_0}} f(x).$$

Equivalently, f is u. s. c. *at x_0 if $f(x_0) \geq \overline{\lim}_{n\to\infty} f(x_n)$, for every sequence $x_n \to x_0$.*

Figure 2.2a and Figure 2.2b, presents examples of functions that are l. s. c. and upper semicontinuous, respectively.



(A) Lower (B) Upper

FIGURE 2.2: Semicontinuous Functions

We say that f is l. s. c. or closed on U if f is l. s. c. at every point $x \in U$. Similarly, f is u. s. c. on U if f is u. s. c. at every point in U. Note also that

$$\underline{\lim_{x \to x_0}} f(x) \leq f(x_0),$$

since x_0 lies in every neighborhood N_δ. This implies, f is l. s. c. at x_0 if and only if

$$f(x_0) = \underline{\lim}_{x \to x_0} f(x).$$

Similarly, the function f is u. s. c. at x_0 if and only if

$$f(x_0) = \overline{\lim}_{x \to x_0} f(x).$$

In addition, any function is l. s. c. at a point x if $f(x) = -\infty$ and similarly any function is u. s. c. at a point x if $f(x) = \infty$.

Since the l. s. c. and u. s. c. functions are, respectively, related with its epigraph and hypograph, we enunciate its definitions.

Definition 2.5. *Let $f : U \to \overline{\mathbb{R}}$ be a function. We define the epigraph of f as the set*

$$\operatorname{epi}(f) := \{(x, t) \in U \times \mathbb{R} : f(x) \leq t\}.$$

Similarly, define the hypograph of f as

$$\operatorname{hypo}(f) := \{(x, t) \in U \times \mathbb{R} : f(x) \geq t\}.$$

The following result relates the semicontinuity of a functions, with its epi or hypo.

Theorem 2.1. *Let $f : U \to \overline{\mathbb{R}}$. The following statements are equivalent:*

a) The function f is l. s. c. *on U,*

b) The set $\operatorname{epi}(f)$ *is a closed subset of $U \times \mathbb{R}$,*

c) The sublevel set $l_\alpha(f) := \{x \in U : f(x) \leq \alpha\}$, is closed for all $\alpha \in \mathbb{R}$.

Proof. We have to prove the following implications $a) \Rightarrow b) \Rightarrow c)$.
$a) \Rightarrow b)$ Suppose that f is a l. s. c. function on U. Let (x_n, t_n) be a sequence on $\operatorname{epi}(f)$ converging to a point (x, t). Since f is l. s. c. at x,

$$f(x) \leq \underline{\lim} f(x_n). \tag{2.1}$$

On the other hand, since $(x_n, t_n) \in \operatorname{epi}(f)$

$$f(x_n) \leq t_n. \tag{2.2}$$

for all $n \in \mathbb{N}$. Combining (2.1) and (2.2) we get

$$f(x) \leq \underline{\lim}\, f(x_n) \leq \lim f(x_n) \leq \lim t_n = t.$$

Thus, $(x,t) \in \operatorname{epi}(f)$. Hence $\operatorname{epi}(f)$ is closed.

b) $\Rightarrow$ *c)* Suppose that $\operatorname{epi}(f)$ is closed and let (x_n) be a sequence in $l_\alpha(f)$ converging to a point $x \in U$. Note that $(x_n, \alpha) \in \operatorname{epi}(f)$ and its limit $(x,\alpha) \in \operatorname{epi}(f)$. Hence $x \in l_\alpha(f)$.

c) $\Rightarrow$ *a)* First, we consider $-\infty < f(x) < \infty$. For this case we proceed by contradiction, assuming that the sublevel set $l_\alpha(f)$ is closed for all $\alpha \in \mathbb{R}$ and f is not l. s. c. Then, there is $\varepsilon > 0$ such that

$$\underline{\lim}_{x \to x_0}\, f(x) = \sup_{\delta > 0} \inf_{N_\delta} f(x) = f(x_0) - 2\varepsilon.$$

Thus, for any $\delta > 0$, we have $\inf_{N_\delta} f(x) \leq f(x_0) - 2\varepsilon < f(x_0) - \varepsilon$. Let $\alpha = f(x_0) - \varepsilon$. Define the sequence $\{x_n\}$ as follows $x_n \in N_\delta(x_0)$, $x_n \to x_0$, such that $f(x_n) \leq f(x_0) - \varepsilon = \alpha$. That is, $x_n \in l_\alpha(f)$. Since $l_\alpha(f)$ is closed, $x_0 \in l_\alpha(f)$ and $f(x_0) < f(x_0) - \varepsilon$, which is a contradiction.

Now, if $f(x) = -\infty$ for a point $x \in U$, then f is l. s. c.. Consider the last case, $f(x) = \infty$ for some $x \in U$. Proceeding by contradiction, suppose that f is not l. s. c. at $x_0 \in E$ and $f(x_0) = \infty$, so there is $\alpha \in \mathbb{R}$ such that $\sup_{\delta > 0} \inf_{x \in N_\delta} f(x) = \alpha$. For any $\delta > 0$, $\inf_{x \in N_\delta} f(x) \leq \alpha$. Let $\beta \in \mathbb{R}$ such that $\alpha < \beta$. Thus, there exist a sequence $\{x_n\} \subset N_\delta$ that converges to x_0 and $f(x_n) \leq \alpha < \beta$. Since $l_\alpha(f)$ is closed, we obtain that $x_0 \in l_\alpha(f)$ and $\infty = f(x_0) \leq \beta$, a contradiction. ∎

Following the same ideas, we can prove an equivalent theorem for u. s. c. functions.

Theorem 2.2. *Let $f : U \to \overline{\mathbb{R}}$. The following statements are equivalent:*

a) The function f is u. s. c. *on U,*

b) The set $\operatorname{hypo}(f)$ *is a closed subset of $U \times \mathbb{R}$,*

c) The sublevel set $l^\alpha(f) := \{x \in U : f(x) \geq \alpha\}$ is closed for all $\alpha \in \mathbb{R}$.

We now prove that the family of l. s. c. functions is closed under the sum operation.

Corollary 2.1. *If the functions $f,\ g\ : U \to \mathbb{R} \cup \{+\infty\}$ are* l. s. c., *then so is $f + g$.*

Proof. We have to prove that $\{x \in U : f(x) + g(x) \leq t\}$ is closed. We claim that

$$\{x \in U : f(x) + g(x) > t\} = \bigcup_{\alpha \in \mathbb{R}} \left(\{x \in U : f(x) > t - \alpha\} \cap \{x \in U : g(x) > \alpha\}\right), \quad (2.3)$$

As usual we prove the two inclusions. Firs, let $x \in \{x \in U : f(x) + g(x) > t\}$ then, there is $\varepsilon > 0$ and some $\alpha \in \mathbb{R}$, such that $f(x) + g(x) = t + 2\varepsilon > t$ and $g(x) = \alpha + \varepsilon > \alpha$. Thus

$$f(x) = t + 2\varepsilon - \alpha - \varepsilon = t - \alpha + \varepsilon > t - \alpha.$$

Conversely, let

$$x \in \bigcup_{\alpha \in \mathbb{R}} (\{x \in U : f(x) > t - \alpha\} \cap \{x \in U : g(x) > \alpha\}).$$

Then $f(x) + g(x) > t - \alpha + \alpha = t$. So, the equality is proved. In the right-hand side of the equality (2.3) we have the arbitrary union of open sets, which implies that $\{x \in U : f(x) + g(x) > t\}$ is open. Thus, the complement $\{x \in U : f(x) + g(x) \leq t\}$, is closed. By Theorem (2.3), $f + g$ is l. s. c. ∎

The following results provide sufficient conditions for the existence of a global minimum for l. s. c. functions.

Theorem 2.3. *Let $f : U \to \mathbb{R} \cup \{+\infty\}$ be* l. s. c.*, defined on a metric space U. If f has a nonempty compact sublevel set $l_\alpha(f)$, then f achieves its global minimum on U.*

Proof. Let $\{x_n\}_{n=1}^{\infty}$ be a minimizing sequence for f, that is

$$f(x_n) \searrow \inf\{f(x) : x \in U\}.$$

Define $\inf_E f := \{f(x) : x \in U\}$. Since $f(x_n) \searrow \inf_E f$ there is $N \in \mathbb{N}$ such that $x_n \in l_\alpha(f)$ for all $n \geq N$, that is $f(x_n) \leq \alpha$, for all $n \geq N$. By hypothesis $l_\alpha(f)$ is a compact set, this implies that $\{x_n\}_{n=N}^{\infty}$ has a convergent subsequence

$$x_{n_k} \to x^* \in l_\alpha(f).$$

Since f is l. s. c. at x^*, we have

$$\inf_E f \leq f(x^*) \leq \varliminf_{n\to\infty} f(x_n) = \lim_{n\to\infty} f(x_n) = \inf_E f.$$

Hence, $f(x^*) = \inf_E f$. ∎

Definition 2.6. *A function $f : D \to \mathbb{R}$ on a subset D of a normed vector space U is called coercive if*

$$f(x) \to \infty \text{ as } \|x\| \to +\infty.$$

Corollary 2.2. *Let $f : K \to \mathbb{R}$ be a* l. s. c. *function defined on a metric space K.*

a) If K is compact, or

b) K is a subset of a finite-dimensional normed vector space U and f is coercive,

then f achieves a global minimum on K.

Proof.

a) Note that all sublevel sets are closed because f is l. s. c. Also, we know that all closed subsets of a compact set are compact. Therefore, the Theorem 2.3 implies that f achieves its global minimum on K.

b) Consider the sublevel set $l_\alpha(f) = \{x \in D : f(x) \leq \alpha\}$. Since f is coercive there is $\delta_\alpha > 0$ such that $\|x\| > \delta_\alpha$, which implies $f(x) > \alpha$. Then, for any $y \in l_\alpha(f)$, $\|y\| \leq \delta_\alpha$. Thus, the sublevel set $l_\alpha(f)$ is bounded and also closed because f is l. s. c. Hence $l_\alpha(f)$ is compact and the conclusion follows from Theorem 2.3.

∎

The Ekeland principle [11] is established for proper l. s. c. functions. A function f is proper if $f(x) \leq \infty$ for at least one point $x \in U$.

2.2 Ekeland's Variational Principle

The aim of this section is to present and prove the Ekeland's ε-Variational Principle [8]. We also see, as an application of this principle, a proof of the Banach fixed point theorem.

We start by introducing a partial order. Let (U, d) be a metric space, and let $f : U \to \mathbf{R}$ be any function. Define the relation $\preceq$ on U by the condition

$$y \preceq x \iff f(y) + d(x, y) \leq f(x).$$

This relation is a partial ordering on U, that is, for all x, y, $z \in U$ we have

i) Reflexivity: $x \preceq x$.

ii) Antisymmetry: $x \preceq y$ and $y \preceq x$ imply $x = y$.

iii) Transitivity: $x \preceq$ and $y \preceq z$ imply $x \preceq z$.

Now, we prove these properties

i) Note that $f(x) + d(x, x) = f(x) \leq f(x)$, then $x \preceq x$.

ii) Suppose that $x \preceq y$ and $y \preceq x$ that is

$$f(x) + d(y, x) \leq f(y), \tag{2.4}$$
$$f(y) + d(x, y) \leq f(x). \tag{2.5}$$

If we add (2.4) and (2.5), then we get $f(x) + f(y) + 2d(x, y) \leq f(x) + f(y)$. Thus, $d(x, y) = 0$, which implies $x = y$.

iii) Suppose that $x \preceq y$ and $y \preceq z$, that is

$$f(x) + d(y, x) \leq f(y), \tag{2.6}$$
$$f(y) + d(y, z) \leq f(z). \tag{2.7}$$

Adding (2.6) and (2.7), we get

$$f(x) + f(y) + d(y, x) + d(y, z) \leq f(y) + f(z).$$

Then $f(x) + d(x, y) + d(y, z) \leq f(z)$. Using the triangle inequality $f(x) + d(x, z) \leq f(z)$, implying that $x \preceq z$.

A point $x \in U$ is a d-point if $y \preceq x$ implies that $y = x$, or equivalently,

$$f(x) < f(y) + d(x, y), \quad \forall y \in U,\ y \neq x.$$

Now, define the set

$$S(x) := \{y \in U : y \preceq x\} = \{y \in U : f(y) + d(x, y) \leq f(x)\}.$$

Note that $x \in S(x)$, since $x \preceq x$, so $S(x) \neq \varnothing$. Since $\preceq$ is a partial order, we claim that $y \preceq x$ if and only if $S(y) \subseteq S(x)$. Suppose that $y \preceq x$ and let $z \in S(y)$, then $z \preceq y$ and

$y \preceq x$. Hence $z \preceq x$ and $z \in S(x)$. Now, suppose that $S(y) \subseteq S(x)$, then $y \in S(x)$. Which implies $y \preceq x$.

Note that if f is a l. s. c. function then $S(x)$ is a closed subset of U. Also, a d-point x is characterized by the condition that $S(x)$ is a singleton, that is, $S(x) = \{x\}$.

The following result from [11] is needed to prove the Ekeland principle.

Theorem 2.4. *Let (U, d) be a metric space. The following condition are equivalent:*

a) (U, d) is complete,

b) For any proper l. s. c. *function $f : U \to \mathbb{R} \cup \{+\infty\}$ bounded from below, and any point $x \in U$, there exists a d-point x_0 satisfying $x \preceq x_0$.*

Proof.
b) $\implies$ *a).* Fix $x_0 \in U$. Let $\{x_n\}_{n=1}^{\infty}$ be a Cauchy sequence in U. Consider the proper l. s. c. function $f(x) := 2 \lim_{n\to\infty} d(x, x_n)$. Let us prove the numerical sequence $\{d(x, x_n)\}$ is a Cauchy sequence in $\mathbb{R}$. Since $\{x_n\}$ is a Cauchy sequence, given ε there is $N(\varepsilon) \in \mathbb{N}$ such that for all $m, n \geq N$, $d(x_m, x_n) < \varepsilon$. It follows from

$$d(x, x_n) \leq d(x, x_m) + d(x_m, x_n),$$
$$d(x, x_m) \leq d(x, x_n) + d(x_n, x_m),$$

that $|d(x, x_m) - d(x, x_n)| \leq d(x_m, x_n) \leq \varepsilon$, for all $m, n \geq N$. Then, $\{d(x, x_n)\}$ is a Cauchy sequence, which implies that f is well-defined.

We claim that f is continuous at $x^* \in U$. Let $\varepsilon > 0$, $x^* \in U$ and take $\delta = \varepsilon/2$ such that $d(x, x^*) < \delta$, with $x \in U$. Then, for any $n \in \mathbb{N}$

$$d(x, x_n) \leq d(x, x^*) + d(x^*, x_n), \tag{2.8}$$

and

$$d(x^*, x_n) \leq d(x^*, x) + d(x, x_n). \tag{2.9}$$

Combining (2.8) and (2.9), we obtain

$$|d(x, x_n) - d(x^*, x_n)| \leq d(x, x^*) \leq \frac{\varepsilon}{2}.$$

Letting $n \to \infty$, we have

$$\left|2 \lim_{n\to\infty} d(x, x_n) - 2 \lim_{n\to\infty} d(x^*, x_n)\right| < \varepsilon.$$

Hence, f is continuous at x^*. Finally,

$$\begin{aligned}\lim_{n\to\infty} f(x_n) &= \lim_{n\to\infty} \left[2 \lim_{m\to\infty} d(x_n, x_m)\right] \\ &= 2 \lim_{n\to\infty} \lim_{m\to\infty} d(x_n, x_m) \\ &= 2 \cdot 0 \\ &= 0.\end{aligned}$$

Let $x \in U$ be a d-point of f, by definition $f(x) < f(y) + d(x, y)$, for all $y \in U$. Then $f(x) < f(x_n) + d(x, x_n)$. Letting $n \to \infty$ we have

$$0 \leq \lim_{n\to\infty} d(x, x_n) = f(x) < f(x_n) + d(x, x_n) = \frac{f(x)}{2},$$

then $f(x) = 0$. That is $d(x, x_n) \to 0$ as $n \to \infty$. Hence, U is a complete metric space.
a) $\implies$ *b)*. Suppose that U is complete and let $f : U \to \mathbb{R} \cup \{+\infty\}$ be a proper l. s. c. bounded below. We assume that $f(x_0) < \infty$ for $x_0 \in U$. Generate a sequence $\{x_n\}_{n=1}^{\infty}$ recursively such that given $x_n \in U$, the term $x_{n+1} \in S(x_n)$. We claim that $\{x_n\}$ is a Cauchy sequence. Note that $x_{n+1} \in S(x_n)$, implies $f(x_{n+1}) + d(x_n, x_{n+1}) \leq f(x_n)$. Then, $0 \leq d(x_n, x_{n+1}) \leq f(x_n) - f(x_{n+1})$. Hence, the sequence $\{f(x_n)\}$ is decreasing. Since f is bounded below, the sequence $\{f(x_n)\}$ converges. In particular, $\{f(x_n)\}$ is a Cauchy sequence.

Now, for any $n, m \in \mathbb{N}$ such that $n > m$ we have $d(x_n, x_m) < f(x_m) - f(x_n)$. Letting $m, n \to \infty$ we see that $d(x_n, x_m) \to 0$. That is, $\{x_n\}$ is a Cauchy sequence in M. Since E is complete, $\{x_n\}$ converges to $x_0 \in U$. Moreover, $x_k \in S(x_n)$ for all $k \geq n$. We also have that $S(x_n)$ and

$$S(x) \subseteq \bigcap_{n=1}^{\infty} S(x_n).$$

We prove that $S(x) = \{x\}$. We choose a sequence $\{\widehat{x}_n\}$ such that $\widehat{x}_{n+1} \in S(\widehat{x}_n)$, and $f(\widehat{x}_{n+1}) \leq \inf\{f(y) : y \in S(\widehat{x}_n)\} + \dfrac{1}{n}$. Let $z \in S(x)$, then $z \preceq x \preceq \widehat{x}_{n-1} \preceq \widehat{x}_n$ and

$$f(z) + d(z, \widehat{x}_n) \leq f(\widehat{x}_n) \leq \inf\{f(y) : y \in \widehat{x}_{n-1}\} + \frac{1}{n}.$$

Thus, letting $n \to \infty$, $d(z, \widehat{x}_n) \to 0$. Hence $\widehat{x}_n \to z = x$, that is $S(x) = \{x\}$. ∎

We now enunciate the Ekeland's ε-Variational principle [8] and prove it following the ideas from [11].

Theorem 2.5 ([11, Thm. 3.2], Ekeland's ε-Variational Principle)**.** *Let (U, d) be a complete metric space, and let $f : U \to \overline{\mathbb{R}}$ be a proper* l. s. c. *function that is bounded from below. Then, for every $\varepsilon > 0$, $\lambda > 0$, and $x \in U$ such that*

$$f(x) \leq \inf_E f + \varepsilon,$$

there exists an element $x_\varepsilon \in U$ satisfying the following properties:

(i) $f(x_\varepsilon) \leq f(x)$,

(ii) $d(x_\varepsilon, x) \leq \lambda$,

(iii) $f(x_\varepsilon) < f(z) + \frac{\varepsilon}{\lambda} d(z, x_\varepsilon)$, *for all* $z \in U, z \neq x_\varepsilon$.

Proof. Note that for $\lambda = 1$ and $\varepsilon = 1$ the properties are

(i) $f(x_\varepsilon) \leq f(x)$,

(ii) $d(x_\varepsilon, x) \leq 1$,

(iii) $f(x_\varepsilon) < f(z) + d(z, x_\varepsilon)$, for all $z \in U, z \neq x_\varepsilon$.

Then, it's enough to prove the result for this values because we can replace d by d/λ and f by f/ε in the inequalities $d(x_\varepsilon, x) \leq 1$ and $f(x_\varepsilon) < f(z) + d(z, x_\varepsilon)$ to get the original properties.

Since (U, d) is a complete metric space and f is a proper l. s. c. function bounded below, the conditions in Theorem 2.4 are satisfied. That is, for $x \in U$ there exists a d-point $\bar{x} \preceq x$. We claim that $\bar{x}$ satisfies the required properties

(i) Since $\bar{x} \preceq x$, we have

$$f(\bar{x}) \leq f(\bar{x}) + d(x, \bar{x}) \leq f(x), \tag{2.10}$$

then $f(\bar{x}) \leq f(x)$.

(ii) By hypothesis $f(x) \leq \inf_{x \in M} f(x) + 1 \leq f(\bar{x}) + 1$. Using the relation (2.10), we get

$$f(\bar{x}) + d(x, \bar{x}) \leq f(x) \leq f(\bar{x}) + 1.$$

Then $d(x, \bar{x}) \leq 1$.

(iii) Since $\bar{x}$ is a d-point $f(\bar{x}) < f(x) + d(x, \bar{x})$, for all $x \in U$, with $x \neq \bar{x}$.

■

If we suppose that $U = \mathbb{R}^n$ and $d(x, y) = \|x - y\|$ in Theorem 2.5, then we have a direct proof.

Proof. As above, suppose that $\lambda = 1$. Fix $\varepsilon > 0$ and choose $x \in \mathbb{R}^n$ such that $f(x) \leq \inf_E f + \varepsilon$. Define $g(z) := f(z) + \varepsilon \|z - x\|$. Since f is l. s. c. and $\|\cdot\|$ is continuous, then g is l. s. c.. The function g is coercive because the function f is bounded below and $\|z - x\| \to \infty$ as $\|z\| \to \infty$.

Consider the set of global minima of g, $K := \{m \in \mathbb{R}^n : g(m) \leq g(x), \forall x \in \mathbb{R}^n\}$. The set K is non-empty by Corollary 2.2 part (b) and closed because K is a sublevel set. Let $x_\varepsilon \in K$ be a point that minimizes g on K, that is $g(x_\varepsilon) \leq g(z)$ for all $z \in K$. By definition of g we have $f(x_\varepsilon) + \varepsilon \|x_\varepsilon - x\| \leq f(z) + \varepsilon \|z - x\|$, for all $z \in \mathbb{R}^n$. If $z = x$, then

$$\begin{aligned} f(x_\varepsilon) + \varepsilon \|x_\varepsilon - x\| &\leq f(x) \\ &\leq \inf_{y \in \mathbb{R}^n} f(y) + \varepsilon \\ &\leq f(x_\varepsilon) + \varepsilon. \end{aligned}$$

From the first inequality we have $f(x_\varepsilon) \leq f(x_\varepsilon) + \varepsilon \|x_\varepsilon - x\| \leq f(x)$, so, $f(x_\varepsilon) \leq f(x)$. Further, $f(x_\varepsilon) + \varepsilon \|x_\varepsilon - x\| \leq f(x_\varepsilon) + \varepsilon$, imply $\|x_\varepsilon - x\| \leq 1$. Hence, (i) and (ii) holds for x and x_ε. To prove (iii) note that if $z \in K$ and $z \neq x_\varepsilon$ then $f(x_\varepsilon) \leq f(z)$ for all $z \in K$. Then $f(x_\varepsilon) \leq f(z) < f(z) + \varepsilon \|z - x_\varepsilon\|$, for all $z \in K$, $z \neq x_\varepsilon$. Now, if $z \notin K$, then

$$\begin{aligned} f(x_\varepsilon) + \varepsilon \|x_\varepsilon - x\| &< f(z) + \varepsilon \|z - x\| \\ &\leq f(z) + \varepsilon(\|z - x_\varepsilon\| + \|x_\varepsilon - x\|). \end{aligned}$$

Thus $f(x_\varepsilon) + \varepsilon \|x_\varepsilon - x\| < f(z) + \varepsilon \|z - x_\varepsilon\| + \varepsilon \|x_\varepsilon - x\|$ implies $f(x_\varepsilon) < f(z) + \varepsilon \|z - x_\varepsilon\|$. Therefore property (iii) holds for every $z \in \mathbb{R}^n$ with $z \neq x_\varepsilon$. ■

Taking a particular value $\lambda = \sqrt{\varepsilon}$ we get the following corollary.

Corollary 2.3. *Let the function f and the point x satisfying the conditions in the Theorem 2.5. Then there exists a point x_ε satisfying the following conditions:*

$$\begin{aligned} &f(x_\varepsilon) \leq f(x), \\ &d(x_\varepsilon, x) \leq \sqrt{\varepsilon}, \\ &f(z) > f(x_\varepsilon) - \sqrt{\varepsilon} d(z, x_\varepsilon), \quad \textit{for all } z \in U, z \neq x_\varepsilon. \end{aligned} \tag{2.11}$$

An application of the Ekeland's ε-Variational Principle is the proof of the Banach Fixed Point Theorem.

Definition 2.7. *Let $\varphi : U \to U$ be a mapping . A point $\bar{x} \in U$ is called a fixed point of φ if $\varphi(\bar{x}) = \bar{x}$.*

The mapping φ is called a contractive mapping if there exists $\alpha \in [0, 1)$ such that $d(\varphi(x), \varphi(y)) \leq \alpha d(x, y)$

Theorem 2.6 (Banach Fixed Point Theorem)**.** *A contractive mapping $\varphi : U \to U$ on a complete metric space (U, d) has a unique fixed point.*

Proof. Let φ be a contractive mappping. Define a function $f(x) := d(x, \varphi(x)) \geq 0$, let $\alpha \in [0, 1)$ and choose $\varepsilon \in (0, 1 - \alpha)$. Note that φ is Lipchitz then f is continuous and bounded from below by definition. Thus, all the hypothesis on the Ekeland's principle are satisfied so, there is $\bar{x} \in U$ such that

$$f(\bar{x}) < f(x) + \varepsilon d(x, \bar{x}) \tag{2.12}$$

for all $x \in M$ with $x = \bar{x}$ and each $\varepsilon \in (0, 1 - \alpha)$.

Proceeding by contraction, we suppose that $\varphi(\bar{x}) \neq \bar{x}$. Then there is $x \in U$ such that $\varphi(\bar{x}) = x$, $x \neq \bar{x}$. By (2.12), we have

$$\begin{aligned} d(\varphi(\bar{x}), \bar{x}) = f(\bar{x}) &< f(x) + \varepsilon d(x, \bar{x}) \\ &= d(\varphi(\bar{x}), \varphi(\varphi(\bar{x}))) + \varepsilon d(\varphi(\bar{x}), \bar{x}) \\ &\leq \alpha d(\bar{x}, \varphi(\bar{x})) + \varepsilon d(\varphi(\bar{x}), \bar{x}) \\ &= (\alpha + \varepsilon) d(\varphi(\bar{x}), \bar{x}). \end{aligned}$$

Then,

$$d(\varphi(\bar{x}), \bar{x}) < (\alpha + \varepsilon) d(\varphi(\bar{x}), \bar{x}) \tag{2.13}$$

Note that $\alpha+\varepsilon < \alpha+1-\alpha = 1$, and from (2.13), $1 < \alpha+\varepsilon$, which leads to a contradiction. Hence $\varphi(\bar{x}) = \bar{x}$.

Now, to prove the uniqueness of the fixed point suppose that there are two fixed points x_1 and x_2, then $d(x_1, x_2) = d(\varphi(x_1), \varphi(x_2)) \leq \alpha d(x_1, x_2) < d(x_1, x_2)$, which is a contradiction. ■

In the next chapter we present the conditions needed to have the existence of an optimal policy.

Chapter 3

The Optimal Control Problem and the Existence of Optimal Policies

The aim of this chapter is to establish the optimal control problem and prove the existence of an optimal policy for that control problem. We follow the ideas presented in [22] by Jiongmin Yong.

We define first a control system. Consider a non-empty closed subset $U \subseteq \mathbb{R}^n$ for $0 \leq t_0 < T < \infty$ define the set

$$\mathcal{U}[t_0, T] := \{u : [t_0, T] \to U \ : \ u(\cdot) \text{ is measurable}\}.$$

Given a function $f : \mathbb{R}_+ \times U \times \mathbb{R}^n \to \mathbb{R}^n$, define, for any pair $(t_0, x_0) \in \mathbb{R}_+ \times \mathbb{R}^n$, a control system by the following initial value problem:

$$\begin{aligned} &\dot{X}(s) = f(s, X(s), u(s)), \quad s \in [t_0, T], \\ &X(t_0) = x_0, \end{aligned} \tag{3.1}$$

where $X(\cdot)$ is the state trajectory and $u(\cdot)$ represents the control.

The open-loop control $u(\cdot) \in \mathcal{U}[t_0, T]$ is called a feasible control on $[t_0, T]$. Under appropiate conditions, for any initial pair (t_0, x_0), and feasible control $u(\cdot)$, the system (3.1) admits a unique solution $X(\cdot) = X(\cdot; t, x, u(\cdot))$ defined on $[t_0, T]$. Note that different choices of $u(\cdot)$ will result in different state trajectories $X(\cdot)$. We refer to $(u(\cdot), X(\cdot))$ as a state-control pair of the control system (3.1).

We now consider a set M in $\mathbb{R}^n$, and define the set of all measurable controls u such that the corresponding trajectory belongs to M:

$$\mathcal{U}_{x_0}^M[t_0,T] = \big\{u(\cdot) \in \mathcal{U}[t_0,T] \,:\, X(T;t_0,x_0,u(\cdot)) \in M\big\}.$$

Define the cost functional

$$J(t_0,x_0;u(\cdot)) := \int_t^T g(s,X(s),u(s))ds + h(T,X(T)),$$

for some maps $g : [t,T] \times U \times \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}$. Now we introduce an optimal control problem with a fixed terminal time and a terminal state constraint.

Optimal Control Problem 3.1. *For given $(t_0,x_0) \in \mathbb{R}_+ \times \mathbb{R}^n$ with $\mathcal{U}_{x_0}^M[t_0,T] \neq \varnothing$, find a control $\bar{u}(\cdot) \in \mathcal{U}_{x_0}^M[t_0,T]$ such that*

$$J(t_0,x_0;\bar{u}(\cdot)) = \inf_{u(\cdot)\in\mathcal{U}_{x_0}^M[t_0,T]} J(t_0,x_0;u(\cdot)). \tag{3.2}$$

In Section 3.1 we present some preliminary results. Latter, in Section 3.2 we prove the existence of an optimal policy.

3.1 Auxiliary results

In this section, we present and prove the Filippov lemma and the Banach-Saks theorem. All of these results are needed to prove the existence of an optimal control.

We now fix some notation. For any $0 \le t_0 < T < \infty$ and $1 \le p < \infty$, define

$$C([t_0,T];\mathbb{R}^n) = \{\varphi : [t_0,T] \to \mathbb{R}^n \,:\, \varphi(\cdot) \text{ is continuous }\},$$
$$L^p([t_0,T];\mathbb{R}^n) = \left\{\varphi : [t_0,T] \to \mathbb{R}^n \,:\, \varphi(\cdot) \text{ is measurable }, \int_{t_0}^T |\varphi(s)|^p\,ds < \infty\right\},$$

which are Banach spaces with the following norms, respectively,

$$\|\varphi(\cdot)\|_{C([t_0,T];\mathbb{R}^n)} = \sup_{s\in[t_0,T]} |\varphi(s)|\,, \text{ for every } \varphi(\cdot) \in C([t_0,T];\mathbb{R}^n),$$
$$\|\varphi(\cdot)\|_{L^p([t_0,T];\mathbb{R}^n)} = \left(\int_{t_0}^T |\varphi(s)|^p\,ds\right)^{\frac{1}{p}}, \text{ for every } \varphi(\cdot) \in L^p([t_0,T];\mathbb{R}^n),$$

Now, we prove the Banach-Saks theorem and the Filippov lemma, following the ideas from [22].

Theorem 3.1 (Banach-Saks, [22, Thm.1.4.3.]). *Let $\varphi_k(\cdot) \in L^2([a,b];\mathbb{R}^n)$ such that*

$$\lim_{N\to\infty} \int_a^b \langle \varphi_k(s) - \bar{\varphi}(s), \eta(s)\rangle ds = 0, \quad \forall\, \eta \in L^2([a,b];\mathbb{R}^n),$$

with $\bar{\varphi}(\cdot) \in L^2([a,b];\mathbb{R}^n)$. Then there is a subsequence $\{\varphi_{k_j}(\cdot)\}$ such that

$$\lim_{N\to\infty} \left\| \frac{1}{N}\sum_{j=1}^{N} \varphi_{k_j}(\cdot) - \bar{\varphi}(\cdot) \right\|_{L^2([a,b];\mathbb{R}^n)} = 0.$$

Proof. First we consider that $\bar{\varphi}(\cdot) = 0$, if not we can consider $\varphi_k' = \varphi_k - \bar{\varphi}$. Let $k_1 = 1$. By the hypothesis of $\varphi_k(\cdot)$, we can find $k_1 < k_2 < \cdots < k_N$ such that

$$\left| \int_a^b \langle \varphi_{k_i}(s), \varphi_{k_j}(s)\rangle ds \right| < \frac{1}{N} \quad 1 \le i < j \le N.$$

Then

$$\begin{aligned}
\left\| \frac{1}{N}\sum_{i=1}^{N} \varphi_{k_i}(\cdot) \right\|^2_{L^2(a,b;\mathbb{R}^2)} &= \frac{1}{N^2} \int_a^b \left| \sum_{i=1}^{N} \varphi_{k_i}(s) \right|^2 ds \\
&= \frac{1}{N^2} \int_a^b \sum_{i,j=1}^{N} \langle \varphi_{k_i}(s), \varphi_{k_j}(s)\rangle ds \\
&= \frac{1}{N^2} \sum_{i,j=1}^{N} \int_a^b \langle \varphi_{k_i}(s), \varphi_{k_j}(s)\rangle ds \\
&= \frac{1}{N^2} \sum_{i=1}^{N} \|\varphi_{k_i}(\cdot)\|^2_{L^2([a,b];\mathbb{R}^n)} + \frac{2}{N^2} \sum_{1\le i<j\le N} \int_a^b \langle \varphi_{k_i}(s), \varphi_{k_j}(s)\rangle ds \\
&\le \frac{1}{N} \sup_{i\ge 1} \|\varphi_{k_i}(\cdot)\|^2_{L^2([a,b];\mathbb{R}^n)} + \frac{2}{N^3} \frac{N(N-1)}{2} \\
&\le \frac{1}{N} \sup_{i\ge 1} \|\varphi_{k_i}(\cdot)\|^2_{L^2([a,b];\mathbb{R}^n)} + \frac{1}{N} \to 0,
\end{aligned}$$

when $N \to \infty$. Now, consider that $\bar{\varphi}(\cdot) \neq 0$, thus $\varphi_k(\cdot) - \bar{\varphi}(\cdot) = 0$, and we can apply the previous steps. ∎

Definition 3.1. *Let $\omega : \mathbb{R}_+ \to \mathbb{R}_+$, be a continuous and increasing function such that $\omega(0) = 0$ and let $f : (X,d) \to (Y,\bar{d})$. Then ω is a modulus of continuity of f if $\bar{d}(f(x), f(y)) \le \omega(d(x,y))$.*

Lemma 3.1 (Filippov)**.** *Let (U,d) be a complete separable metric space. Let $g:[t_0,T]\times U\to\mathbb{R}^n$ be a map which is measurable in $t\in[t_0,T]$ and*

$$|g(t,u)-g(t,v)|\le\omega(d(u,v)),\ \forall u,v\in U,\ t\in[t_0,T],$$

where ω is the modulus of continuity of g. Moreover, $0\in g(t,U)$ a.e., $t\in[t_0,T]$. Then there exists a measurable map $u:[t_0,T]\to U$, such that

$$g(t,u(t))=0\quad a.e.\quad t\in[t_0,T]. \tag{3.3}$$

Proof. We consider first the case when $d(u,v)<1$ for all $u,v\in U$. Define

$$\Gamma(t):=\{u\in U\,:\,g(t,u)=0\},\ t\in[t_0,T].$$

Since $0\in g(t,U)$ a.e. $t\in[t_0,T]$, the set $\Gamma(t)\neq\varnothing$. Let $U_0:=\{v_j\,:\,j\ge1\}$ be a countable dense subset of U. We claim that for any $u\in U$ and $0\le c<1$,

$$\underbrace{\{t\in[t_0,T]\,:\,d(u,\Gamma(t))\le c\}}_{:=A}=\underbrace{\bigcap_{i=1}^{\infty}\bigcup_{j=1}^{\infty}\left\{t\in[t_0,T]:\,d(u,v_j)\le c+\frac{1}{i},\,|g(t,v_j)|\le\frac{1}{i}\right\}}_{:=B},$$

where

$$d(u,\Gamma(t)):=\inf_{v\in\Gamma(t)}d(u,v).$$

If $t\in A$ then there is a sequence $u_k\in\Gamma(t)$, i.e. $g(t,u_k)=0$, such that

$$d(u,u_k)\le d(u,\Gamma(t))+\frac{1}{k}\le c+\frac{1}{k}.$$

Since U_0 is a dense subset of U, there exists a sequence $v_{j_k}\in U_0$ such that $d(u_k,v_{j_k})<\frac{1}{k}$. Using the triangle inequality we get

$$d(u,v_{j_k})\le d(u,u_k)+d(u_k,v_{j_k})\le c+\frac{2}{k}. \tag{3.4}$$

By hypothesis, we have

$$|g(t,v_{j_k})|=|g(t,v_{j_k})-g(t,u_k)|\le\omega(d(v_{j_k},u_k))\le\omega\left(\frac{1}{k}\right). \tag{3.5}$$

Hence, using the inequality (3.4) we get

$$\lim_{k\to\infty} d(u, v_{j_k}) \le c,$$

and by (3.5) and the continuity of ω we obtain

$$\lim_{k\to\infty} g(t, v_{j_k}) = 0.$$

Thus, for any $i \ge 1$, there exists $j \ge 1$, such that

$$\begin{aligned} d(u, v_j) &\le c + \frac{1}{i}, \\ |g(t, v_j) - 0| &\le \frac{1}{i}. \end{aligned}$$

Hence, $A \subseteq B$. Conversely, let $t \in B$ for all $i \ge 1$ there is $j \ge i$ such that $d(u, v_j) \le c + \frac{1}{i}$ and $|g(t, v_j)| \le \frac{1}{i}$. Let $v \in \Gamma(t)$, since U_0 is dense there is $v_j \in U_0$ such that $d(v, v_j) < \frac{1}{i}$ for all $j \ge 1$. Then $d(u, \Gamma(t)) \le d(u, v) \le d(u, v_j) + d(v_j, v) < c + \frac{2}{i}$ for all $i \ge 1$. Hence $t \in A$ and $B \subseteq A$.

Note that B is measurable, because g is measurable. Then A is measurable. On the other hand, note that

$$\begin{aligned} \{t \in [t_0, T] \,:\, d(u, \Gamma(t)) \le c\} &= [t_0, T], \quad \forall\, c \ge 1, \\ \{t \in [t_0, T] \,:\, d(u, \Gamma(t)) \le c\} &= \varnothing, \quad \forall\, c < 0. \end{aligned}$$

Note that inverse image of the set $(-\infty, c]$, under the mapping $t \mapsto d(u, \Gamma(t))$, is

$$\begin{cases} \varnothing, & c < 0, \\ A, & 0 \le c < 1, \\ [t_0, T], & c \ge 1. \end{cases}$$

Therefore the mapping $t \mapsto d(u, \Gamma(t))$ is measurable.

We now construct recursively a sequence u_k. For every $t \in [t_0, T]$ define $u_0(t) := v_1 \in U_0$. Note that $u_0(t)$ is measurable and $d(u_0(t), \Gamma(t)) < 1$, for $t \in [t_0, T]$. Next, choose $u_{k-1}(\cdot)$ such that

$$\begin{aligned} d(u_{k-1}(t), \Gamma(t)) &\le 2^{1-k}, \\ d(u_{k-1}(t), u_{k-2}(t)) &\le 2^{2-k}, \end{aligned} \qquad t \in [t_0, T]. \tag{3.6}$$

Define for fixed i and k the sets

$$C_i^k := \{t \in [t_0, T] \,:\, d(v_i, \Gamma(t)) < 2^{-k}\},$$
$$D_i^k := \{t \in [t_0, T] \,:\, d(v_i, u_{k-1}(t)) < 2^{1-k}\}.$$

Since $t \mapsto d(v_i, \Gamma(t))$ is measurable, C_i^k is measurable. Since $u_{k-1}(t)$ is measurable, by construction, and d is continuous, the set D_i^k is measurable. Let $A_i^k = C_i^k \cap D_i^k$, for $k, i \geq 1$. Then A_i^k is also measurable. We prove now that

$$[t_0, T] = \bigcup_{i=1}^{\infty} A_i^k, \quad \forall\, k \geq 1. \tag{3.7}$$

For any $t \in [t_0, T]$, by (3.6), there exists a $u \in \Gamma(t)$ such that

$$d(u_{k-1}(t), u) < 2^{1-k}.$$

By the density of U_0 in U, there exists $i \geq 1$ such that $d(v_i, u) < 2^{-k}$ which implies

$$d(v_i, \Gamma(t)) \leq d(v_i, u) < 2^{-k}.$$

Also,

$$\begin{aligned} d(v_i, u_{k-1}(t)) &\leq d(v_i, u) + d(u, u_{k-1}(t)) \\ &< 2^{-k} + 2^{1-k} \\ &< 2^{1-k} + 2^{1-k} \\ &= 2^{2-k}. \end{aligned}$$

Then

$$\begin{cases} d(v_i, \Gamma(t)) < 2^{-k}, \\ d(v_i, u_{k-1}(t)) < 2^{1-k}, \end{cases}$$

which means $t \in A_i^k$. Note that $A_i^k = C_i^k \cap D_i^k \subseteq [t_0, T]$. Therefore, the relation (3.7) is proved.

Define $u_k(\cdot) : [t_0, T] \to U_0 \subseteq U$ as follows:

$$u_k(t) = v_i, \quad \forall\, t \in A_i^k \setminus \bigcup_{j=1}^{i-1} A_j^k.$$

Since $t \in C_i^k$, we have

$$d(u_k(t), \Gamma(t)) < 2^{-k},$$

and $t \in D_i^k$ implies

$$d(u_k(t), u_{k-1}(t)) < 2^{1-k}.$$

This completes the construction of the sequence $\{u_k(\cdot)\}$. Notice that the relation (3.6) holds for every $k \geq 1$. Which also implies that for each $t \in [t_0, T]$, $\{u_k(t)\}$ is a Cauchy sequence in U. By the completeness of U we obtain

$$\lim_{k\to\infty} u_k(t) = u(t), \quad t \in [t_0, T].$$

Since u_k is measurable, the function $u(\cdot)$ is measurable. Then by the closeness of $\Gamma(t)$, we have

$$u(t) \in \Gamma(t), \quad \forall\, t \in [t_0, T].$$

This means (3.3) holds for almost every $t \in [t_0, T]$.

In the case $d(u, v) \geq 1$, consider an equivalent metric $\bar{d}(u, v) := \frac{d(u,v)}{1+d(u,v)} < 1$ for every $u, v \in U$. ■

3.2 Existence theory for optimal policies

In this section we present some conditions and prove the existence of an optimal policy for the Optimal Control Problem 3.1.

We introduce the following assumption

Condition 1. *The maps $f : \mathbb{R}_+ \times U \times \mathbb{R}^n \to \mathbb{R}^n$ and $g : \mathbb{R}_+ \times U \times \mathbb{R}^n \to \mathbb{R}$ are measurable and there exists a constant $L > 0$ such that*

$$\begin{aligned} &|f(t, u(t), x_1) - f(t, u(t), x_2)| \leq L\,|x_1 - x_2|\,, \quad (t, u(t)) \in \mathbb{R}_+ \times U, \quad \forall\, x_1, x_2 \in \mathbb{R}^n, \\ &|f(t, u(t), 0)| \leq L \quad \forall\, (t, u(t)) \in \mathbb{R}_+ \times U. \end{aligned} \tag{3.8}$$

and

$$|g(t, u(t), x_1) - g(t, u(t), x_2)| \leq L\,|x_1 - x_2|\,, \quad (t, u) \in \mathbb{R}_+ \times U, \quad \forall\, x_1, x_2 \in \mathbb{R}^n, \tag{3.9}$$

Note that combining the inequalities in (3.8) we get

$$|f(t,u,x)| \le L(1+|x|), \quad (t,u,x) \in \mathbb{R}_+ \times U \times \mathbb{R}^n. \tag{3.10}$$

This condition is usually called the Lipschitz condition of the function f. A key feature of the above is that the bond of $|f(t,u,x)|$, depending on $|x|$, is uniform in u.

Proposition 3.1 ([22, Prop.2.1.1, p.37]). *Suppose that Condition 1 hold. Then, for any $(t_0,x_0) \in \mathbb{R}_+ \times \mathbb{R}^n$, and $u(\cdot) \in \mathcal{U}[t_0,T]$ with $T < \infty$, there exits a unique solution $X(\cdot) \equiv X(\cdot;t_0,x_0,u(\cdot))$ to the state equation (3.1). Moreover, the following estimates hold for every $s \in [t_0,T]$ and $u(\cdot) \in \mathcal{U}[t_0,T]$*

$$\begin{aligned} &|X(s;t_0,x_0,u(\cdot))| \le e^{L(s-t_0)}(1+|x_0|) - 1, \\ &|X(s;t_0,x_0,u(\cdot)) - x_0| \le [e^{L(s-t_0)} - 1](1+|x_0|). \end{aligned} \tag{3.11}$$

Further, for any $t \in \mathbb{R}_+$, $x_1, x_2 \in \mathbb{R}^n$, and $u(\cdot) \in \mathcal{U}[t_0,T]$,

$$|X(s;t_0,x_2,u(\cdot)) - X(s;t_0,x_1,u(\cdot))| \le e^{L(s-t_0)}|x_2 - x_1|, \tag{3.12}$$

for every $s \in [t_0,T]$.

Proof. Let $\delta < \frac{1}{L}$, for any $X(\cdot) \in C([t_0,t_0+\delta];\mathbb{R}^n)$ we define

$$[\mathcal{S}X(\cdot)](s) := x_0 + \int_{t_0}^{s} f(r,u(r),X(r))dr, \quad s \in [t_0,t_0+\delta].$$

By Condition 1 we have

$$\begin{aligned} |[\mathcal{S}X_1(\cdot)](s) - [\mathcal{S}X_2(\cdot)](s)| &= \left| \int_{t_0}^{s} f(r,u(r),X_1(r))dr - \int_{t_0}^{s} f(r,u(r),X_2(r))dr \right| \\ &\le L \left| \int_{t_0}^{s} |X_1(r) - X_2(r)|\, dr \right| \\ &\le L \int_{t_0}^{s} \sup_{s\in[t_0,t_0+\delta]} |X_1(s) - X_2(s)|\, dt \\ &= L(s-t_0)\, \|X_1(\cdot) - X_2(\cdot)\|_{C([t_0,t_0+\delta];\mathbb{R}^n)} \\ &\le L(t_0+\delta-t_0)\, \|X_1(\cdot) - X_2(\cdot)\|_{C([t_0,t_0+\delta];\mathbb{R}^n)} \\ &= L\delta\, \|X_1(\cdot) - X_2(\cdot)\|_{C([t_0,t_0+\delta];\mathbb{R}^n)}, \end{aligned}$$

for any $X_1(\cdot), X_2(\cdot) \in C([t_0, t_0+\delta]; \mathbb{R}^n)$ and for all $s \in [t_0, t_0+\delta]$. Hence

$$\|[\mathcal{S}X_1(\cdot)] - [\mathcal{S}X_2(\cdot)]\|_{C([t_0,t_0+\delta];\mathbb{R}^n)} \leq \delta L \, \|X_1(\cdot) - X_2(\cdot)\|_{C([t_0,t_0+\delta];\mathbb{R}^n)} .$$

By the above inequality we have that $\mathcal{S} : C([t_0, t_0+\delta]; \mathbb{R}^n) \to C([t_0, t_0+\delta]; \mathbb{R}^n)$ is contractive. Therefore, the Banach Fixed Point Theorem 2.6 implies that the control system (3.1) admits a unique solution in $[t_0, t_0+\delta]$. We can repeat the same procedure for the set $[t_0+\delta, t_0+2\delta]$ and so on. Then, the system 3.1 admits a unique solution $X(\cdot)$ on $[t_0, T]$.

Now, by the inequality (3.10) we have

$$\begin{aligned} |X(s)| &= \left| x_0 + \int_{t_0}^{s} f(r, u(r), X(r)) dr \right| \\ &\leq |x_0| + L \int_{t_0}^{s} (1 + |X(r)|) dr \end{aligned}$$

for all $s \in [t_0, T]$. Define $\theta(s) := |x_0| + L \int_{t_0}^{s} (1 + |X(r)|) dr$, then by the fundamental theorem of calculus

$$\dot{\theta}(s) = L + L|X(s)| \leq L + L\theta(s),$$

which leads to

$$\theta(s) \leq e^{L(s-t_0)}|x_0| + L \int_{t_0}^{s} e^{L(s-r)} dr = e^{L(s-t_0)}|x_0| + e^{L(s-t_0)} - 1.$$

Thus, $|X(s)| \leq e^{L(s-t_0)}(1 + |x_0|) - 1$ and we obtain the first estimate in (3.11). Next, we apply the first estimate

$$\begin{aligned} |X(s) - x_0| &= \left| \int_{t_0}^{s} f(r, u(r), X(r)) dr \right| \\ &\leq \int_{t_0}^{s} |f(r, u(r), X(r))| dr \\ &\leq L \int_{t_0}^{s} (1 + |X(r)|) dr \\ &\leq L \int_{t_0}^{s} e^{L(r-t_0)} (1 + |x_0|) dr \\ &= (1 + |x_0|)[e^{L(s-t_0)} - 1]. \end{aligned}$$

This prove the second estimate in (3.11). Finally, for $X_1, X_2 \in \mathbb{R}^n$, let us denote $X_i(\cdot) = X(\cdot; t_0, x_i, u(\cdot))$. Then

$$|X_1(s) - X_2(s)| \leq |x_1 - x_2| + L \int_{t_0}^{s} |X_1(r) - X_2(r)| dr.$$

By Gronwall's Inequality A.2, we obtain

$$\begin{aligned} |X_1(s) - X_2(s)| &\leq |x_1 - x_2| \, e^{\int_{t_0}^{s} L dr} \\ &= |x_1 - x_2| \, e^{L(s-t_0)}, \end{aligned}$$

which proves (3.12). ∎

Note that the bounds of the estimations in (3.11) does not depend of the control u, so the estimates are uniform in $u(\cdot) \in \mathcal{U}[t_0, T]$.

Extending the Definition 3.1, in the following condition we consider a modulus of continuity $\omega : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$, which is increasing in each argument, and $\omega(r, 0) = 0$ for every $r \geq 0$.

Condition 2. *The maps $g : \mathbb{R}_+ \times U \times \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}$ are measurable and there exists a local modulus of continuity ω such that*

$$|g(s, u, x_1) - g(s, u, x_2)| + |h(x_1) - h(x_2)| \leq \omega(|x_1| \vee |x_2|, |x_1 - x_2|),$$

for every $(s, u) \in \mathbb{R}_+ \times U, x_1, x_2 \in \mathbb{R}^n$, where $|x_1| \vee |x_2| = \max\{|x_1|, |x_2|\}$, and

$$\sup_{(s,u) \in \mathbb{R}_+ \times U} |g(s, u, 0)| \equiv g_0 < \infty.$$

For any $(t, x) \in [t_0, T] \times \mathbb{R}^n$, we introduce the following set

$$\mathbb{E}(t, x) = \{(z^0, z) \in \mathbb{R} \times \mathbb{R}^n \, : \, z^0 \geq g(t, u, x), \, z = f(t, u, x), \, u \in U\}.$$

To fix the idea see Figure 3.1 and suppose that u is fixed. The set $\mathbb{E}$ is represented by the orange line for each value of u. So the the set $\mathbb{E}$ is the area in the right of the blue curve.

We now introduce the last condition needed to prove the existence of the optimal control.

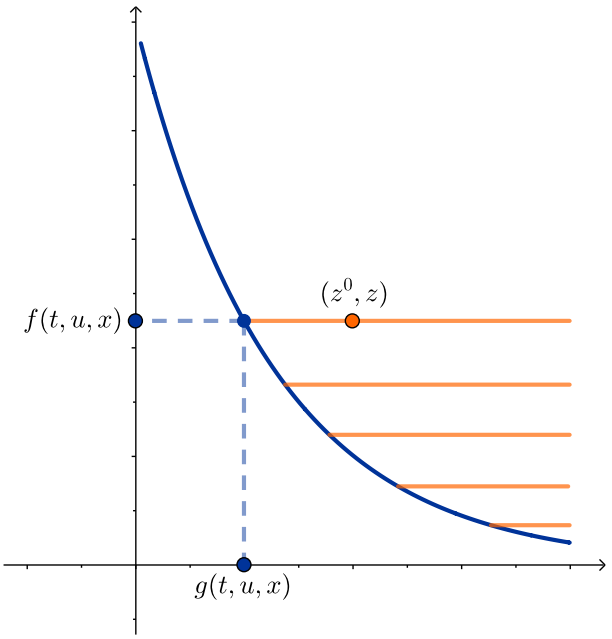


FIGURE 3.1: Convex hull $\mathbb{E}(t,x)$

Definition 3.2. *Consider a set E, then the closed convex hull of the set E is the smallest closed convex set containing E, that is, the intersection of all closed convex sets containing E. We denote the closed convex hull as $\bar{\text{co}}(E)$.*

Condition 3. *For almost all $t \in [t_0, T]$, the following Cesari property holds for any $x \in \mathbb{R}^n$,*

$$\bigcap_{\delta>0} \bar{\text{co}}\big[\mathbb{E}(t, B_\delta(x))\big] = \mathbb{E}(t,x),$$

where $B_\delta(x)$ is the open ball centered at x with radius $\delta > 0$.

Observe that if $\mathbb{E}(t,x)$ has the Cesari property at x, then $\mathbb{E}(t,x)$ is convex and closed.

Theorem 3.2 ([22, Thm.2.2.1, p. 40])**.** *Let Conditions 1-3 hold. Let $M \subseteq \mathbb{R}^n$ be a non-empty closed set. Consider $(t_0, x_0) \in [0,T] \times \mathbb{R}^n$ as $(0, x_0)$ and $\mathcal{U}^M_{x_0}[0,T] \neq \varnothing$. Then, the Optimal Control Problem 3.1 admits at least one optimal pair.*

Proof. Let $u_k(\cdot) \in \mathcal{U}^M_{x_0}[0,T]$ be a minimizing sequence and define

$$x_\tau := X_k(\tau) = X(\tau, 0, x_0, u_k(\cdot)), \quad 0 \leq \tau < s \leq T.$$

By the inequalities in (3.11) we have

$$|X_k(s)| \leq e^{Ls}(1 + |x_0|) - 1, \quad s \in [0,T], \quad k \geq 1, \tag{3.13}$$

and

$$\begin{aligned}|X_k(s) - X_k(\tau)| &= |X(s;\tau,x_\tau,u_k(\cdot)) - X(\tau;0,x_0,u_k(\cdot))| \\ &= |X_k(\tau) - x_\tau| \\ &\leq [e^{L(s-\tau)} - 1][1 + X(\tau;0,x_0,u_k(\cdot))] \\ &\leq [e^{L(s-\tau)} - 1]e^{L\tau}(1 + |x_0|).\end{aligned}$$

Note that inequality (3.13) implies that the sequence $\{X_k(\cdot)\}$ is uniformly bounded. Additionally, let $\varepsilon = [e^{L(s-\tau)} - 1]e^{L\tau}(1 + |x_0|)$, then for all $|s - \tau| < \delta$ and $k \geq 1$ we have that

$$|X_k(s) - X_k(\tau)| < \varepsilon.$$

Hence, the sequence $\{X_k(\cdot)\}$ is equi-continuous. Therefore, by Arzela-Ascoli Theorem A.2, there is a convergent subsequence. To simplify notation we consider $\{X_k(\cdot)\}$ as the subsequence, which is convergent to some $\bar{X}(\cdot) \in C([0,T];\mathbb{R}^n)$. On the other hand, by inequality (3.10)

$$|f(s,u_k(s),X_k(s))| \leq L(1 + |X_k(s)|) \leq Le^{Ls}(1 + |x_0|).$$

By inequality (3.13) and Condition 2, we have for all $s \in [0,T]$, $k \geq 1$

$$\begin{aligned}|g(s,u_k(s),X_k(s))| &\leq |g(s,u_k(s),0)| + |g(s,u_k(s),X_k(s)) - g(s,u_k(s),0)| \\ &\leq g_0 + \omega\big(\,|X_k(s)|\,,|X_k(s)|\,\big) \\ &\leq g_0 + \omega(e^{LT}(1 + |x_0|), e^{LT}(1 + |x_0|)) \\ &\leq K,\end{aligned}$$

where $K \geq 0$ is a generic constant. Hence, by extracting a subsequence if necessary, we may assume that

$$g(\cdot,u_k(\cdot),X_k(\cdot)) \to \bar{g}(\cdot), \quad \text{weakly in } L^2([0,T];\mathbb{R}),$$

and

$$f(\cdot,u_k(\cdot),X_k(\cdot)) \to \bar{f}(\cdot), \quad \text{weakly in } L^2([0,T];\mathbb{R}^n),$$

for some $\bar{g}(\cdot)$ and $\bar{f}(\cdot)$. Then by Banach-Saks Theorem 3.1 we have

$$\begin{aligned}\tilde{g}_k(\cdot) &:= \frac{1}{k}\sum_{i=1}^{k} g(\cdot, u_i(\cdot), X_i(\cdot)) \to \bar{g}(\cdot), \text{ strongly in } L^2([0,T];\mathbb{R}),\\ \tilde{f}_k(\cdot) &:= \frac{1}{k}\sum_{i=1}^{k} f(\cdot, u_i(\cdot), X_i(\cdot)) \to \bar{f}(\cdot), \text{ strongly in } L^2([0,T];\mathbb{R}^n).\end{aligned} \tag{3.14}$$

On the other hand, by Condition 1 we have that $X_k(\cdot) \to \bar{X}(\cdot)$ in $C([0,T];\mathbb{R}^n)$, we have

$$\begin{aligned}\left|\tilde{f}_k(s) - \frac{1}{k}\sum_{i=1}^{k} f(s, u_i(s), \bar{X}(s))\right| &\le \frac{1}{k}\sum_{i=1}^{k}\left|f(s, u_i(s), X_i(s)) - f(s, u_i(s), \bar{X}(s))\right|\\ &\le \frac{L}{k}\sum_{i=1}^{k}\left|X_i(s) - \bar{X}(s)\right|.\end{aligned}$$

We claim that if $X_k(\cdot) \to \bar{X}(\cdot)$ then

$$\frac{L}{k}\sum_{i=1}^{k}\left|X_k(s) - \bar{X}(s)\right| \to 0,$$

uniformly in $s \in [0,T]$ when $k \to \infty$. Let $\varepsilon > 0$ and $0 < \delta < \frac{\varepsilon}{2L}$. Since $X_k(\cdot) \to \bar{X}(\cdot)$ there is N_δ such that for all $k \ge N_\delta$, $\left|X_k(\cdot) - \bar{X}(\cdot)\right|$. Then

$$\begin{aligned}\frac{L}{k}\sum_{i=1}^{k}\left|X_k(s) - \bar{X}(s)\right| &= \frac{L}{k}\sum_{i=1}^{N}\left|X_i(s) - \bar{X}(s)\right| + \frac{L}{k}\sum_{i=N+1}^{k}\left|X_i(s) - \bar{X}(s)\right|\\ &\le \frac{L}{k}\sum_{i=1}^{N}\left\|X_i(\cdot) - \bar{X}(\cdot)\right\| + \frac{k-N-1}{k}L\delta\\ &\le \frac{L}{k}\sum_{i=1}^{N}\left|X_i(s) - \bar{X}(s)\right| + L\delta\\ &\le \frac{L}{k}\sum_{i=1}^{N}\left\|X_i(\cdot) - \bar{X}(\cdot)\right\|_{C([0,T];\mathbb{R}^n)} + \frac{\varepsilon}{2}\\ &= \frac{L}{k}N\left\|X_i(\cdot) - \bar{X}(\cdot)\right\|_{C([0,T];\mathbb{R}^n)} + \frac{\varepsilon}{2}.\end{aligned}$$

Now, let N' such that $\frac{L}{N'}N\left\|X_i(\cdot)-\bar{X}(\cdot)\right\|_{C([0,T];\mathbb{R}^n)}<\frac{\varepsilon}{2}$ and $N'>N_\delta$. Hence, if $k>N'$, then

$$\begin{aligned}\frac{L}{k}\sum_{i=1}^{k}\left|X_k(s)-\bar{X}(s)\right| &\leq \frac{L}{k}N\left\|X_i(\cdot)-\bar{X}(\cdot)\right\|_{C([0,T];\mathbb{R}^n)}+\frac{\varepsilon}{2}\\ &<\frac{\varepsilon}{2}+\frac{\varepsilon}{2}\\ &=\varepsilon.\end{aligned}$$

Similarly, by Condition 2

$$\begin{aligned}\left|\tilde{g}_k(s)-\frac{1}{k}\sum_{i=1}^{k}g(s,u_i(s),\bar{X}(s)\right| &\leq \frac{1}{k}\sum_{i=1}^{k}\left|g(s,u_i(s),X_i(s))-g(s,u_i(s),\bar{X}(s))\right|\\ &\leq \frac{1}{k}\sum_{i=1}^{k}\omega(|X_i(s)|\vee|\bar{X}(s)|,|X_i(s)-\bar{X}(s)|)\to 0,\end{aligned}$$

uniformly in $s\in[0,T]$ when $k\to\infty$. Next, by the definition of $\mathbb{E}(t_0,x_0)$, we have

$$\begin{pmatrix}g(s,u_i(s),X_i(s))\\ f(s,u_i(s),X_i(s))\end{pmatrix}\in\mathbb{E}(s,X_i(s)),\quad i\geq 1,\ s\in[0,T].$$

Hence, for any $\delta>0$, there exits a $K_\delta>0$ such that

$$\begin{pmatrix}\tilde{g}_k(s)\\ \tilde{f}_k(s)\end{pmatrix}\in\bar{co}\mathbb{E}(s,B_\delta(\bar{X}(s))),\ K\geq K_\delta,\ s\in[0,T]. \tag{3.15}$$

Combining (3.14) and (3.15), using (C3), we obtain

$$\begin{pmatrix}\bar{g}(s)\\ \bar{f}(s)\end{pmatrix}=\lim_{k\to\infty}\begin{pmatrix}\tilde{g}_k(s)\\ \tilde{f}_k(s)\end{pmatrix}\in\bigcap_{\delta>0}\bar{co}\mathbb{E}(s,B_\delta(\bar{X}(s)))=\mathbb{E}(s,\bar{X}(s)).$$

Note that the function

$$F(s,u):=\begin{pmatrix}g(s,u(s),\bar{X}(s))-\bar{g}(s)\\ f(s,u(s),\bar{X}(s))-\bar{f}(s)\end{pmatrix}$$

is measurable. Then, by Filippov Lemma 3.1, there exits a $\bar{u}(\cdot)\in\mathcal{U}[0,T]$ such that

$$\begin{aligned}\bar{g}(s)&=g(s,\bar{u}(s),\bar{X}(s)),\\ \bar{f}(s)&=f(s,\bar{u}(s),\bar{X}(s)),\end{aligned}$$

for all $s \in [0,T]$. This means $\bar{X}(\cdot) = X(\cdot; t_0, x_0, \bar{u})$. On the other hand, since

$$\bar{X}_k(T) \equiv X(T; t_0, x_0, \bar{u}_k(\cdot)) \in M, \quad k \geq 1,$$

one has

$$\bar{X}(T) \equiv X(T; t_0, x_0, \bar{u}(\cdot)) \in M,$$

which means that $\bar{u}(\cdot) \in \mathcal{U}_x^M[0,T]$. Finally, by Fatou's Lemma [18, p. 86, thm. 9]

$$\begin{aligned}
J(\bar{u}(\cdot)) &\leq \int_0^T \bar{g}(s)ds + h(\bar{X}(T)) \\
&\leq \varlimsup_{k\to\infty} \left[\int_0^T \tilde{g}_k(s)ds + h(X_k(T))\right] \\
&= \varlimsup_{k\to\infty} \left[\int_0^T \frac{1}{k}\sum_{i=1}^k g(s, u_i(s), X_i(s))ds + \frac{1}{k}\sum_{i=1}^k h(X_k(T))\right] \\
&= \varlimsup_{k\to\infty} \frac{1}{k}\sum_{i=1}^k \left[\int_0^T g(s, u_i(s), X_i(s))ds + h(X_k(T))\right] \\
&= \varlimsup_{k\to\infty} \frac{1}{k}\sum_{i=1}^k J(u_i(\cdot)) \\
&= \lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^k J(u_i(\cdot)) \\
&= \lim_{k\to\infty} J(u_k(\cdot)) \\
&= \inf_{u(\cdot)\in\mathcal{U}_x^M[0,T]} J(u(\cdot)).
\end{aligned}$$

This means that $(\bar{u}(\cdot), \bar{X}(\cdot))$ is an optimal pair. ∎

Chapter 4

Pontryagin's Maximum Principle

In this chapter we define the Ekeland distance presented by Ivar Ekeland in 1974 [8]. We also prove previous results needed to prove the Pontryagin's Maximum Principle, introduced by Lev Pontryagin and coworkers in 1956.

Consider the control system defined in (3.1)

$$\begin{aligned} &\dot{X}(s) = f(s, X(s), u(s)), \quad s \in [t_0, T], \\ &X(t_0) = x_0, \end{aligned}$$

with terminal state constraint $X(T, t_0, x_0, u(\cdot)) \in M \subset \mathbb{R}^n$ and the cost functional

$$J(u(\cdot)) = J(t_0, x_0; u(\cdot)) = \int_{t_0}^{T} g(s, X(s), u(s))ds + h(T, X(T)).$$

Recall the sets

$$\mathcal{U}[t_0, T] := \{u : [t_0, T] \to U \ : \ u(\cdot) \text{ is measurable}\},$$

and

$$\mathcal{U}_{x_0}^{M}[t_0, T] = \big\{u(\cdot) \in \mathcal{U}[t_0, T] \ : \ X(T; t_0, x_0, u(\cdot)) \in M\big\}.$$

Finally, we recall the Optimal Control Problem 3.1

Optimal Control Problem. *For a given pair* $(t_0, x_0) \in \mathbb{R}_+ \times \mathbb{R}^n$ *with* $\mathcal{U}_{x_0}^{M}[t_0, T] \neq \varnothing$, *find a control* $\bar{u}(\cdot) \in \mathcal{U}_{x_0}^{M}[t_0, T]$ *such that*

$$J(\bar{u}(\cdot)) = \inf_{u(\cdot) \in \mathcal{U}_{x_0}^{M}[t_0, T]} J(u(\cdot)).$$

4.1 The Ekeland Distance

Consider the set of all measurable controls $\mathcal{U}[t_0, T]$ and the Lebesgue measure λ, then the Ekeland distance is defined by

$$\rho(u(\cdot), v(\cdot)) := \lambda\big(\{s \in [t_0, T] : u(s) \neq v(s)\}\big), \quad u(\cdot), v(\cdot) \in \mathcal{U}[t_0, T].$$

We claim that $\mathcal{U}[t_0, T]$ is a metric space, under the Ekeland distance. By the definition of the Lebesgue measure, we have that $\rho(u(\cdot), v(\cdot)) \geq 0$, for all $u(\cdot), v(\cdot) \in \mathcal{U}[t_0, T]$. Note that

$$\rho(u(\cdot), v(\cdot)) = \lambda\big(\{s \in [t_0, T] \, : \, u(s) \neq v(s)\}\big) = 0,$$

if and only if $u(s) = v(s) \quad \lambda$-a.e. $s \in [t_0, T]$. Moreover, by the definition of the Ekeland distance $\rho(u(\cdot), v(\cdot)) = \rho(v(\cdot), u(\cdot))$ for all $u, v \in \mathcal{U}[t_0, T]$. Finally, we claim that

$$\rho(u(\cdot), v(\cdot)) \leq \rho(u(\cdot), w(\cdot)) + \rho(w(\cdot), v(\cdot)), \quad \forall\, u, v, w \in \mathcal{U}[t_0, T].$$

Note that we have the next inclusion

$$\{s \in [t_0, T] : u(s) = w(s)\} \cap \{s \in [t_0, T] : w(s) = v(s)\} \subseteq \{s \in [t_0, T] : u(s) = v(s)\},$$

by the complement, we get

$$\{s \in [t_0, T] : u(s) \neq v(s)\} \subseteq \{s \in [t_0, T] : u(s) \neq w(s)\} \cup \{s \in [t_0, T] : w(s) \neq v(s)\}.$$

Then

$$\begin{aligned}
\rho(u(\cdot), v(\cdot)) &\leq \lambda\big(\{s \in [t_0, T] : u(s) \neq w(s)\} \cup \{s \in [t_0, T] : w(s) \neq v(s)\}\big) \\
&\leq \lambda\big(\{s \in [t_0, T] : u(s) \neq w(s)\}\big) + \lambda\big(\{s \in [t_0, T] : w(s) \neq v(s)\}\big) \\
&\leq \rho\big(u(\cdot), w(\cdot)\big) + \rho\big(w(\cdot), v(\cdot)\big).
\end{aligned}$$

Hence $(\mathcal{U}[t_0, T], \rho)$ is a metric space.

Lemma 4.1 ([8], Lemma 7.2)**.** *According to the Ekeland distance, $(\mathcal{U}[t_0, T], \rho)$ is a complete metric space.*

Proof. To prove the completeness of $(\mathcal{U}[t_0, T], \rho)$, we use the usual method, that is, take a Cauchy sequence in $\mathcal{U}[t_0, T]$, and prove that a subsequence of it converges.

Let $\{u_n\}_{n=1}^{\infty} \subseteq \mathcal{U}[t_0, T]$ be a Cauchy sequence. Take a subsequence $\{u_{n_k}\}_{k=1}^{\infty}$ such that

$$\rho(u_{n_k}(\cdot), u_{n_{k+1}}(\cdot)) < \frac{1}{2^{k+1}}(T - t_0)$$

Now, we prove that $\{u_{n_k}(s)\}$ converges in $\mathcal{U}[t_0, T]$. Define the set

$$A_k := \bigcup_{p \geq k} \left\{ s \in [t_0, T] : u_{n_p}(s) \neq u_{n_{p+1}}(s) \right\}, \quad \forall k \in \mathbb{N}.$$

Note that $A_{k+1} \subseteq A_k$ and ${A_k}^{\mathsf{c}} \subseteq {A_{k+1}}^{\mathsf{c}}$ for each $k \in \mathbb{N}$. Now, for $k \in \mathbb{N}$

$$\begin{aligned}
\lambda(A_k) &= \lambda\left(\bigcup_{p \geq k} \{s \in [t_0, T] : u_{n_p}(s) \neq u_{n_{p+1}}(s)\}\right) \\
&\leq \sum_{p \geq k} \lambda(\{s \in [t_0, T] : u_{n_p}(s) \neq u_{n_{p+1}}(s)\}) \\
&= \sum_{p \geq k} \rho(u_{n_p}(\cdot), u_{n_{p+1}}(\cdot)) \\
&\leq \sum_{p \geq k} \frac{T - t_0}{2^{p+1}} \\
&= (T - t_0) \sum_{p=k+1}^{\infty} \frac{T - t_0}{2^p} \\
&= (T - t_0) \left[\sum_{p=0}^{\infty} \frac{1}{2^p} - \sum_{p=0}^{k} \frac{1}{2^p}\right] \\
&= (T - t_0) \left[\frac{1}{1 - \frac{1}{2}} - \frac{\frac{1}{2^{k+1}}}{1 - \frac{1}{2}}\right] \\
&= \frac{T - t_0}{2^k}.
\end{aligned}$$

Define for $s \in [t_0, T]$, the function $\bar{u}(s)$ as follows

$$\bar{u}(s) := \begin{cases} u_{n_1}(s), & s \in {A_1}^{\mathsf{c}} \\ u_{n_2}(s), & s \in {A_2}^{\mathsf{c}} \\ \vdots \qquad\qquad \vdots & \\ u_{n_k}(s), & s \in {A_k}^{\mathsf{c}} \\ \vdots \qquad\qquad \vdots & \end{cases},$$

Note that for each $s \in A_1{}^{\mathsf{c}}$

$$\bar{u}(s) = u_{n_1}(s) = u_{n_l}(s), \quad l \geq 1,\ s \in A_1{}^{\mathsf{c}},$$

and

$$\lim_{l\to\infty} u_{n_l}(s) = u_{n_1}(s) = \bar{u}(s).$$

An equivalent way to write the function below is

$$\bar{u}(s) = \sum_{k=1}^{\infty} u_{n_k}(s)\mathbb{1}_{A_k{}^{\mathsf{c}}}(s),$$

then $\bar{u}$ is measurable. Therefore, by construction, $u_{n_k}(s) \to \bar{u}(s)$ when $k \to \infty$. ∎

Proposition 4.1. *Let Condition 1 hold. The mapping $u(\cdot) \mapsto X(\cdot; t_0, x_0, u(\cdot))$ from $(\mathcal{U}[t_0, T], \rho)$ into $C([t_0, T]; \mathbb{R}^n)$ is uniformly continuous.*

Proof. Let $u, v \in \mathcal{U}[t_0, T]$. From the Control System (3.1)

$$X_u(s) = X(s, t_0, x_0, u(s)) = x_0 + \int_{t_0}^{s} f(r, u(r), X_u(r))dr,$$

and

$$X_v(s) = X(s, t_0, x_0, v(s)) = x_0 + \int_{t_0}^{s} f(r, v(r), X_v(r))dr.$$

Then

$$\begin{aligned}
|X_u(s) - X_v(s)| &\leq \int_{t_0}^{s} |f(r, u(r), X_u(r)) - f(r, v(r), X_v(r))|\, dr \\
&= \int_{A} |f(r, u(r), X_u(r)) - f(r, v(r), X_v(r))|\, dr \\
&+ \int_{A^{\mathsf{c}}} |f(r, u(r), X_u(r)) - f(r, v(r), X_v(r))|\, dr,
\end{aligned}$$

where $A = \{r \in [t_0, s] : u(r) \neq v(r)\}$. Applying Condition 1 and Proposition 3.1 we get

$$\begin{aligned}
|X_u(s) - X_v(s)| &\leq \int_{A} L(1 + |X_u(r)|)dr + \int_{A} L(1 + |X_v(r)|)dr \\
&+ \int_{A^{\mathsf{c}}} L\,|X_u(r)) - X_v(r)|\, dr \\
&\leq 2Le^{L(T-t_0)}(1 + x_0)\lambda(A) + \int_{t_0}^{T} L\,|X_u(r)) - X_v(r)|\, dr.
\end{aligned}$$

Let $K_1 = 2Le^{L(T-t_0)}(1+x_0)$. Applying the Gronwall inequality (A.2) we obtain

$$\begin{aligned}|X_u(s) - X_v(s)| &\leq K_1\rho(u,v) + \int_{t_0}^{T} L\,|X_u(r)) - X_v(r)|\,dr \\ &\leq K_1\rho(u,v)e^{L(T-t_0)}.\end{aligned}$$

Take $K = K_1 e^{L(T-t_0)}$, then

$$|X_u(s) - X_v(s)| \leq K\rho(u,v). \tag{4.1}$$

Given $\varepsilon > 0$, take $\delta > \frac{\varepsilon}{K}$. If $\rho(u,v) \leq \delta$, then, by (4.1)

$$|X_u(s) - X_v(s)| < \varepsilon, \quad \forall\ s \in [t_0, T].$$

Hence, $\|X_u(s) - X_v(s)\| < \varepsilon$. Therefore the mapping is uniformly continuous. ∎

Theorem 4.1. *Let Condition 1 hold. The mapping*

$$u \mapsto \int_{t_0}^{T} g(r, u(r), X_u(r))dr$$

is continuous.

Proof. Let $u, v \in \mathcal{U}[t_0, T]$. Note that

$$\begin{aligned}\int_{t_0}^{T} |g(r,u(r),X_u(r)) - g(r,v(r),X_v(r))|\,dr &= \int_{A} |g(r,u(r),X_u(r)) - g(r,v(r),X_v(r))|\,dr \\ &+ \int_{A^c} |g(r,u(r),X_u(r)) - g(r,v(r),X_v(r))|\,dr,\end{aligned}$$

where $A = \{r \in [t_0, T] : u(r) \neq v(r)\}$. Then, by Condition 1 and Proposition 3.1 we get

$$\begin{aligned}
&\left| \int_{t_0}^{T} [g(r, u(r), X_u(r)) - g(r, v(r), X_v(r))] dr \right| \\
&\qquad \leq \int_A |g(r, u(r), X_u(r)) - g(r, v(r), X_v(r))| \, dr \\
&\qquad + \int_{A^c} |g(r, u(r), X_u(r)) - g(r, v(r), X_v(r))| \, dr, \\
&\qquad \leq \int_A L(1 + |X_u(r)|) dr + \int_A L(1 + |X_v(r)|) dr + \int_{A^c} |X_u(r) - X_v(r)| \, dr \\
&\qquad \leq 2Le^{L(T-t_0)}(1 + |x_0|)\lambda(A) + \int_{t_0}^{T} |X_u(r) - X_v(r)| \, dr \\
&\qquad \leq K\rho(u, v) + \int_{t_0}^{T} |X_u(r) - X_v(r)| \, dr.
\end{aligned}$$

Thus, by Proposition 4.1

$$\begin{aligned}
\left| \int_{t_0}^{T} [g(r, u(r), X_u(r)) - g(r, v(r), X_v(r))] dr \right| &\leq K\rho(u, v) + K\rho(u, v) \\
&= 2\rho(u, v).
\end{aligned}$$

Given $\varepsilon > 0$, take $\delta > \frac{\varepsilon}{2K}$. If $\rho(u, v) < \delta$, then

$$\left| \int_{t_0}^{T} [g(r, u(r), X_u(r)) - g(r, v(r), X_v(r))] dr \right| < \varepsilon, \quad \forall\, r \in [t_0, T].$$

■

We now prove the spike variation lemma and some results needed in the proof of the Pontryagin principle.

Lemma 4.2 (see [22], Lemma 1.4.6.)**.** *Suppose $f(\cdot) \in L^1([0, T]; \mathbb{R}^n)$ and for $0 < \delta < 1$, let*

$$\mathbb{E}_\delta = \{E \subseteq [0, T] \, : \, \lambda(E) = \delta T\},$$

where λ is the Lebesgue measure. Define $g : [0, T] \to \mathbb{R}^n$ as

$$g_E(t) = \int_0^t \left(1 - \frac{1}{\delta} \mathbb{1}_E(s)\right) f(s) ds.$$

Then

$$\inf_{E \in \mathbb{E}_\delta} \|g_E(t)\|_{C([0,T];\mathbb{R}^n)} = 0.$$

Proof. By Proposition A.1, for any $\varepsilon > 0$ there exists an $f_\varepsilon \in C([0,T];\mathbb{R}^n)$ such that

$$\int_0^T |f(r) - f_\varepsilon(r)|\, dr < \varepsilon. \tag{4.2}$$

Since f_ε is finite, we can find a partition $0 = t_0 < t_1 < \cdots < t_{k-1} < t_k = T$ of $[0,T]$ such that

$$\|f_\varepsilon(\cdot)\|_{C([0,T];\mathbb{R}^n)} \max_{1\le i\le k}(t_i - t_{i-1}) < \frac{\varepsilon}{k}. \tag{4.3}$$

Define the step function $\bar{f}_\varepsilon(\cdot)$ as

$$\bar{f}_\varepsilon(r) = \sum_{i=1}^{k} f_\varepsilon(t_i)\mathbb{1}_{(t_{i-1},t_i]}(r), \quad r \in [0,T]. \tag{4.4}$$

First, we prove that

$$\int_0^T \left|f_\varepsilon(r) - \bar{f}_\varepsilon(r)\right| dr < \varepsilon. \tag{4.5}$$

Using eq. (4.4) we get

$$\begin{aligned}
\int_0^T \left|f_\varepsilon(r) - \bar{f}_\varepsilon(r)\right| dr &= \int_0^T \left|f_\varepsilon(r) - \sum_{i=1}^{k} f_\varepsilon(t_i)\mathbb{1}_{(t_{i-1},t_i]}(r)\right| dr \\
&= \int_0^{t_1} \left|f_\varepsilon(r) - \mathbb{1}_{(t_0,t_1]}(r) f_\varepsilon(t_1)\right| dr + \cdots \\
&+ \int_{t_{k-1}}^{T} \left|f_\varepsilon(r) - \mathbb{1}_{(t_{k-1},T]}(r) f_\varepsilon(T)\right| dr \\
&= \int_0^{t_1} |f_\varepsilon(r) - f_\varepsilon(t_1)|\, dr + \ldots + \int_{t_{k-1}}^{T} |f_\varepsilon(r) - f_\varepsilon(T)|\, dr \\
&\le \int_0^{t_1} \|f_\varepsilon(\cdot)\|_{C([0,T];\mathbb{R}^n)}\, dr + \ldots + \int_{t_{k-1}}^{T} \|f_\varepsilon(\cdot)\|_{C([0,T];\mathbb{R}^n)}\, dr \\
&\le k\, \|f_\varepsilon(\cdot)\|_{C([0,T];\mathbb{R}^n)} \max_{1\le i\le k}(t_i - t_{i-1}).
\end{aligned}$$

Then,

$$\int_0^T \left|f_\varepsilon(r) - \bar{f}_\varepsilon(r)\right| dr < \varepsilon.$$

Now, let

$$E_\delta = \bigcup_{i=1}^{k} [t_{i-1}, t_{i-1} + \delta(t_i - t_{i-1})]. \tag{4.6}$$

Note that $\lambda(E_\delta) = \sum_{i\ge k} \delta(t_i - t_{i-1}) = \delta T$.

We prove that the integral

$$I_i := \int_{t_{i-1}}^{t_i} \left(1 - \frac{1}{\delta}\mathbb{1}_{E_\delta}(r)\right) \bar{f}_\varepsilon(r)dr = 0, \tag{4.7}$$

for all subinterval $(t_{i-1}, t_i]$ which not enclose s. Substituting (4.4) in (4.7) yields

$$\begin{aligned}
I_i &= \int_{t_{i-1}}^{t_i} \left(1 - \frac{1}{\delta}\mathbb{1}_{E_\delta}(r)\right) \sum_{i=1}^{k} f_\varepsilon(t_i)\mathbb{1}_{(t_{i-1},t_i]}(r)dr \\
&= \int_{t_{i-1}}^{t_i} \left(1 - \frac{1}{\delta}\mathbb{1}_{E_\delta}(r)\right) f_\varepsilon(t_i)dr \\
&= f_\varepsilon(t_i) \int_{t_{i-1}}^{t_i} \left(1 - \frac{1}{\delta}\mathbb{1}_{[t_{i-1},t_{i-1}+\delta(t_i-t_{i-1})]}(r)\right) \\
&= f_\varepsilon(t_i) \left[(t_i - t_{i-1}) - \frac{1}{\delta}\delta(t_i - t_{i-1})\right] \\
&= 0.
\end{aligned}$$

Now we consider $s \in (t_{j-1}, t_j]$ and estimate

$$I_j := \left| \int_{t_{j-1}}^{s} \left(1 - \frac{1}{\delta}\mathbb{1}_{E_\delta}(r)\right) \bar{f}_\varepsilon(r)dr \right|.$$

By definition of $\bar{f}_\varepsilon$ we have

$$\begin{aligned}
I_j &= \left| \int_{t_{j-1}}^{s} \left(1 - \frac{1}{\delta}\mathbb{1}_{E_\delta}(r)\right) \bar{f}_\varepsilon(r)dr \right| \\
&= |f_\varepsilon(t_j)| \left| \int_{t_{j-1}}^{s} \left(1 - \frac{1}{\delta}\mathbb{1}_{[t_{j-1},t_{j-1}+\delta(t_j-t_{j-1})]}(r)\right) dr \right| \\
&= |f_\varepsilon(t_j)| \left| s - t_{j-1} - \frac{1}{\delta}\left\{(s - t_{j-1}) \wedge [\delta(t_j - t_{j-1})]\right\} \right| \\
&\leq |f_\varepsilon(t_j)| \, (t_j - t_{j-1}) \\
&< \|f_\varepsilon(\cdot)\| \max_{1\leq i\leq k}(t_i - t_{i-1}) \\
&< \frac{\varepsilon}{k} < \varepsilon.
\end{aligned}$$

Thus,

$$\left| \int_{0}^{s} \left(1 - \frac{1}{\delta}\mathbb{1}_{E_\delta}(r)\right) \bar{f}_\varepsilon(r)dr \right| < \varepsilon. \tag{4.8}$$

Now,

$$\begin{aligned}\left|\int_0^s \left(1-\frac{1}{\delta}\mathbb{1}_{E_\delta}(r)\right) f(r)dr\right| &\leq \left|\int_0^s \left(1-\frac{1}{\delta}\mathbb{1}_{E_\delta}(r)\right)(f(r)-f_\varepsilon(r))dr\right| \\ &+ \left|\int_0^s \left(1-\frac{1}{\delta}\mathbb{1}_{E_\delta}(r)\right)(f_\varepsilon(r)-\bar{f}_\varepsilon(r))dr\right| \\ &+ \left|\int_0^s \left(1-\frac{1}{\delta}\mathbb{1}_{E_\delta}(r)\right)\bar{f}_\varepsilon(r)dr\right|.\end{aligned}$$

Note that, inequality (4.2)

$$\begin{aligned}\left|\int_0^s \left(1-\frac{1}{\delta}\mathbb{1}_{E_\delta}(r)\right)(f(r)-f_\varepsilon(r))dr\right| &\leq \int_0^s \left|1-\frac{1}{\delta}\mathbb{1}_{E_\delta}(r)\right| |f(r)-f_\varepsilon(r)|\,dr \\ &\leq \frac{1+\delta}{\delta}\int_0^s |f(r)-f_\varepsilon(r)|\,dr \\ &< \frac{(1+\delta)\varepsilon}{\delta},\end{aligned}$$

and by the inequality (4.5)

$$\left|\int_0^s \left(1-\frac{1}{\delta}\mathbb{1}_{E_\delta}(r)\right)(f_\varepsilon(r)-\bar{f}_\varepsilon(r))dr\right| < \frac{(1+\delta)\varepsilon}{\delta}.$$

Note that given $\varepsilon > 0$ we obtain a set E_δ such that

$$\left|\int_0^s \left(1-\frac{1}{\delta}\mathbb{1}_{E_\delta}(r)\right) f(r)dr\right| < \frac{2(1+\delta)\varepsilon}{\delta} + \varepsilon.$$

Hence

$$\|g_E(t)\|_{C([0,T];\mathbb{R}^n)} = \sup\left|\int_0^s \left(1-\frac{1}{\delta}\mathbb{1}_{E_\delta}(r)\right) f(r)dr\right| < K\varepsilon.$$

Therefore, $\inf\|g_E(t)\| = 0$. ∎

Corollary 4.1. *Suppose $f(\cdot) \in L^1([0,T];\mathbb{R}^n)$ and $0 < \delta < 1$. Then, there exists E_δ and a function $r_\delta \in L^1([0,T];\mathbb{R}^n)$ such that $\lambda(E_\delta) = \delta T$ and*

$$\delta\int_0^\tau f(s)ds = \int_0^\tau \mathbb{1}_{E_\delta}(s)f(s)ds + r_\delta(\tau), \quad \tau \in [0,T],$$

with $|r_\delta(\tau)| < \delta^2$ for all $\tau \in [0,T]$.

Proof. Since $0 < \delta < 1$, by the spike variation Lemma 4.2 there is $E_\delta \in \mathbb{E}_\delta$ such that

$$\sup_{\tau\in[0,T]}\left|\int_0^\tau f(s)ds - \frac{1}{\delta}\int_0^\tau \mathbb{1}_{E_\delta}(s)f(s)ds\right| < \delta,$$

that is

$$\left| \int_0^\tau f(s)ds - \frac{1}{\delta} \int_0^\tau \mathbb{1}_{E_\delta}(s) f(s)ds \right| < \delta, \quad \forall\, \tau \in [0,T].$$

Then

$$\left| \delta \int_0^\tau f(s)ds - \int_0^\tau \mathbb{1}_{E_\delta}(s) f(s)ds \right| < \delta^2, \quad \forall\, \tau \in [0,T].$$

Take

$$r_\delta(\tau) := \delta \int_0^\tau f(s)ds - \int_0^\tau \mathbb{1}_{E_\delta}(s) f(s)ds, \quad \tau \in [0,T].$$

Hence $|r_\delta(\tau)| < \delta^2$ and

$$\delta \int_0^\tau f(s)ds = \int_0^\tau \mathbb{1}_{E_\delta}(s) f(s)ds + r_\delta(\tau),$$

for all $\tau \in [0,T]$.

■

Proposition 4.2 ([22], Proposition 1.4.8)**.** *Let $M \subseteq \mathbb{R}^n$ be a non-empty closed convex set. Then there exists a map $P_M : \mathbb{R}^n \to M$ such that*

(i) $|x - P_M(x)| = \min_{y \in M} |x - y| =: d_M(x)$.

(ii) For $z \in M$, $z = P_M(x)$ if and only if

$$\langle x - z, y - z \rangle \leq 0, \quad \forall\ y \in M. \tag{4.9}$$

(iii) $|P_M(x_1) - P_M(x_2)| \leq |x_1 - x_2|$, *for all x_1, x_2 in $\mathbb{R}^n$.*

(iv) $\nabla_x d_M(x)^2 = 2(x - P_M(x))$.

Proof.

(i) Fix $x \in \mathbb{R}^n$. Let $\{z_k\} \subseteq M$ be a minimizing sequence such that $\lim_{k\to\infty} |x - z_k| = d_M(x)$ for any $x \in \mathbb{R}^n$. Since $\{|x - z_k|\}$ is convergent, we have that $\{z_k\}$ is bounded. We assume that $z_k \to \bar{z} \in M$, otherwise we can extract a convergent subsequence. Then

$$\lim_{k\to\infty} |x - z_k| = |x - \bar{z}| = d_M(x).$$

Suppose that there is another $\bar{y} \in M$ such that $|x - \bar{y}| = d_M(x)$. By the convexity of M, $\dfrac{\bar{y} + \bar{z}}{2} \in M$. Thus

$$\begin{aligned}
(d_M(x))^2 &\leq \left| x - \frac{\bar{y} + \bar{z}}{2} \right|^2 \\
&= \frac{1}{4} |2x - \bar{y} - \bar{z}|^2 \\
&= \frac{1}{4} |x - \bar{y} + x - \bar{z}|^2 \\
&= \frac{1}{4}(|(x - \bar{y}) + (x - \bar{z})|^2 + |(x - \bar{y}) - (x - \bar{z})|^2 - |\bar{y} - \bar{z}|^2) \\
&= \frac{1}{4}(2 |x - \bar{y}|^2 + 2 |x - \bar{z}|^2 - |\bar{y} - \bar{z}|^2) \\
&= [d_M(x)]^2 - \frac{1}{4} |\bar{y} - \bar{z}|^2 ,
\end{aligned}$$

thus, $\bar{y} = \bar{z}$.

(ii) Suppose that $P_M(x) \in M$. Then for any $y \in M$ and $\alpha \in (0, 1)$, we have

$$P_M(x) + \alpha(y - P_M(x)) = (1 - \alpha)P_M(x) + \alpha y \in M.$$

Then,

$$|P_M(x) - x|^2 \leq |P_M(x) + \alpha(y - P_M(x)) - x|^2 ,$$

which implies

$$\begin{aligned}
0 &\leq |P_M(x) - x| + |\alpha(y - P_M(x))|^2 - |P_M(x) - x|^2 \\
&= 2\alpha\langle P_M(x) - x, y - P_M(x)\rangle + \alpha^2 |y - P_M(x)|^2 \\
&= -2\alpha\langle P_M(x) - x, y - P_M(x)\rangle + \alpha^2 |y - P_M(x)|^2 .
\end{aligned}$$

Dividing by α and multiplying by -1 we have

$$\langle P_M(x) - x, y - P_M(x)\rangle - \alpha |y - P_M(x)|^2 \leq 0.$$

Letting $\alpha \to 0$ we get

$$\langle x - P_M(x), y - P_M(x)\rangle \leq 0, \quad \forall y \in M. \tag{4.10}$$

Now, suppose that, for $z \in M$

$$\langle x - z, y - z\rangle \leq 0, \quad \forall y \in M.$$

Note $|y-x|^2 = |z-x|^2 + |y-z|^2 + 2\langle y-z, z-x\rangle$, then

$$|y-x|^2 - |z-x|^2 = |y-z|^2 + 2\langle y-z, z-x\rangle,$$

for all $y \in M$. Thus $|y-x| \geq |z-x|$ for all $y \in M$. By definition of infimum $z = P_M(x)$.

(iii) From the inequality (4.10), for any $x_1, x_2 \in \mathbf{R}^n$, we have

$$\langle P_M(x_1) - P_M(x_2), x_2 - P_M(x_2)\rangle \leq 0,$$

and

$$\langle P_M(x_2) - P_M(x_1), x_1 - P_M(x_1)\rangle = \langle P_M(x_1) - P_M(x_2), P_M(x_2) - x_1\rangle \leq 0.$$

Adding the both inequalities we obtain

$$\langle P_M(x_1) - P_M(x_2), x_2 - P_M(x_2) - x_1 + P_M(x_1)\rangle \leq 0,$$

thus

$$\langle P_M(x_1) - P_M(x_2), P_M(x_1) - P_M(x_2) - (x_1 - x_2)\rangle \leq 0.$$

The last inequality implies

$$|P_M(x_1) - P_M(x_2)|^2 \leq |P_M(x_1) - P_M(x_2)|\,|x_1 - x_2|\,.$$

(iv) Note that

$$d_M^2(x) = \min\{|x-y|^2 \,:\, y \in Y\},$$

with $Y = M \cap \{z \,:\, |z - P_M(x)| \leq 1\}$ and the set of minimizers is $Y(x) = \{P_M(x)\}$. So, by Danskin Theorem A.3

$$\begin{aligned}\nabla_x d_M^2(x) &= \nabla_x(|x-y|^2)|_{y=P_M(x)} \\ &= 2(x-y)|_{y=P_M(x)} \\ &= 2(x - P_M(x)).\end{aligned}$$

■

4.2 Proof of Pontraying's Maximum Principle

Originally the result is establish as the Pontryagin Maximum Principle but, usually, in optimization the problems are establish as minimization problems. Following these ideas we prove the minimum version of the Pontryagin Maximum Principle.

We first introduce the last condition

Condition 4. *The map*

$$x \mapsto \big(f(t,u,x), g(t,u,x), h(x)\big),$$

is differentiable, the map

$$(x,u) \mapsto \big(f(t,u,x), f_x(t,u,x), g(t,u,x), g_x(t,u,x), h_x(x)\big),$$

is continuous, bounded and

$$\begin{aligned} |f_x(t,x_1,u) - f_x(t,x_2,u)| &\leq L\,|x_1 - x_2|\,, \\ |g_x(t,x_1,u) - g_x(t,x_2,u)| &\leq L\,|x_1 - x_2|\,, \end{aligned}$$

for some constant $L > 0$.

Now, according to the Optimal Control Problem 3.1 we consider the adjoint equation

$$\dot{\psi}(s) = -f_x(s,\bar{X}(s),\bar{u}(s))^\top \psi(s) - \psi^0 g_x(s,\bar{X}(s),\bar{u}(s))^\top, s \in [0,T], \tag{4.11}$$

with

$$\psi^0 \geq 0, \tag{4.12}$$

and

$$\left|\psi^0\right|^2 + \left|\psi(T) - \psi^0 h_x(\bar{X}(T))^\top\right|^2 = 1. \tag{4.13}$$

We also define the Hamiltonian function regarding to the Optimal Control Problem 3.1 as

$$H(s,x,u,\psi^0,\psi) := \psi^0 g(s,x,u) + \langle \psi, f(s,x,u)\rangle, \tag{4.14}$$

with $(s,x,u,\psi^0,\psi) \in [0,T] \times \mathbb{R}^n \times U \times \mathbb{R} \times \mathbb{R}^n$.

Theorem 4.2 (Pontryagin Minimum Principle)**.** *Assume Conditions 1, 2, and 4. Let M be a non-empty closed convex set. Suppose* $(\bar{X}(\cdot),\bar{u}(\cdot))$ *is an optimal pair of the Optimal*

Control Problem 3.1 for the initial pair $(0, x_0)$ *and* ψ *is the solution of the adjoint equation* (4.11). *Then the following conditions holds:*

(P-1) *Minimum condition:*

$$H(s, \bar{X}(s), \bar{u}, \psi^0, \psi(s)) = \min_{u \in U} H(s, \bar{X}(s), u, \psi^0, \psi(s)), \qquad a.e.\ s \in [0, T], \tag{4.15}$$

(P-2) *Transversality condition:*

$$\langle \psi(T) - \psi^0 h_x(\bar{X}(T))^\top, y - \bar{X}(T) \rangle \leq 0, \quad \forall y \in M. \tag{4.16}$$

Proof. We prove this theorem, following the next steps. **Step**-1 Introduce an auxiliary cost functional J_ε. **Step**-2 Apply the Corollary (2.3) of the E.V.P to the functional J_ε, in order to obtain an ε-optimal pair (x, u^ε). **Step**-3 From the corollary 4.1 of the spike variation lemma we obtain the necessary conditions for the ε-optimal pair. **Step**-4 Take the limit $\varepsilon \to 0$ to obtain the necessary conditions for the original problem.

Step-1 If $\bar{u}$ is the optimal control and $\bar{X}$ is the corresponding path, then, without loss of generality, we may assume that

$$J(\bar{u}) = J(\bar{u}(\cdot)) = \int_0^T g(s, \bar{u}(s), \bar{X}(s))ds + h(\bar{X}(t)) = 0,$$

otherwise, we can consider the functional $J(u) - J(\bar{u})$. Let $\varepsilon > 0$, $X(T) = X(T; 0, x_0, u(\cdot))$ and define the functional

$$J_\varepsilon(u) = \left[(J(u) + \varepsilon)^2 + d_M^2 \left(X(T) \right) \right]^{1/2} \geq 0, \tag{4.17}$$

where

$$d_M(x) = \min_{y \in M}(|x - y|).$$

for all $x \in \mathbb{R}^n$.

Step-2 First, we have that $\mathcal{U}[0, T]$ is a complete metric space by Lemma 4.1. Also, note that J_ε is continuous by Theorem 4.1, bounded below and

$$\begin{aligned} J_\varepsilon(\bar{u}) &= \varepsilon \\ &\leq \inf_{u \in \mathcal{U}_x^M} J_\varepsilon(u) + \varepsilon. \end{aligned}$$

By Corollary 2.3 there is $u^\varepsilon \in \mathcal{U}[0,T]$ such that

(i) $J_\varepsilon(u^\varepsilon) \leq J_\varepsilon(\bar{u})$,

(ii) $\rho(u^\varepsilon, \bar{u}) \leq \sqrt{\varepsilon}$,

(iii) $J_\varepsilon(u^\varepsilon) \leq J_\varepsilon(u) + \sqrt{\varepsilon}\rho(u, u^\varepsilon)$.

Thus, u^ε is a minimum of the map

$$u \longmapsto J_\varepsilon(u) + \sqrt{\varepsilon}\rho(u, u^\varepsilon).$$

Step-3 We obtain the necessary conditions for the ε-pair $(x^\varepsilon(\cdot), u^\varepsilon(\cdot))$. Let $\varepsilon > 0$ be fixed. For each $0 < \delta < 1$ and $u \in \mathcal{U}$ we apply the Corollary 4.1 to the map

$$\tau \longmapsto \begin{pmatrix} g(\tau, X^\varepsilon(\tau), u(\tau)) - g(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau)) \\ f(\tau, X^\varepsilon(\tau), u(\tau)) - f(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau)) \end{pmatrix},$$

then, there is $E_\delta^\varepsilon \in \mathbb{E}_\delta$ with $\lambda(E_\delta^\varepsilon) = \delta T$ and a function $r_\delta \in L^1([0,T];\mathbb{R}^n)$ such that

$$\begin{aligned} &\delta \int_0^s \begin{pmatrix} g(s, X^\varepsilon(\tau), u(\tau)) - g(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau)) \\ f(\tau, X^\varepsilon(\tau), u(\tau)) - f(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau)) \end{pmatrix} d\tau \\ &= \int_0^s \mathbb{1}_{E_\delta^\varepsilon}(\tau) \begin{pmatrix} g(s, X^\varepsilon(\tau), u(\tau)) - g(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau)) \\ f(\tau, X^\varepsilon(\tau), u(\tau)) - f(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau)) \end{pmatrix} d\tau + \begin{pmatrix} r_\delta^{0,\varepsilon}(s) \\ r_\delta^\varepsilon(s) \end{pmatrix} \end{aligned} \tag{4.18}$$

where $\left|r_\delta^{0,\varepsilon}(s)\right| + |r_\delta^\varepsilon(s)| \leq \delta^2$ for all $s \in [0,T]$ and $X^\varepsilon(\cdot) = X(\cdot, 0, x_0, u^\varepsilon(\cdot))$. Thus, given u and δ, we define the spike variation u_δ^ε of the optimal control u^ε as

$$u_\delta^\varepsilon(s) = \begin{cases} u^\varepsilon(s) & \text{if} \quad s \in [0,T] \setminus E_\delta^\varepsilon \\ u(s) & \text{if} \quad s \in E_\delta^\varepsilon \end{cases} \tag{4.19}$$

Now, define $X_\delta^\varepsilon(\cdot) := X(\cdot; 0, x_0, u_\delta^\varepsilon(\cdot))$ and

$$Y_\delta^\varepsilon(s) := \frac{X_\delta^\varepsilon(s) - X^\varepsilon(s)}{\delta}, \quad s \in [0,T].$$

Then, by the equation (4.18)

$$\begin{aligned}
Y_\delta^\varepsilon(s) &= \frac{1}{\delta}\int_0^s \big[f(\tau, X_\delta^\varepsilon(\tau), u_\delta^\varepsilon(\tau)) - f(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau))\big]d\tau \\
&= \frac{1}{\delta}\int_0^s \big[f(\tau, X_\delta^\varepsilon(\tau), u_\delta^\varepsilon(\tau)) - f(\tau, X^\varepsilon(\tau), u_\delta^\varepsilon(\tau))\big]d\tau \\
&+ \frac{1}{\delta}\int_0^s \big[f(\tau, X^\varepsilon(\tau), u_\delta^\varepsilon(\tau)) - f(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau))\big]d\tau \\
&= \frac{1}{\delta}\int_0^s \big[f(\tau, X_\delta^\varepsilon(\tau), u_\delta^\varepsilon(\tau)) - f(\tau, X^\varepsilon(\tau), u_\delta^\varepsilon(\tau))\big]d\tau \\
&+ \frac{1}{\delta}\int_{[t,s]\cap E_\delta^\varepsilon} \big[f(\tau, X^\varepsilon(\tau), u(\tau)) - f(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau))\big]d\tau \\
&= \int_0^s \left[\int_0^1 f_x(\tau, X^\varepsilon(\tau) + \theta(X_\delta^\varepsilon(\tau) - X^\varepsilon(\tau)), u_\delta^\varepsilon(\tau))d\theta\right] Y_\delta^\varepsilon(\tau)d\tau \\
&+ \int_0^s \big[f(\tau, X^\varepsilon(\tau), u(\tau)) - f(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau))\big]d\tau - \frac{r_\delta^\varepsilon(s)}{\delta}.
\end{aligned}$$

Consider the following initial value problem

$$\begin{aligned}
\dot{Y}^\varepsilon(s) &= f_x(s, X^\varepsilon(s), u^\varepsilon(s))Y^\varepsilon(s) \\
&\quad + f(s, X^\varepsilon(s), u(s)) - f(s, X^\varepsilon(s), u^\varepsilon(s)), \qquad s \in [0, T], \\
Y^\varepsilon(0) &= 0.
\end{aligned}$$

Then the solution Y^ε is

$$Y^\varepsilon(s) = \int_0^s f_x(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau))Y^\varepsilon(\tau)d\tau + \int_0^s \big[f(\tau, X^\varepsilon(\tau), u(\tau)) - f(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau))\big]d\tau.$$

By Theorem A.4 we have that

$$\lim_{\delta\to 0} \|Y_\delta^\varepsilon(\cdot) - Y^\varepsilon(\cdot)\|_{C([0,T];\mathbb{R}^n)} = 0. \tag{4.20}$$

On the other hand, note that

$$\begin{aligned}
-T\sqrt{\varepsilon} &= -\sqrt{\varepsilon}\frac{\lambda(E^{\varepsilon}_{\delta})}{\delta} \\
&= -\sqrt{\varepsilon}\frac{\rho(u^{\varepsilon}_{\delta}(\cdot), u^{\varepsilon}(\cdot))}{\delta} \\
&\leq \frac{1}{\delta}\big[J_{\varepsilon}(u^{\varepsilon}_{\delta}(\cdot)) - J_{\varepsilon}(u^{\varepsilon}(\cdot))\big] \\
&= \frac{1}{\delta}\frac{\big[J_{\varepsilon}(u^{\varepsilon}_{\delta}(\cdot))\big]^2 - \big[J_{\varepsilon}(u^{\varepsilon}(\cdot))\big]^2}{J_{\varepsilon}(u^{\varepsilon}_{\delta}(\cdot)) + J_{\varepsilon}(u^{\varepsilon}_{\delta}(\cdot))} \\
&= \frac{\big[J(u^{\varepsilon}_{\delta}(\cdot)) + \varepsilon\big]^2 - \big[J(u^{\varepsilon}(\cdot)) + \varepsilon\big]^2}{\delta[J_{\varepsilon}(u^{\varepsilon}_{\delta}(\cdot)) + J_{\varepsilon}(u^{\varepsilon}_{\delta}(\cdot))]} + \frac{d(X^{\varepsilon}_{\delta}(T), M)^2 - d(X^{\varepsilon}(T), M)^2}{\delta[J_{\varepsilon}(u^{\varepsilon}_{\delta}(\cdot)) + J_{\varepsilon}(u^{\varepsilon}_{\delta}(\cdot))]}.
\end{aligned}$$

Define

$$\begin{aligned}
\psi^{0,\varepsilon}_{\delta} &:= \frac{[J(u^{\varepsilon}_{\delta}(\cdot)) + \varepsilon] + [J(u^{\varepsilon}(\cdot)) + \varepsilon]}{J_{\varepsilon}(u^{\varepsilon}_{\delta}(\cdot)) + J_{\varepsilon}(u^{\varepsilon}(\cdot))}, \\
\psi^{\varepsilon}_{\delta} &:= \frac{\int_0^1 \nabla_x(d^2_M)(X^{\varepsilon}(T) + \theta[X^{\varepsilon}_{\delta}(T) - X^{\varepsilon}(T)])d\theta}{J_{\varepsilon}(u^{\varepsilon}_{\delta}(\cdot)) + J_{\varepsilon}(u^{\varepsilon}(\cdot))}.
\end{aligned}$$

Thus, using the same method in we have

$$\begin{aligned}
-T\sqrt{\varepsilon} \leq \psi^{0,\varepsilon}_{\delta}\Big\{\frac{1}{\delta}\int_0^T [g(s, X^{\varepsilon}_{\delta}(s), u^{\varepsilon}_{\delta}) - g(s, X^{\varepsilon}(s), u^{\varepsilon})]ds + \frac{h(X^{\varepsilon}_{\delta}(T)) - h(X^{\varepsilon}(T))}{\delta}\Big\} \\
+ \psi^{\varepsilon}_{\delta} Y^{\varepsilon}_{\delta}(T),
\end{aligned} \tag{4.21}$$

By the definition of u^{ε}_{δ} in (4.19), $u^{\varepsilon}_{\delta}(s) = u^{\varepsilon}(s)$ a.e. $s \in [0, T]$ when $\delta \to 0$. Then, by Proposition 4.2 part (iv) and Condition 4 we have

$$\begin{aligned}
\lim_{\delta\to 0} \psi^{0,\varepsilon}_{\delta} &= \frac{J(u^{\varepsilon}(\cdot)) + \varepsilon}{J_{\varepsilon}(u^{\varepsilon}(\cdot))} =: \psi^{0,\varepsilon}, \\
\lim_{\delta\to 0} \psi^{\varepsilon}_{\delta} &= \frac{X^{\varepsilon}(T) - P_M(X^{\varepsilon}(T))}{J_{\varepsilon}(u^{\varepsilon}(\cdot))} =: \psi^{\varepsilon}.
\end{aligned}$$

Note that

$$\begin{aligned}
\big|\psi^{0,\varepsilon}\big|^2 + |\psi^{\varepsilon}|^2 &= \frac{(J(u^{\varepsilon}(\cdot)) + \varepsilon)^2 + (X^{\varepsilon}(T) - P_M(X^{\varepsilon}(T))^2}{J_{\varepsilon}(u^{\varepsilon}(\cdot))^2} \\
&= \frac{(J(u^{\varepsilon}(\cdot)) + \varepsilon)^2 + d_M(X^{\varepsilon}(T))^2}{J_{\varepsilon}(u^{\varepsilon}(\cdot))^2} \\
&= 1,
\end{aligned}$$

thus

$$\big|\psi^{\varepsilon,0}\big|^2 + |\psi^{\varepsilon}|^2 = 1, \quad \forall\, \epsilon > 0. \tag{4.22}$$

Also, from the second statement in Proposition 4.2 we have

$$\langle X^\varepsilon(T) - P_M(X^\varepsilon(T)), y - X^\varepsilon(T)\rangle \le 0, \quad \forall\, y \in M.$$

Multiplying by $\frac{1}{J_\varepsilon(u^\varepsilon)}$ we get

$$\langle \psi^\varepsilon, y - X^\varepsilon(T)\rangle \le 0, \quad \forall y \in M. \tag{4.23}$$

From the inequality (4.21), applying the same method in we get

$$\begin{aligned}
&\frac{1}{\delta}\int_0^T \Big[g(\tau, X^\varepsilon_\delta(\tau), u^\varepsilon_\delta(\tau)) - g(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau))\Big]d\tau \\
&= \frac{1}{\delta}\int_0^T \Big[g(\tau, X^\varepsilon_\delta(\tau), u^\varepsilon_\delta(\tau)) - g(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau))\Big]d\tau \\
&+ \frac{1}{\delta}\int\limits_{[t,s]\cap E^\varepsilon_\delta} \Big[g(\tau, X^\varepsilon(\tau), u(\tau)) - g(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau))\Big]d\tau \\
&= \int_0^T\int_0^1 g_x(\tau, X^\varepsilon + \theta[X^\varepsilon_\delta(\tau) - X^\varepsilon(\tau)], u^\varepsilon_\delta(\tau))Y^\varepsilon_\delta(\tau)d\theta d\tau \\
&+ \int_0^T \Big[g(\tau, X^\varepsilon(\tau), u(\tau)) - g(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau))\Big]d\tau - \frac{r^{\varepsilon,0}_\delta(\tau)}{\delta}.
\end{aligned}$$

Letting $\delta \to 0$ we have $u^\varepsilon_\delta(s) = u^\varepsilon(s)$ a.e. $s \in [0,T]$. Thus, by Condition 4 and (4.20) we obtain

$$\begin{aligned}
&\frac{1}{\delta}\int_0^T \Big[g(\tau, X^\varepsilon_\delta(\tau), u^\varepsilon_\delta(\tau)) - g(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau))\Big]d\tau \\
&\to \int_0^T \Big[g_x(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau))Y^\varepsilon(\tau) + g(\tau, X^\varepsilon(\tau), u(\tau)) - g(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau))\Big]d\tau.
\end{aligned} \tag{4.24}$$

Now,

$$\begin{aligned}
\frac{h(X^\varepsilon_\delta(T)) - h(X^\varepsilon(T))}{\delta} &= \frac{1}{\delta}\int_0^1 h_x\big(X^\varepsilon(T) + \theta[X^\varepsilon_\delta(T) - X^\varepsilon(T)]\big)d\theta \cdot [X^\varepsilon_\delta(T) - X^\varepsilon(T)] \\
&= \int_0^1 h_x\big(X^\varepsilon(T) + \theta[X^\varepsilon_\delta(T) - X^\varepsilon(T)]\big)d\theta \cdot \frac{X^\varepsilon_\delta(T) - X^\varepsilon(T)}{\delta}.
\end{aligned}$$

If we let $\delta \to 0$, we get

$$\frac{h(X^\varepsilon_\delta(T)) - h(X^\varepsilon(T))}{\delta} \to \int_0^1 h_x\big(X^\varepsilon(T)\big)d\theta \cdot Y^\varepsilon(T) = h_x\big(X^\varepsilon(T)\big) \cdot Y^\varepsilon(T) \tag{4.25}$$

Consequently, letting $\delta \to 0$ in (4.21), by (4.24) and (4.25) we obtain

$$
\begin{aligned}
-T\sqrt{\varepsilon} \leq \psi^{0,\varepsilon} \int_0^T \Big[& g_x(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau)) Y^\varepsilon(\tau) + g(\tau, X^\varepsilon(\tau), u(\tau)) \\
& - g(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau)) \Big] d\tau + \Big[\psi^{0,\varepsilon} h_x(X^\varepsilon(T)) + \psi^\varepsilon \Big] Y^\varepsilon(T).
\end{aligned} \tag{4.26}
$$

Step-4 Note that (4.22) and (4.23) are valid for every ε, in particular for the limit. Thus, we define $(\psi^0, \bar{\psi}) := \lim_{\varepsilon \to 0} (\psi^{0,\varepsilon}, \psi^\varepsilon)$, such that

$$
\begin{aligned}
& \left|\psi^0\right|^2 + \left|\bar{\psi}\right|^2 = 1, \quad \psi^0 \geq 0, \\
& \langle \bar{\psi}, y - \bar{X}(T) \rangle \leq 0, \quad \forall\, y \in M.
\end{aligned} \tag{4.27}
$$

On the other hand, denote by Y the solution to the initial value problem

$$
\begin{aligned}
& \dot{Y}(s) = f_x(s, \bar{X}(s), \bar{u}(s))^\top Y(s) + f(s, \bar{X}(s), u(s)) - f(s, \bar{X}(s), \bar{u}(s)), \\
& Y(0) = 0,
\end{aligned}
$$

with $s \in [0, T]$. Then, by Theorem A.5

$$
\lim_{\varepsilon \to 0} \|Y^\varepsilon(\cdot) - Y(\cdot)\|_{C([0,T];\mathbb{R}^n)} = 0.
$$

Consider the problem of terminal value

$$
\begin{aligned}
& \dot{\psi}(s) = -f_x(s, \bar{X}(s), \bar{u}(s))^\top \psi(s) - \psi^0 g_x(s, \bar{X}(s), \bar{u}(s))^\top, \quad s \in [0, T], \\
& \psi(T) = \psi + \psi^0 h_x(\bar{X}(T))^\top.
\end{aligned} \tag{4.28}
$$

The terminal condition satisfies (4.27), that is

$$
\left|\psi^0\right|^2 + \left|\psi(T) - \psi^0 h_x(\bar{X}(T))^\top\right|^2 = 1, \quad \psi^0 \geq 0
$$

and

$$
\left\langle \psi(T) - \psi^0 h_x(\bar{X}(T))^\top,\, y - \bar{X}(T) \right\rangle \leq 0 \quad \forall y \in M. \tag{4.29}
$$

This prove (4.12), (4.13) and (4.16).

On the other hand, according to the initial value problem for Y and ψ consider the following derivative

$$\begin{aligned}\frac{d}{ds}\langle Y(s),\psi(s)\rangle &= \langle Y(s),\dot{\psi}(s)\rangle + \langle \dot{Y}(s),\psi(s)\rangle \\ &= \big\langle Y(s), -f_x(s,\bar{X}(s),\bar{u}(s))^\top\psi(s) - \psi^0 g_x(s,\bar{X}(s),\bar{u}(s))^\top\big\rangle \\ &+ \big\langle f_x(s,\bar{X}(s),\bar{u}(s))^\top Y(s) + f(s,\bar{X}(s),u(s)) - f(s,\bar{X}(s),\bar{u}(s)),\psi(s)\big\rangle \\ &= \big\langle Y(s), -\psi^0 g_x(s,\bar{X}(s),\bar{u}(s))^\top\big\rangle \\ &+ \big\langle f(s,\bar{X}(s),u(s)) - f(s,\bar{X}(s),\bar{u}(s)),\psi(s)\big\rangle.\end{aligned}$$

Integrating we obtain

$$\begin{aligned}\langle Y(T),\psi(T)\rangle - \langle Y(0),\psi(0)\rangle = \int_0^T \Big[&\big\langle Y(s), -\psi^0 g_x(s,\bar{X}(s),\bar{u}(s))^\top\big\rangle \\ &+ \big\langle f(s,\bar{X}(s),u(s)) - f(s,\bar{X}(s),\bar{u}(s)),\psi(s)\big\rangle\Big]ds,\end{aligned}$$

then

$$\begin{aligned}\langle Y(T),\psi + \psi^0 h_x(\bar{X}(T))\rangle = \int_0^T \Big[&\big\langle Y(s), -\psi^0 g_x(s,\bar{X}(s),\bar{u}(s))^\top\big\rangle \\ &+ \big\langle f(s,\bar{X}(s),u(s)) - f(s,\bar{X}(s),\bar{u}(s)),\psi(s)\big\rangle\Big]ds.\end{aligned} \tag{4.30}$$

Now, letting $\varepsilon \to 0$ in (4.26), we get

$$\begin{aligned}0 \le \int_0^T \psi^0\Big[g_x(s,\bar{X}(s),\bar{u}(s))Y(s) + g(s,\bar{X}(s),u(s)) - g(s,\bar{X}(s),\bar{u}(s))\Big]ds \\ + \Big[\psi^0 h_x(\bar{X}(T)) + \psi(s)\Big]Y(T).\end{aligned} \tag{4.31}$$

Applying the definition of the Hamiltonian, $H(s,x,u,\psi^0,\psi) = \psi^0 g(s,x,u) + \langle\psi, f(s,x,u)\rangle$, we get

$$\begin{aligned}H(\bar{X}(s),u(s)) - H(\bar{X}(s),\bar{u}(s)) &= H(s,\bar{X}(s),u(s),\psi^0,\psi(s)) - H(s,\bar{X}(s),\bar{u}(s),\psi^0,\psi(s)) \\ &= \psi^0\big[g(s,\bar{X}(s),u(s)) - g(s,\bar{X}(s),\bar{u}(s))\big] \\ &+ \big\langle\psi(s), f(s,\bar{X}(s),u(s)) - f(s,\bar{X}(s),\bar{u}(s))\big\rangle.\end{aligned}$$

Hence, by (4.30) and (4.31)

$$\begin{aligned}\int_0^T [H(\bar{X}(s),u(s)) - H(\bar{X}(s),\bar{u}(s))]ds &= \int_0^T \Big\{\psi^0\Big[g(s,\bar{X}(s),u(s)) - g(s,\bar{X}(s),\bar{u}(s))\Big] \\ &\quad + \big\langle \psi(s), f(s,\bar{X}(s),u(s)) - f(s,\bar{X}(s),\bar{u}(s))\big\rangle\Big\}ds \\ &= \int_0^T \Big\{\psi^0\Big[g(s,\bar{X}(s),u(s)) - g(s,\bar{X}(s),\bar{u}(s))\Big] \\ &\quad - \Big\langle Y(s), -\psi^0 g_x(s,\bar{X}(s),\bar{u}(s))^\top\Big\rangle\Big\}ds \\ &\quad + \Big\langle Y(T), \psi + \psi^0 h_x(\bar{X}(T))\Big\rangle \\ &\geq 0.\end{aligned}$$

Define now the operator $W : U \to \mathbb{R}$

$$W[u](s) := H(\bar{X}(s),u(s)) - H(\bar{X}(s),\bar{u}(s)), \quad s \in [0,T].$$

Since U is separable, let $\mathcal{A} = \{u_i,\ i \in \mathbb{N}\}$ be dense subset of U. Fix $i \in \mathbb{N}$, $t \in [0,T]$ and $\eta > 0$. Define

$$u_\eta(s) = \begin{cases} u_i, & s \in [t-\eta, t+\eta], \\ \bar{u}(s), & otherwise. \end{cases}$$

By the definition of the Hamiltonian, $W[u](s)$ is measurable. Then, by Theorem A.6

$$\begin{aligned}0 \leq \lim_{\eta\to 0} \frac{1}{2\eta}\int_0^T W[u_\eta](s)ds &= \lim_{\eta\to 0}\frac{1}{2\eta}\Big[\int_0^{t-\eta} W[u_\eta](s)ds + \int_{t-\eta}^{t+\eta} W[u_\eta](s)ds + \int_{t+\eta}^{T} W[u_\eta](s)ds\Big] \\ &= \lim_{\eta\to 0}\frac{1}{2\eta}\int_{t-\eta}^{t+\eta}[H(\bar{X}(s),u_i) - H(\bar{X}(s),\bar{u}(s))]ds \\ &= H(\bar{X}(t),u_i) - H(\bar{X}(t),\bar{u}(t)),\end{aligned}$$

for a.e. $t \in [0,T]$. Hence, for all $i \in \mathbb{N}$

$$H(\bar{X}(t),u_i) \geq H(\bar{X}(t),\bar{u}(t)), \quad a.e.\ t \in [0,T]. \tag{4.32}$$

Let $u \in U$. Since $\mathcal{A}$ is dense there exists a subsequence $\{u_{i_k}\} \subseteq \mathcal{A}$ such that $u_{i_k} \to u$ when $k \to \infty$. By (4.32)

$$H(\bar{X}(t),u_{i_k}) \geq H(\bar{X}(t),\bar{u}(t)), \quad \forall\, k \in \mathbb{N}.$$

Since $H(\bar{X}(t),\cdot)$ is continuous we get

$$H(\bar{X}(t),u)=\lim_{k\to\infty}H(\bar{X}(t),u_{i_k})\geq H(\bar{X}(t),\bar{u}(t)).$$

Therefore

$$H(t,\bar{X}(t),\bar{u}(t),\psi^0,\psi(t))=\min_{u\in U}H(t,\bar{X}(t),u,\psi^0,\psi(t)),\quad a.e.\ t\in[0,T].$$

■

Remark 4.1. *Note that in the case $M=\mathbb{R}^n$ the transversality condition becomes*

$$\psi(T)=\psi^0 h_x(\bar{X}(T))^\top.$$

In the last chapter we apply the Pontryagin principle to epidemiological problems. Because of this, we introduce a particular case of the Pontryagin Principle. Let $M\subseteq\mathbb{R}^n$ and $\psi^0=1$, and consider the optimal problem

$$\begin{aligned}&\max_{u(\cdot)}J(u(\cdot))=\max_{u(\cdot)}\left[\int_0^T g(s,x(s),u(s))dt+h(x(T))\right]\\ &\text{subject to}\ \ x'(s)=f(s,x(s),u(s)),\ \ x(0)=x_0.\end{aligned}\tag{4.33}$$

Then, we have the following version

Theorem 4.3. *If $(\bar{u}(\cdot),\bar{x}(\cdot))$ is an optimal pair for the optimal control problem 4.33, then there exists a piecewise differentiable adjoint variable $\psi(\cdot)$ such that*

$$H(s,\bar{x}(t),\bar{u}(t),\psi^0,\psi(t))=\max_{u\in U}H(s,\bar{X}(s),u,\psi^0,\psi(t)),$$

at each time t, where the Hamiltonian H is

$$H(s,x(s),u(s),\psi(s))=g(s,x(s),u(s))+\big\langle\psi(s),f(s,x(s),u(s))\big\rangle,\tag{4.34}$$

and $\psi(t)$ is the solution of the adjoint equation

$$\begin{aligned}\dot{\psi}(s)&=-\frac{\partial H(s,x(s),u(s),\psi(s))}{\partial x}\\ &=-\Big[g_x(s,x(s),u(s))+\big\langle\psi(s),f_x(s,x(s),u(s))\big\rangle\Big],\end{aligned}\tag{4.35}$$

with a transversality condition $\psi(T) = h_x(\bar{x}(T))$*. Moreover, we have the optimality condition*

$$\frac{\partial H(s, \bar{x}(s), \bar{u}(s), \psi(s))}{\partial u} = 0,$$

that is

$$g_u(s, \bar{x}(s), \bar{u}(s)) + f_u(s, \bar{x}(s), \bar{u}(s))\psi(s) = 0. \tag{4.36}$$

Chapter 5

The Forward-Backward Sweep Method

The Pontryagin principle allows us to transform the control problem to a problem of solving a system of ordinary differential equations. To this end, we need to solve forward the dynamic with a given initial condition, and backward the adjoint equations with a transversality condition. using the fourth order Runge-Kutta-Felberg method we construct the forward-backward sweep method. This way of solving is called the Forward-Backward Sweep method. The purpose of this Chapter is to present the Runge-Kutta methods, then the Forward-Backward Sweep method and some examples of these methods.

Here we present the Runge-Kutta methods following the ideas of [10]. Consider the following initial value problem (I.V.P.)

$$\begin{aligned} x'(t) &= f(t, x(t)), \qquad t \in [t_0, T], \\ x(t_0) &= x_0, \end{aligned} \tag{5.1}$$

The Runge-Kutta (RK) methods are used to approximate the solution of the initial value problem described by (5.1). Let $P = \{t_0, t_1, \ldots, t_N\}$ be a partition of the interval $[t_0, T]$. These methods compute the slopes of nearby points at a time t_n and then calculate the average of these slopes to approximate the solution at the next time t_{n+1}.

The general s-stage RK method is defined as

$$x_{n+1} = x_n + h \sum_{i=1}^{s} b_i k_i, \tag{5.2}$$

where the terms k_i are computed from the following evaluation in the right-hand side of the equation (5.1):

$$k_i = f\left(t_n + c_i h, x_n + h\sum_{j=1}^{s} a_{i,j}k_j\right), \qquad i = 1, \cdots s, \tag{5.3}$$

with

$$c_i = \sum_{j=1}^{s} a_{i,j}, \qquad i = 1, \cdots, s. \tag{5.4}$$

c	A
	b

TABLE 5.1: General form of a Butcher array.

These parameters can be displayed in a table known as the Butcher array (see Table 5.1). The vector c indicates the positions within the step of the stages values. The matrix A indicates the dependence of the stages on the derivatives found at other stages. And b is a vector of weights, showing how the final result depends on the derivatives computed at various stages.

To specify a particular method, we need to choose the stage s, and then provide the coefficients a_{ij} with $i, j = 1, \ldots, s$, the weight b_i and the terms c_i, with $i = 1, \ldots, s$. Thus, given s-stages, the method depends on $s^2 + s$ parameters $\{a_{i,j}, b_j\}$.

We focus on the explicit RK methods. In these methods the upper-triangular components of the matrix A are zero. In this case, the Butcher array is as follows:

c_1	0	0	$\cdots$	0	0
c_2	$a_{2,1}$	0	$\cdots$	0	0
c_3	$a_{3,1}$	$a_{3,2}$	$\cdots$	0	$\vdots$
$\vdots$	$\vdots$	$\vdots$		0	0
c_s	$a_{s,1}$	$a_{s,2}$	$\cdots$	$a_{s,s-1}$	0
	b_1	b_2	$\cdots$	b_{s-1}	b_s

TABLE 5.2: The Butcher array for an explicit RK method

We obtain the parameters in the RK methods, based on the definition below.

Definition 5.1 (Local Truncation Error, [10, def. 9.3])**.** *The Local Truncation Error, T_{n+1} of an RK method is defined as the difference between the exact $x(t_{n+1})$ and the*

numerical solution x_{n+1} *of the initial value problem* (5.1) *at* $t = t_{n+1}$*:*

$$T_{n+1} = x(t_{n+1}) - x_{n+1},$$

under the assumption that $x_n = x(t_n)$*. If* $T_{n+1} = \mathcal{O}(h^{p+1})$*, the method is said to be of order* p*.*

To clarify this idea, we present an example of a one-stage RK method, that is, we choose $s = 1$. Following the definition of the general RK method (5.2) and (5.3), we get

$$\begin{aligned} x_{n+1} &= x_n + hb_1k_1, \\ k_1 &= f(t_n + c_1h, x_n + ha_{1,1}k_1). \end{aligned}$$

Since the method is explicit, we have that $c_1 = a_{1,1} = 0$. Also, $k_1 = f(t_n, x_n) = f_n$, then $x_{n+1} = x_n + hb_1f_n$. To find the coefficient b_1, we have to compare this expression with the expression for $x(t_{n+1})$. First, We use the Taylor expansion A.1 at t_{n+1} of order 3:

$$x(t_{n+1}) = x(t_n) + hx'(t_n) + \frac{1}{2}h^2x''(t_n) + \mathcal{O}(h^3).$$

Now, we differentiate the equation $x'(t) = f(t, x(t))$ respect to t, in order to find the term $x''(t_n)$. By the chain rule, we obtain

$$x''(t) = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial x}x'(t) = f_t + f_xf.$$

Then,

$$\begin{aligned} T_{n+1} &= x(t_{n+1}) - x_{n+1} \\ &= x(t_n) + hx'(t_n) + \frac{1}{2}h^2x''(t_n) - x_n - hb_1f_n + \mathcal{O}(h^3) \\ &= x_n + hf_n + \frac{1}{2}h^2(f_t + ff_x)|_{t=t_n} - x_n - hb_1f_n + \mathcal{O}(h^3) \\ &= h(1 - b_1)f_n + \frac{1}{2}h^2(f_t + ff_x)|_{t=t_n} + \mathcal{O}(h^3), \end{aligned}$$

If T_{n+1} is consistent of order $p = 1$, then $b_1 = 1$. This choosing gives the only first-order one-stage explicit RK method, the Euler's method

$$x_{n+1} = x_n + hf_n.$$

We now present an example of how to approximate the solution of an initial value problem using a two-stage RK method. Consider the IVP

$$\begin{aligned} x'(t) &= (1-2t)x(t), \qquad t > 0, \\ x(0) &= 1. \end{aligned} \tag{5.5}$$

According to the two-stage RK method,

$$\begin{aligned} t_{n+1} &= t_n + h, \\ x_{n+1} &= x_n + hk_2, \end{aligned}$$

and

$$\begin{aligned} k_1 &= f(t_n, x_n), \\ k_2 &= f\left(t_n + \frac{1}{2}h, x_n + \frac{1}{2}hk_1\right), \end{aligned} \tag{5.6}$$

we calculate the approximate solution of 5.5 choosing $h = 0.2$. Thus, substituting 5.6 in 5.5, we obtain

$$\begin{aligned} k_1 &= (1-2t_n)x_n, \\ k_2 &= (1-(2t_n+h))\left(x_n + \frac{1}{2}hk_1\right). \end{aligned}$$

So, for $n = 0$, we have

$$\begin{aligned} k_1 &= (1-2t_0)x_0 = (1-2(0))(1) = 1, \\ k_2 &= (1-2t_0-h)(x_0 + 0.5hk_1) = (1-0.2)(1+0.5(0.2)(1)) = 0.88, \\ t_1 &= t_0 + h = 0 + 0.2 = 0.2, \\ x_1 &= x_0 + hk_2 = 1 + (0.2)(0.88) = 1.176. \end{aligned}$$

Now, for $n = 1$, yields

$$\begin{aligned} k_1 &= (1-2t_1)x_1 = (1-2(0.2))(1.176) = 0.7056, \\ k_2 &= (1-2t_1-h)(x_1 + 0.5hk_1) = (1-2(0.2)-0.2)(1.176+0.5(0.2)(0.7056)) = 0.4986, \\ t_2 &= t_1 + h = 0.2 + 0.2 = 0.4, \\ x_2 &= x_1 + hk_2 = 1.176 + (0.2)(0.4986) = 1.2757, \end{aligned}$$

and so on. This is an example of an explicit RK method.

The Runge-Kutta-Felberg method

Consider the initial value problem (5.1). Letting $s = 4$ we get an explicit four-stage RK method: (5.1).

$$x_{n+1} = x_n + h(b_1k_1 + b_2k_2 + b_3k_3 + b_4k_4) \tag{5.7}$$

where

$$\begin{aligned} k_1 &= f(t_n, x_n), \\ k_2 &= f(t_n + c_2h, x_n + ha_{2,1}k_1), \\ k_3 &= f\left(t_n + c_3h, x_n + h\sum_{j=1}^{2} a_{3,j}k_j\right), \\ k_4 &= f\left(t_n + c_4h, x_n + h\sum_{j=1}^{3} a_{1,j}k_j\right). \end{aligned} \tag{5.8}$$

To obtain the parameters we have to satisfy the following conditions (see p. 90, [4])

$$\begin{aligned} b_1 + b_2 + b_3 + b_4 &= 1, \\ b_2c_2 + b_3c_3 + b_4c_4 &= \frac{1}{2}, \\ b_2{c_2}^2 + b_3{c_3}^2 + b_4{c_4}^2 &= \frac{1}{3}, \\ b_3a_{3,2}c_2 + b_4a_{4,2}c_2 + b_4a_{4,3}c_3 &= \frac{1}{6}, \\ b_2{c_2}^3 + b_3{c_3}^3 + b_4{c_4}^3 &= \frac{1}{4}, \\ b_3c_3a_{3,2}c_2 + b_4c_4a_{4,2}c_2 + b_4c_4a_{4,3}c_3 &= \frac{1}{8}, \\ b_3a_{3,2}{c_2}^2 + b_4a_{4,2}{c_2}^2 + b_4a_{4,3}{c_3}^2 &= \frac{1}{12}, \\ b_4a_{4,3}a_{3,2}c_2 &= \frac{1}{24}. \end{aligned}$$

These systems of nonlinear equations a infinite number of solutions. Table 5.3 shows a popular solution. This set of parameters conform the Runge Kutta-Felberg method of fourth order. This method is common used in packages from R, Julia and MATLAB.

Substituting the parameters of Table 5.3 in (5.7) and (5.8), we obtain the Runge Kutta-Felberg method

$$x_{n+1} = x_n + \frac{1}{6}h(k_1 + 2k_2 + 2k_3 + k_4) \tag{5.9}$$

0	0			
$\frac{1}{2}$	$\frac{1}{2}$	0		
$\frac{1}{2}$	0	$\frac{1}{2}$	0	
1	0	0	1	0
	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$

TABLE 5.3: The Butcher array for an explicit four stage RK method

where

$$\begin{aligned} k_1 &= f(t_n, x_n), \\ k_2 &= f\left(t_n + \frac{1}{2}h, x_n + \frac{1}{2}hk_1\right), \\ k_3 &= f\left(t_n + \frac{1}{2}h, x_n + \frac{1}{2}hk_2\right), \\ k_4 &= f\left(t_n + h, x_n + hk_3\right). \end{aligned} \tag{5.10}$$

According to the Pontryagin principle, we now present the forward backward sweep method following [13]. As we say at the beginning of this chapter, this method consists on solving forward in time the dynamics given by the control system and backward in time the adjoint equations $\dot{\psi}$ with a transversality condition. We describe the steps of the forward-backward sweep method as follows

Step 1. Make an initial guess for u over the interval.

Step 2. Using the initial condition $x_1 = x(t_0) = a$ and the values for u, solve x forward in time according to its differential equation in the control system.

Step 3. Using the transversality condition $\psi_{N+1} = \psi(t_1) = 0$ and the values for u and x, solve ψ backward in time according to its differential equation.

Step 4. Update u by entering the new x and ψ values into the characterization of the optimal control.

Step 5. Check convergence. If the values of the variables in this iteration and the last iteration and the last iteration are negligibly close, output the current values as solutions. If values are not close, return to Step 2.

Using Python, we made an implementation of this method [17], which follows the Algorithm 1.

Algorithm 1 Forward Backward Sweep

INPUT: $t_0, t_f, n_{max}, x_0, h, a, r, m, \epsilon, \psi_f$
OUTPUT: x^*, u^*, ψ

procedure FORWARD BACKWARD SWEEP($g, \psi_{\text{function}}, u, x_0, \psi_f, h, n_{max}$)
  while test $> \epsilon$ **do**
    $u_{\text{old}} \leftarrow u$
    $x_{\text{old}} \leftarrow x$
    $x \leftarrow$ RUNGE_KUTTA_FORWARD(f, u, x_0, h, n_{max})
    $\psi_{\text{old}} \leftarrow \psi$
    $\psi \leftarrow$ RUNGE_KUTTA_BACKWARD($\psi_{\text{function}}, x, \psi_f, h, n_{max}$)
    $u_1 \leftarrow$ OPTIMALITY_CONDITION(u, x, ψ)
    $u \leftarrow \dfrac{u_1 + u_{old}}{2}$
    $test_1 \leftarrow \dfrac{||u - u_{\text{old}}||}{||u||}$
    $test_2 \leftarrow \dfrac{||x - x_{\text{old}}||}{||x||}$
    $test_3 \leftarrow \dfrac{||\psi - \psi_{\text{old}}||}{||\psi||}$
    test $\leftarrow \max\{test_1, test_2, test_3\}$
  return x^*, u^*, ψ ▷ Optimal pair

To fix ideas we present the following example. Consider the maximization problem below

$$\begin{aligned} &\max_u \int_0^1 Ax(t) - B(u(t))^2 dt, \\ \text{subject to } &x'(t) = -\frac{1}{2}(x(t))^2 + Cu(t), \quad x(0) = x_0 > -2, \\ &A \geq 0, \ B > 0, \end{aligned}$$

for $t \in [0, 1]$. By (4.34), the Hamiltonian is

$$H(t, x, u, \psi) = Ax(t) - Bu(t)^2 - \frac{1}{2}\psi(t)x(t)^2 + C\psi(t)u(t).$$

Using the optimality condition,

$$\frac{\partial H}{\partial u} = -2Bu(t) + C\psi(t) = 0,$$

at $\bar{u}(t)$, we have that $\bar{u}(t) = \frac{C\psi(t)}{2B}$. Now, by the definition of the adjoint equation (4.35), $\psi'(t) = -(g_x(t, x, u) + \psi(t) f_x(t, x, u))$. Thus, the problem to solve, using the Algorithm 1,

is

$$x'(t) = -\frac{1}{2}x(t)^2 + Cu(t), \quad x(0) = x_0,$$
$$\psi'(t) = -A + x(t)\psi(t),$$

with the transversality condition $\psi(1) = 0$.

Algorithm 2 Evolutionary Algorithms

$Y \leftarrow \mathbf{Y}_0(Np, \mathcal{V})$
while (the stopping criterion has not been met) **do**
  $M \leftarrow \mathbf{M}(Y)$
  $C \leftarrow \mathbf{C}(Y, C)$
  $Y \leftarrow \mathbf{S}(Y, C, fob)$
$\mathbf{y}_{best} \leftarrow \mathbf{Best}(Y, fob)$

Chapter 6

Applications of Optimal Control Problems

The objective of this chapter is to present applications in biology to fix the ideas established in previous chapters. We follow the Labs [14] presented in [13] but with our Python implementations [17]. We present multidimensional problems with one and two controls.

6.1 Chemotherapy for the HIV

The Human Immunodeficiency Virus (HIV) is a condition that targets the immune system and weakens people's defense system against other infections, such as the tuberculosis and some types of cancer. At this moment the HIV has no cure and that's the reason it is one of the biggest problems in public health. However, there exist treatments to control this sickness, like drugs or chemotherapy that try to suppress the infectivity of the virus.

We consider a model for HIV reported in [5] which describes the interaction between the immune system and the HIV virus. Let $T(t)$ be the concentration of uninfected $CD4^+T$ cells and $T_i(t)$ the infected $CD4^+T$ cells. Let $V(t)$ be the concentration of free

infectious virus particles. The dynamic of this problem is

$$\begin{aligned}
\frac{dT}{dt} &= \frac{s}{1+V} - \mu_1 T + rT\left(1 - \frac{T+T_i}{T_{max}}\right) - kVT,\\
\frac{dT_i}{dt} &= kVT - \mu_2 T_i,\\
\frac{dV}{dt} &= N\mu_2 T_i - \mu_3 V,
\end{aligned}$$

with initial conditions $T(0) = T_0$, $T_i(0) = T_i^0$ and $V(0) = V_0$. The term $\frac{s}{1+V(t)}$ represents the rate of generation of new $CD4^+T$ cells. We consider r as the growth rate of T cells per day. This growth is assumed to be logistic, with a maximum level T_{max}. The term kVT models the rate that free virus V infects $CD4^+T$ cells. Once the infection occurs, replication of the virus is initiated, then N represents the average number of virus particles produced before the host cell dies. The death rates of uninfected $CD4^+$ cells T, infected $CD4^+$ cells T_i and free virus particles V are μ_1, μ_2 and μ_3, respectively. Note that

$$\begin{aligned}
\left.\frac{dT}{dt}\right|_{T=0} &= \frac{s}{1+V} \geq 0,\\
\left.\frac{dT_i}{dt}\right|_{T_i=0} &= kVT \geq 0,\\
\left.\frac{dV}{dt}\right|_{V=0} &= N\mu_2 T_i \geq 0.
\end{aligned}$$

In this example, we consider as treatment the chemotherapy of reverse transcriptase inhibitors, like azidothymidine (AZT), which affects the infectivity of the virus. The term $1 - u(t)$ with $0 \leq u(t) \leq 1$ represents the strength of the chemotherapy. That is, if the control $u(t) = 0$, then we have the maximal use of chemotherapy. In the other case, when $u(t) = 1$ there is no chemotherapy. Thus, we consider the following control problem:

$$\max_u J(u) = \max_u \int_0^{t_f} \left[AT(t) - (1-u(t))^2\right] dt,$$

subject to

$$\begin{aligned}
&\frac{dT}{dt} = \frac{s}{1+V} - \mu_1 T + rT\left(1 - \frac{T+T_i}{T_{max}}\right) - u(t)kVT,\\
&\frac{dT_i}{dt} = u(t)kVT - \mu_2 T_i,\\
&\frac{dV}{dt} = N\mu_2 T_i - \mu_3 V,\\
&T(0) = T_0, T_i(0) = T_i^0, V(0) = V_0, A \geq 0.
\end{aligned} \tag{6.1}$$

where A is a weight parameter. We want to maximize the number of T cells and minimize the 'cost' of the chemotherapy to the body.

Let $x(t) = (T(t), T_i(t), V(t))^\top$. Note that each function in the right-hand side of ODE in (6.1) and the function $AT(t) - (1-u(t))^2$ are continuously differentiable. Thus, by the Existence Theorem 3.2 there is, at least, one optimal pair $(\bar{u}, \bar{x})$. Moreover, we can apply the Pontryagin Maximum Principle 4.3. According to the definition of the Hamiltonian 4.34 we have that

$$H(t, x(t), u(t), \psi(t)) = AT(t) - (1-u(t))^2 + \big\langle \psi(t), f(t, x(t), u(t)) \big\rangle,$$

where f represents the right-hand side of the ODE in the control problem (6.1). Then, according to the definition of the adjoint equation (4.35), we obtain

$$\begin{aligned}
\dot{\psi}_T &= -A + \psi_T \left[\mu_1 - r\left(1 - \frac{T_i}{T_{max}}\right)\right] - \psi_{T_i} ukV \\
\dot{\psi}_{T_i} &= \psi_T \frac{rT}{T_{max}} + \psi_{T_i}\mu_2 - \psi_V N \mu_2 \\
\dot{\psi}_V &= \psi_T \left(\frac{s}{(1+V)^2} + ukT \right) - \psi_{T_i} ukT + \psi_V \mu_3.
\end{aligned}$$

with transversality conditions $\psi_T(t_f) = 0$, $\psi_{T_i}(t_f) = 0$ and $\psi_V(t_f) = 0$. From the optimality condition (4.36) we have that

$$\frac{\partial H}{\partial u}(\bar{u}) = 2(1-\bar{u}) + (\psi_{T_i} - \psi_T) kVT = 0$$

Thus,

$$\bar{u} = 1 + \frac{(\psi_{T_i} - \psi_T) kVT}{2}.$$

Since the control is bounded, the optimality condition reads

$$\bar{u} = \min\left\{ \max\left\{ 0, 1 + \frac{(\psi_{T_i} - \psi_T) kVT}{2} \right\}, 1 \right\}.$$

For the simulations, we consider the parameters from [5] which are presented in the following table.

For these initial conditions we can observe, in Figure 6.1, that the control suggests to apply the strongest dose of chemotherapy in the first 15 days. Then, between days 15 and 20 of treatment we have to critically reduce the chemotherapy dose to the half. From this moment we have to gradually decrease the dose. With this schedule of treatment,

Parameters	**Values**		
Death rate of T cells	μ_1	0.02	days^{-1}
Death rate of T_i cells	μ_2	0.2	days^{-1}
Death rate of V cells	μ_3	4.4	days^{-1}
Infection rate	k	2.4×10^{-5}	$\text{mm}^3\text{days}^{-1}$
Growth rate of T cells	r	0.03	days^{-1}
Average number of virus particles produced	N	300	
Maximum growth level	$T_{\max}$	1500	mm^{-3}
	s	10	$\text{mm}^{-3}\text{days}^{-1}$
	A	0.2	
	$T(0)$	806.4	mm^{-3}
	$T_i(0)$	0.04	mm^{-3}
	$V(0)$	1.5	mm^{-3}

TABLE 6.1: Values for the parameters and initial conditions.

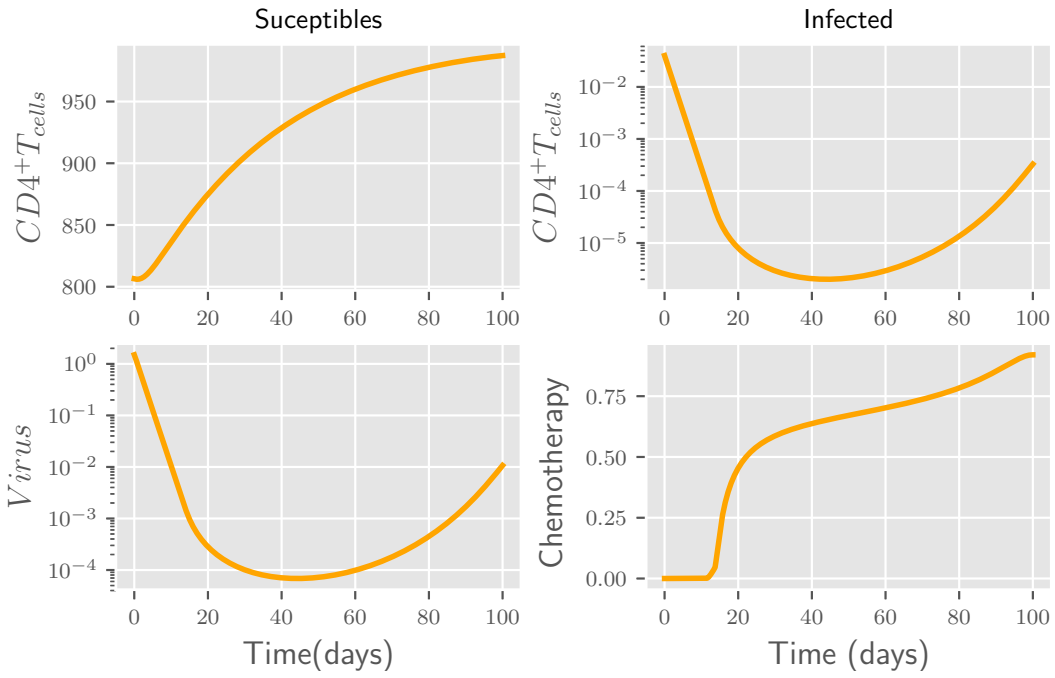


FIGURE 6.1: The horizontal axis represents time t on days. The vertical axis represents, in each case, the states susceptible T_{cells}, infected T_{cells}, the Virus and the chemotherapy, control u.

we see that the virus population and the infected cells decreases in the first 20 days and stays at a low level for a few days.

6.2 Multidrug-resistant Tuberculosis

The Tuberculosis (TB) is a disease caused by a bacteria called *Mycobacterium tuberculosis* affecting, principally, the lungs. This disease is one of the top 10 causes of death worldwide and is a leading killer of HIV-positive people. Fortunately, it is a treatable and curable disease. But, failures in treatment such as inappropriate use of medicines, can cause a drug resistance to TB. So, the Multidrug-resistant tuberculosis (MDR-TB) is a form of TB caused by bacteria that do not respond to isoniazid and rifampicin, the two most powerful anti-TB drugs. The MDR-TB is still curable, but is expensive and requires chemotherapy.

We consider the following class populations, the variable S represent the susceptible individuals, L_1 and L_2 represent the latent class for the TB and MDR-TB population, respectively. The infectioius population are I_1 for the TB and I_2 for the MDR-TB. Finally, we consider T as the class of effectively treated population. The following is an MDR-TB model based on [7]

$$\begin{aligned}
\dot{S} &= \Lambda - \beta_1 S\frac{I_1}{N} - \beta_3 S\frac{I_2}{N} - \mu S \\
\dot{L}_1 &= \beta_1 S\frac{I_1}{N} - (\mu + k_1 + r_1)L_1 + pr_2 I_1 + \beta_2 T\frac{I_1}{N} - \beta_3 L_1\frac{I_2}{N} \\
\dot{I}_1 &= k_1 L_1 - (\mu + d_1 + r_2)I_1 \\
\dot{L}_2 &= qr_2 I_1 - (\mu + k_2)L_2 + \beta_3(S + L_1 + T)\frac{I_2}{N} \\
\dot{I}_2 &= k_2 L_2 - (\mu + d_2)I_2 \\
\dot{T} &= r_1 L_1 + (1 - (p + q))r_2 I_1 - \beta_2 T\frac{I_1}{N} - \beta_3 T\frac{I_2}{N},
\end{aligned}$$

where the parameters are describe in Section 6.2. For this model we get the basic reproduction number considering the disease-free equilibrium $x_0 = \left(\frac{\Lambda}{\mu}, 0, 0, 0, 0, 0\right)$ and the functions $\mathscr{F}$ and $\mathscr{V}$:

$$\mathscr{F} = \begin{pmatrix} \beta_1 S\frac{I_1}{N} + \beta_2 T\frac{I_1}{N} + pr_2 I_1 \\ \beta_3(S + L_1 + T)\frac{I_2}{N} + qr_2 I_1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

Parameters	**Description**
Λ	Recruitment rate
β_1	Probability that a susceptible individual become infected by one infectious individual per contact per unit of time.
β_2	Probability that a recovered individual become infected by one infectious individual per contact per unit of time.
β_3	Probability that uninfected individuals become infected by one resistant-TB infectious individual per contact per unit of time.
k_1	Rate at which an individual leaves the latent class of TB by becoming infectious.
k_2	Rate at which an individual leaves the latent class of MDR-TB by becoming infectious.
μ	Per-capita natural death rate.
d_1	Per-capita disease induced death rate for TB.
d_2	Per-capita disease induced death rate for MDR-TB.
r_1	Treatment rate of individuals with latent TB.
r_2	Treatment rate of individuals with infectious TB.
$p+q$	Proportion of treated infectious individuals that did not complete their treatment.

TABLE 6.2: Description of parameters for the MDR-TB model

$$\mathscr{V} = \begin{pmatrix} (\mu+k_1+r_1)L_1 + \beta_3 L_1 \frac{I_2}{N} \\ (\mu+k_2)L_2 \\ -k_1 L_1 + (\mu+d_1+r_2)I_1 \\ -k_2 L_2 + (\mu+d_2)I_2 \\ -\Lambda + \beta_1 S\frac{I_1}{N} + \beta_3 S\frac{I_2}{N} + \mu S \\ -r_1 L_1 + (p+q-1)r_2 I_1 + \beta_2 T\frac{I_1}{N} + \beta_3 T\frac{I_2}{N} \end{pmatrix}.$$

With this functions we construct the next generation matrix on x_0:

$$FV^{-1}(x_0) = \begin{pmatrix} \frac{k_1(\beta_1+pr_2)}{(\mu+k_1+r_1)(\mu+d_1+r_2)} & 0 & \frac{\beta_1+pr_2}{\mu+d_1+r_2} & 0 \\ \frac{k_1 q r_2}{(\mu+k_1+r_1)(\mu+d_1+r_2)} & \frac{k_2\beta_3}{(\mu+k_2)(\mu+d_2)} & \frac{qr_2}{\mu+d_1+r_2} & \frac{\beta_3}{\mu+d_2} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Then, calculating the eigenvalues of $\det(FV^{-1}(x_0) - \lambda I_d)$, we get that

$$\mathscr{R}_1 = \frac{k_1(\beta_1+pr_2)}{(\mu+k_1+r_1)(\mu+d_1+r_2)}, \qquad \mathscr{R}_2 = \frac{k_2\beta_3}{(\mu+k_2)(\mu+d_2)}.$$

By the above, the basic reproduction number is $\mathscr{R}_0 = \max\{\mathscr{R}_1, \mathscr{R}_2\}$. As Castillo-Chavez

established in [7], the disease free equilibrium is stable when $\mathscr{R}_0 < 1$ and it is unstable if $\mathscr{R}_0 > 1$

We now consider a controlled version of the MDR-TB model presented above from [12], So we have the same compartments and parameters. In this case we have two controls u_1 and u_2. The control $u_1(t)$ (case finding) represents the fraction of typical TB latent individuals that are identified and put under treatment. The term $1-u_2(t)$ (case holding), represents the measures to avoid the failure of treatment. Thus the control problem is

$$\min_{u\in\Omega} J(u_1(t), u_2(t)) = \min_{u\in\Omega} \int_0^{t_f} \left[L_2(t) + I_2(t) + \frac{B_1}{2}u_1^2(t) + \frac{B_2}{2}u_2^2(t) \right] dt$$

subject to

$$\begin{aligned}
&\dot{S} = \Lambda - \beta_1 S\frac{I_1}{N} - \beta_3 S\frac{I_2}{N} - \mu S\\
&\dot{L}_1 = \beta_1 S\frac{I_1}{N} - (\mu + k_1 + u_1(t)r_1)L_1 + (1-u_2(t))pr_2 I_1 + \beta_2 T\frac{I_1}{N} - \beta_3 L_1\frac{I_2}{N}\\
&\dot{I}_1 = k_1 L_1 - (\mu + d_1 + r_2)I_1\\
&\dot{L}_2 = (1-u_2(t))qr_2 I_1 - (\mu + k_2)L_2 + \beta_3(S + L_1 + T)\frac{I_2}{N}\\
&\dot{I}_2 = k_2 L_2 - (\mu + d_2)I_2\\
&\dot{T} = u_1(t)r_1 L_1 + (1 - (1-u_2(t))(p+q))r_2 I_1 - \beta_2 T\frac{I_1}{N} - \beta_3 T\frac{I_2}{N}\\
&S(0) = S_0, L_1(0) = L_1^0, I_1(0) = I_1^0, L_2(0) = L_2^0, I_2(0) = I_2^0, T(0) = T_0,\\
&B_1, B_2 \geq 0.
\end{aligned} \tag{6.2}$$

where $\Omega = \{(u_1, u_2) \in L^1(0, t_f) | a_i \leq u_i \leq b_i, \}$ with a_i, b_i fixed positive constants. The terms B_1 and B_2 represent the balancing cost factors.

Let $x = (S, L_1, I_1, L_2, I_2, T)$ and let f denote the right-hand side of the control system the optimal control problem 6.2. Since the cost functional and f are continuously differentiable we can apply the Existence Theorem 3.2 and the Pontryagin Principle 4.3. That is, there is a optimal pair $(\bar{u}, \bar{x})$ that minimizes the cost functional and we can write the Hamiltonian as

$$H = L_2 + I_2 + \frac{B_1}{2}u_1^2 + \frac{B_2}{2}u_2^2 + \langle \psi, f \rangle.$$

Moreover, there exist adjoint functions 13 $\psi_1(t), \ldots, \psi_6(t)$ such that

$$\begin{aligned}
\dot{\psi_1} &= \psi_1\left(\beta_1\frac{I_1}{N} + \beta_3\frac{I_2}{N} + \mu\right) - \psi_2\beta_1\frac{I_1}{N} - \psi_4\beta_3\frac{I_2}{N} \\
\dot{\psi_2} &= \psi_2\left(\mu + k_1 + u_1r_1 + \beta_3\frac{I_2}{N}\right) - \psi_3k_1 - \psi_4\beta_3\frac{I_2}{N} - \psi_6\left(u_1r_1\right) \\
\dot{\psi_3} &= \psi_1\beta_1\frac{S}{N} - \psi_2\left(\beta_1\frac{S}{N} + (1-u_2)pr_2 + \beta_2\frac{T}{N}\right) + \psi_3\left(\mu + d_1 + r_2\right) - \psi_4\left(1-u_2\right)qr_2 \\
&\quad - \psi_6\left((1-(1-u_2)(p+q))r_2 - \beta_2\frac{T}{N}\right) \\
\dot{\psi_4} &= -1 + \psi_4\left(\mu + k_2\right) - \psi_5k_2 \\
\dot{\psi_5} &= -1 + \psi_1\beta_3\frac{S}{N} + \psi_2\beta_3\frac{L_1}{N} - \psi_4\beta_3\frac{S+L_1+T}{N} + \psi_5\left(\mu + d_2\right) + \psi_6\beta_3\frac{T}{N} \\
\dot{\psi_6} &= -\psi_2\beta_2\frac{I_1}{N} - \psi_4\beta_3\frac{I_2}{N} - \psi_6\left(\beta_2\frac{I_1}{N} + \beta_3\frac{I_2}{N} + \mu\right).
\end{aligned}$$

with $\psi_i(t_f) = 0$, for each $i = 1, \ldots, 6$. According to the definition of the optimality condition we have that

$$\frac{\partial H}{\partial u_1}(\bar{u}_1) = B_1\bar{u}_1 - \psi_2r_1L_1 + \psi_6r_1L_1 = 0$$

and

$$\frac{\partial H}{\partial u_2}(\bar{u}_2) = B_2\bar{u}_2 - \psi_4qr_2I_1 + \psi_6(p+q)r_2I_1 = 0$$

Thus, the optimal control is given by

$$\bar{u}_1 = \min\left\{\max\left\{a_1, \frac{1}{B_1}(\psi_2 - \psi_6)r_1L_1\right\}, b_1\right\}$$

and

$$\bar{u}_2 = \min\left\{\max\left\{a_2, \frac{1}{B_1}\left(\psi_2p + \psi_4q - \psi_6(p+q)r_2I_1\right).\right\}, b_2\right\}$$

In Figure 6.2 we can observe that the infected population without these countermeasures is growing linearly, in contrast with the controlled model the infected population remains almost constant in the beginning. After 3 years the controlled MDR-TB shows that the infected population grows faster but stays low.

Parameters	**Values**
β_1	13
β_2	13
β_3	0.0131, 0.0217, 0.029, 0.0436
μ	0.0143
d_1	0
d_2	0
k_1	0.5
k_2	1
r_1	2
r_2	1
p	0.4
q	0.1
N	6000, 12000, 30000
Λ	μN
t_f	5 years
B_1	50
B_2	500
Lower bound for controls	0.05
Upper bound for controls	0.95

TABLE 6.3: Values of the parameters

States	**Values**
$S(0)$	$(76/120)N$
$L_1(0)$	$(36/120)N$
$I_1(0)$	$(4/120)N$
$L_2(0)$	$(2/120)N$
$I_2(0)$	$(1/120)N$
$T(0)$	$(1/120)N$

TABLE 6.4: Initial conditions

6.3 Quarantine and Isolation for the SARS

The Severe Acute Respiratory Syndrome SARS is a viral disease, highly contagious. This disease emerged in China in 2002 and has quickly spread over the years. The main problem with this sickness is that there's no vaccine or medicine to fight it. So the principal measure to fight the disease, is to control the spread of it. The two measures to control it are isolation of the population who presents symptoms, and quarantine for those are asymptomatic but has been in contact with the disease. This measure reduce the contact with the infected population and so the SARS can be controlled.

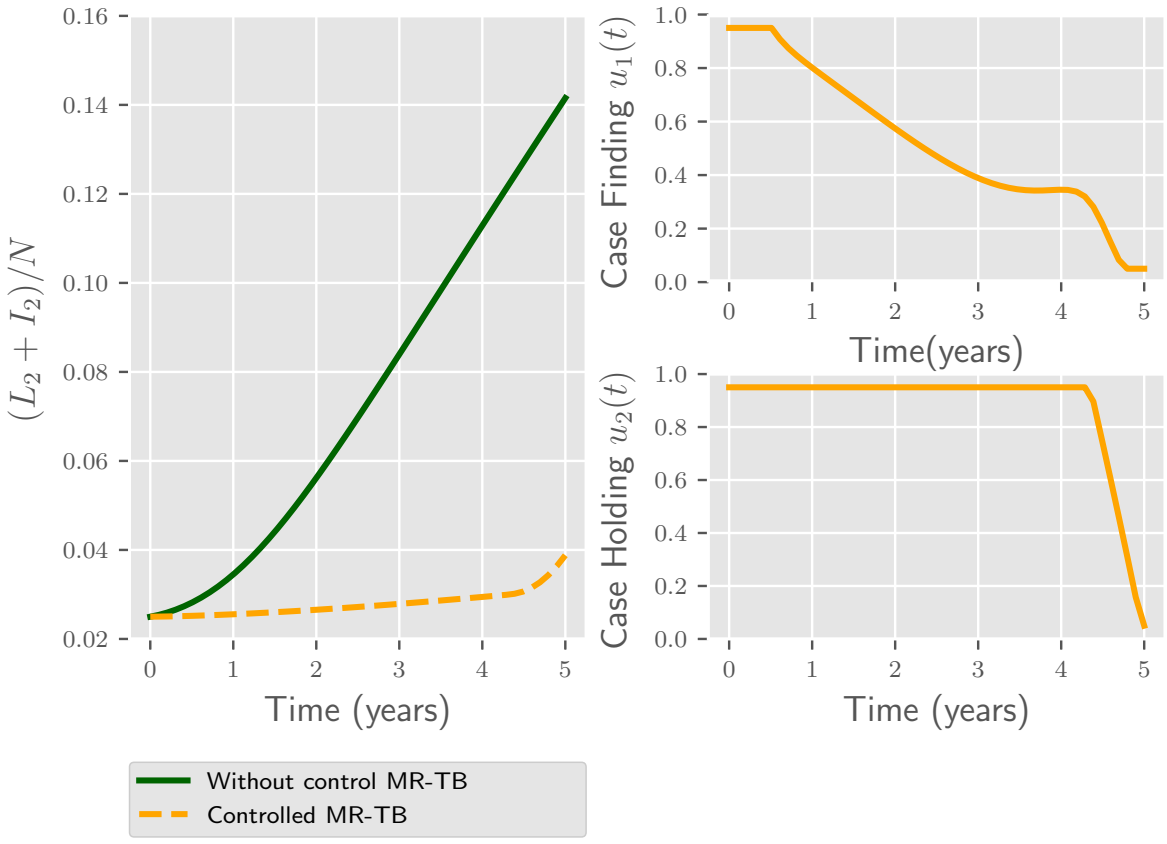


FIGURE 6.2: In the left side, the green line represents the uncotrolled state of MDR-TB infected population (I/N) and the orange dashed line represents the controlled state. In the right side, the two controls are plotted The parameters that were used are:

Here we present the dynamical model for SARS based on [21]. In this first model we present a constant control. The class S represents the susceptible individuals; E, the asymptomatic individuals who have been exposed to the virus but do not present clinical symptoms of SARS; the quarantine individuals are represented by Q; I, symptomatic individuals; J represent the isolated individuals; and R, the recovered individuals.

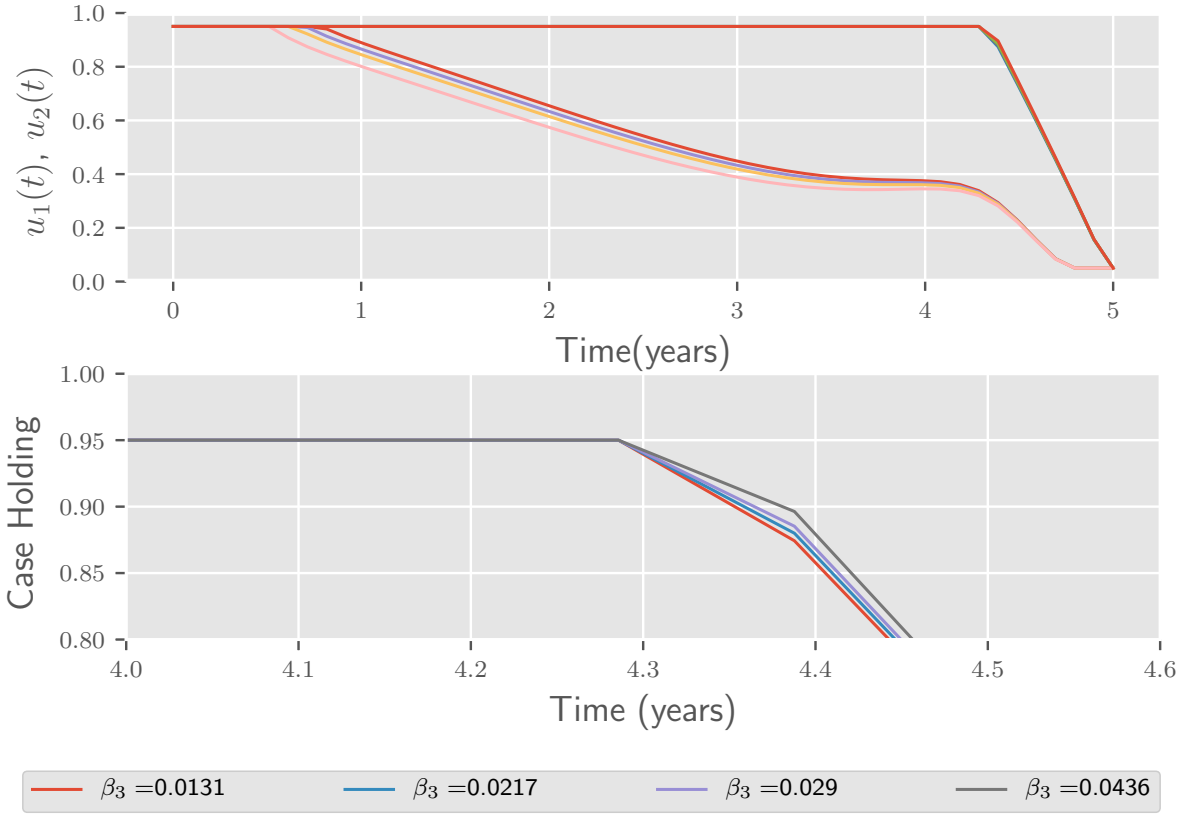


FIGURE 6.3: A comparison of the controls modifying the probability of getting infected by a resistant-TB infected.

$$
\begin{aligned}
\frac{dS}{dt} &= \Lambda - \frac{S\left(\beta I + \varepsilon_E \beta E + \varepsilon_Q \beta Q + \varepsilon_J \beta J\right)}{N} - \mu S, \\
\frac{dE}{dt} &= p + \frac{\beta S\left(\beta I + \varepsilon_E \beta E + \varepsilon_Q \beta Q + \varepsilon_J \beta J\right)}{N} - (u_1 + k_1 + \mu)E, \\
\frac{dQ}{dt} &= u_1 E - (k_2 + \mu)Q, \\
\frac{dI}{dt} &= k_1 E - (u_2 + d_1 + \sigma_1 + \mu)I, \\
\frac{dJ}{dt} &= u_2 I + k_2 Q - (d_2 + \sigma_2 + \mu)J, \\
\frac{dR}{dt} &= \sigma_1 I + \sigma_2 J - \mu R.
\end{aligned}
\tag{6.3}
$$

In the model we have the recruitment rate Λ, the natural death rate $\mu > 0$. A net inflow of asymptomatic individuals into the region at a rate p per unit of time. This parameter includes new births, immigration and emigration. We set p to zero for simplicity. The transmission coefficients for these four classes of infected individuals (I, E, Q, J) are β, $\varepsilon_E\beta$, $\varepsilon_Q\beta$ and $\varepsilon_J\beta$, respectively. An asymptomatic individual is transferred into the

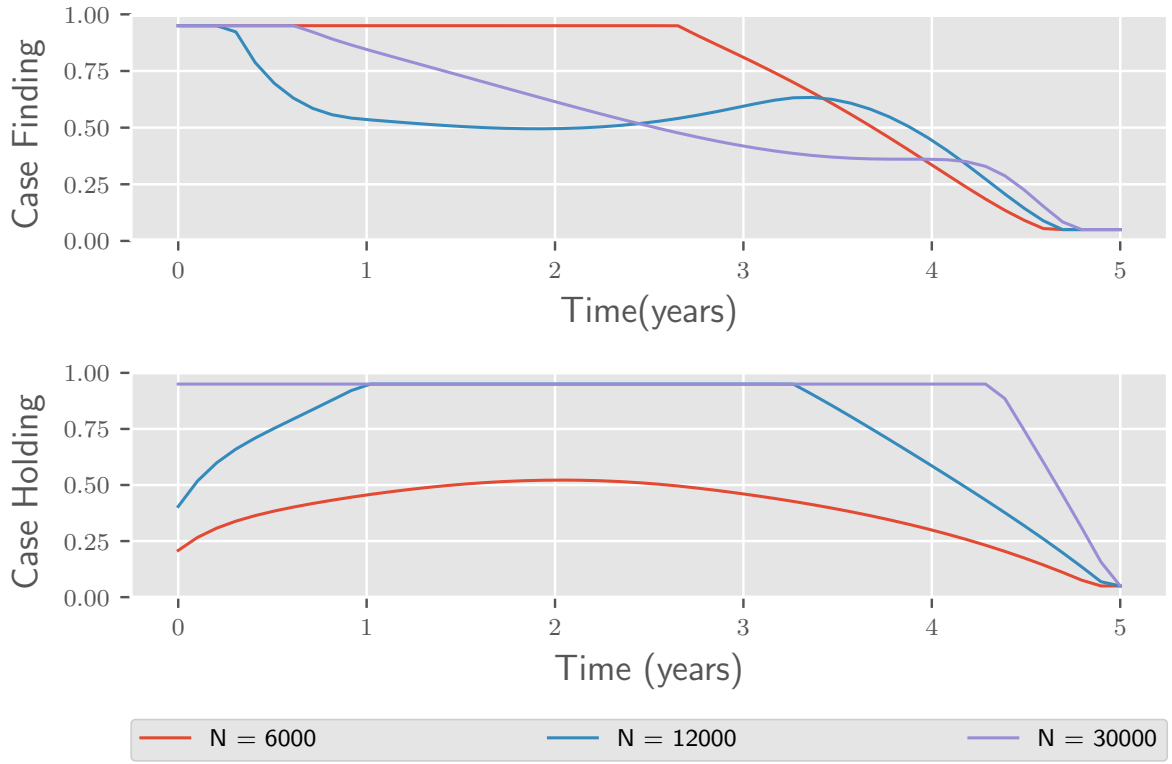


FIGURE 6.4: A comparison of the controls modifying the size of the population.

symptomatic class at a rate k_1 and a quarantined individual is transferred into the isolated class at a rate k_2. The per-capita death rates induced by the disease are d_1 and d_2. The per-capita recovery rates of symptomatic and isolated individuals are σ_1 and σ_2, respectively. The control parameter u_1 represents the rate of quarantining of people who have been in contact with an infected individual by a quarantine program. The control parameter u_2 represents the rate of isolating of symptomatic individuals by an isolation program.

Now, we suppose that the controls are not constant, this mean that the rate of quarantining $u_1(t)$ and the rate of isolation $u_2(t)$ can change in different moments of time. Define $\Omega := \{(u_1, u_2) \in L^1(0, t_f) \, : \, a_i \leq u_i \leq b_i, \, i = 1, 2\}$ with a_i, b_i fixed positive constants. We consider the following control problem

$$\min_{u\in\Omega}\int_0^{t_f}\left[B_1E(t)+B_2Q(t)+B_3I(t)+B_4J(t)+\frac{C_1}{2}u_1^2(t)+\frac{C_2}{2}u_2^2(t)\right]dt.$$

subject to

$$\begin{aligned}
\frac{dS}{dt} &= \Lambda-\frac{S\left(\beta I+\varepsilon_E\beta E+\varepsilon_Q\beta Q+\varepsilon_J\beta J\right)}{N}-\mu S,\\
\frac{dE}{dt} &= p+\frac{\beta S\left(\beta I+\varepsilon_E\beta E+\varepsilon_Q\beta Q+\varepsilon_J\beta J\right)}{N}-(u_1(t)+k_1+\mu)E,\\
\frac{dQ}{dt} &= u_1(t)E-(k_2+\mu)Q,\\
\frac{dI}{dt} &= k_1E-(u_2(t)+d_1+\sigma_1+\mu)I,\\
\frac{dJ}{dt} &= u_2(t)I+k_2Q-(d_2+\sigma_2+\mu)J,\\
\frac{dR}{dt} &= \sigma_1I+\sigma_2J-\mu R,\\
S(0)=S_0, E(0)=E_0, Q(0)=Q_0, I(0)&=I_0, J(0)=J_0, R(0)=R_0\\
B_1,B_2,B_3,B_4,C_1,C_2 &\geq 0
\end{aligned}\tag{6.4}$$

where the coefficients B_1, B_2, B_3, B_4, C_1 and C_2 are balancing cost factors due to size and importance of each term in the cost functional.

Since the right-hand side of the ODE in (6.4) and integrand in the cost functional are continuously differentiable we can assure the existence of an optimal pair $(\bar{u},\bar{x})$, where $x(t)=(S(t),E(t),Q(t),I(t),J(t))$ and we can apply the Pontryagin's Minimum Principle 4.3. According to the definition of the Hamiltonian, we have

$$H=B_1E(t)+B_2Q(t)+B_3I(t)+B_4J(t)+\frac{C_1}{2}u_1^2(t)+\frac{C_2}{2}u_2^2(t)+\langle\psi(t),f(t,x(t),u(t)\rangle \tag{6.5}$$

where f represents the right-hand of the control system in (6.4). From this principle we also obtain the adjoint equations

$$\frac{d\psi_i}{dt}=-\frac{\partial H}{\partial x_i},\quad \psi_i(t_f)=0. \tag{6.6}$$

with x_i representing the i-th state variable and the optimality condition

$$\frac{\partial H}{\partial u_i}(\bar{u}_i)=0. \tag{6.7}$$

By the equation (6.6) we get

$$\begin{aligned}
\frac{d\psi_1}{dt} &= \psi_1\left(\frac{\beta I + \varepsilon_E\beta E + \varepsilon_Q\beta Q + \varepsilon_J\beta J}{N} + \mu\right) - \frac{\psi_2(\beta I + \varepsilon_E\beta E + \varepsilon_Q\beta Q + \varepsilon_J\beta J}{N},\\
\frac{d\psi_2}{dt} &= -B_1 + \psi_1\frac{\varepsilon_E\beta}{N}S - \psi_2\left(\frac{\varepsilon_E\beta}{N}S - (u_1(t) + k_1 + \mu)\right) - \psi_3 u_1(t) - \psi_4 k_1,\\
\frac{d\psi_3}{dt} &= -B_2 + \psi_1\frac{\varepsilon_Q\beta}{N}S - \psi_2\frac{\varepsilon_Q\beta}{N}S + \psi_3(k_2 + \mu) - \psi_5 k_2,\\
\frac{d\psi_4}{dt} &= -B_3 + \psi_1\frac{\beta}{N}S - \psi_2\frac{\beta}{N}S + \psi_4(u_2(t) + d_1 + \sigma_1 + \mu) - \psi_5 u_2(t) - \psi_6\sigma_1,\\
\frac{d\psi_5}{dt} &= -B_4 + \psi_1\frac{\varepsilon_J\beta}{N}S - \psi_2\frac{\varepsilon_J\beta}{N}S + \psi_5(d_2 + \sigma_2 + \mu) - \psi_6\sigma_2,\\
\frac{d\psi_6}{dt} &= \psi_6\mu.
\end{aligned}$$

By the optimality condition (6.7) we obtain

$$\begin{aligned}
C_1u_1(t) - \psi_2E + \psi_3E = 0,\\
C_2u_2(t) - \psi_4I + \psi_5I = 0.
\end{aligned}$$

Since the controls u_1 and u_2 are bounded, the optimality condition yields

$$\begin{aligned}
u_1(t) &= \min\left\{\max\left\{a_1, \frac{1}{C_1}(\psi_2 - \psi_3)E\right\}, b_1\right\},\\
u_2(t) &= \min\left\{\max\left\{a_2, \frac{1}{C_2}(\psi_4 - \psi_5)I\right\}, b_2\right\}.
\end{aligned}$$

Figure 6.5 suggest to quarantine and isolate as many as possible of the asymptomatic in the first 170 days and the symptomatic population in the first 50 days. This in order to minimize the infected population. For the asymptomatic population, after the 170th day the quarantine control has to steadily decrease, this means that the recruitment of the asymptomatic population has to be less through the pass of the days. The same idea is applied to the isolation control, but in this case, after the 50th day the decrease has to be bigger in comparison to the quarantine case. With this schedule we can observe in Figure 6.6 that the infected population formed by the asymptomatic and symptomatic population decreases rapidly after the 10th day until the day 150. After this point, the population starts to decrease steadily. If we compare to dynamics without control, we see that the infected population declines slowly after the day 20.

Parameters	**Description**	**Values**
β	0.2	Transmission coefficient
ε_E, ε_Q, ε_J	0.3, 0, 0.1	Modification parameter for exposed, quarantine and isolation classes.
μ	0.000034	Natural death rate.
Λ	408.09	Constant recruitment rate .
p	0	Net inflow of asymptomatic individuals .
k_1	0.1	Transfer rate from class of asymptomatic to symptomatic.
k_2	0.125	Transfer rate from the quarantine class to isolation.
d_1, d_2	0.0079, 0.0068	Per-capita disease induced death rates for the symptomatic individuals and isolated individuals.
σ_1, σ_2	0.0337, 0.0386	Per-capita recovery rates for the symptomatic individuals and isolated individuals
t_f	365 days	Final time
B_1, B_2, B_3, B_4	1	Respectively cost for E,Q,I,J classes.
C_1, C_2	300, 600	Costs for Isolation and Quarantine policies.
a_i, b_i	0.05, 0.5	Bounds for the each control.

TABLE 6.5: Parameter description and values for the SARS model (6.3).

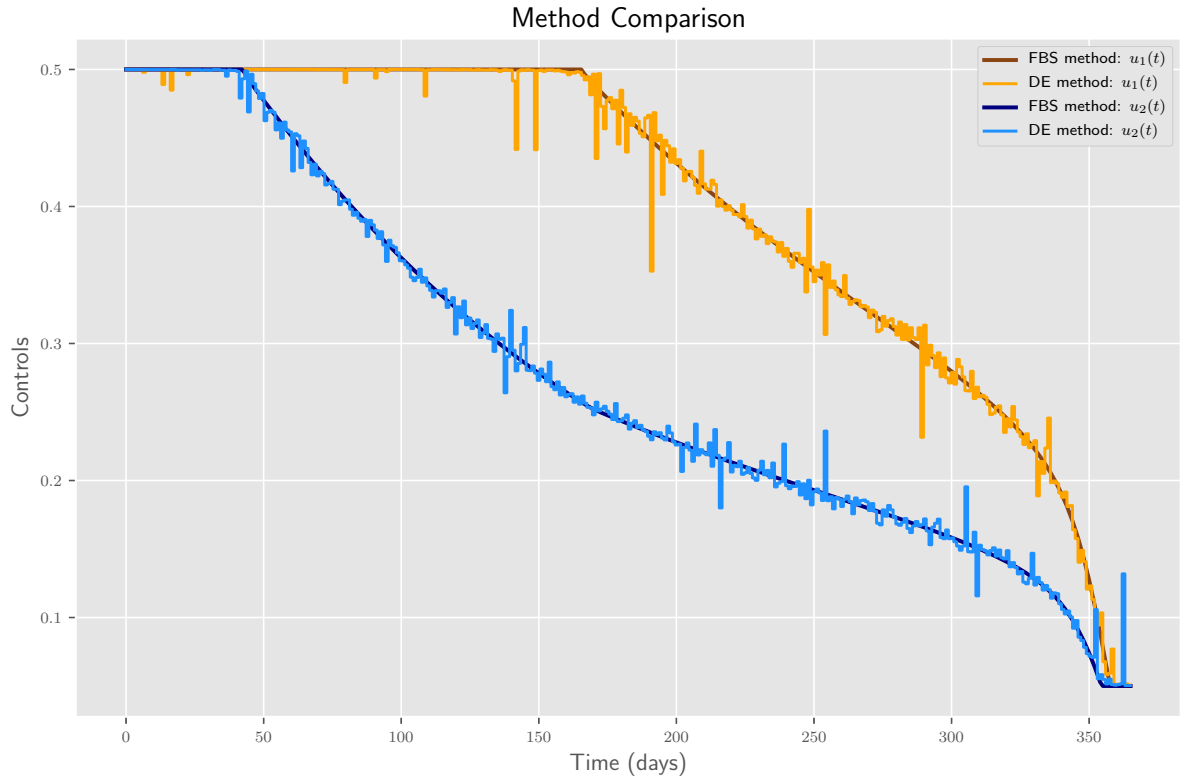


FIGURE 6.5: Comparison of the optimal controls obtained by the forward-backward sweep method and sub-optimal controls obtained by the differential evolution method. The initial values are $S_0 = 12$ million, $E_0 = 1565$, $Q_0 = 292$, $I_0 = 695$, $J_0 = 326$, 20.

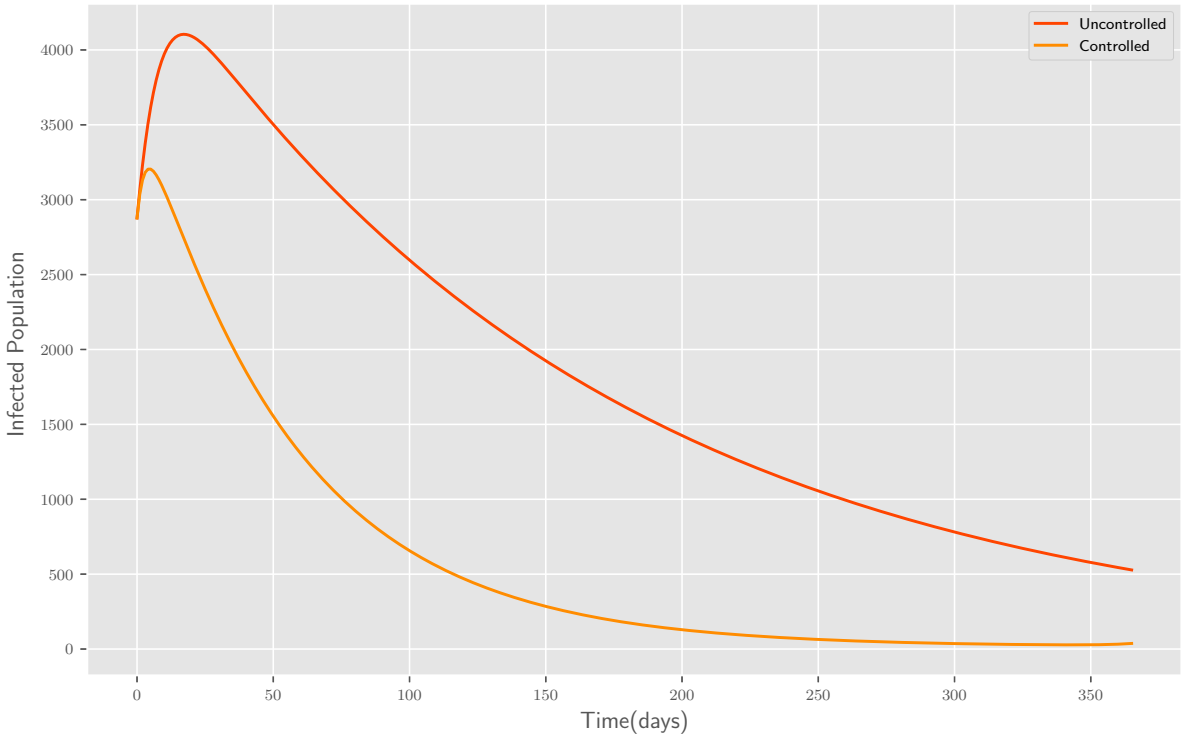


FIGURE 6.6: Comparison of the dynamics between the controlled problem and the uncontrolled. The initial values are $S_0 = 12$ million, $E_0 = 1565$, $Q_0 = 292$, $I_0 = 695$, $J_0 = 326$, 20.

Chapter 7

Conclusions and perspectives

We have reviewed the proof related to the so called Pontryagin's Maximum Principle. Appealing to the Ekeland's variational principle and other auxiliary results we have clarified most of the details of the proof of this seminal principle. According to the Pontryagin principle and its proof we have understood the forward-backward sweep method. Consequently, we made a GitHub repository [17], with the Python implementation code that approximates the solution of optimally controlled biological models reported in the literature.

Following the ideas of Suzanne Lenhart [13], we have presented controlled models reported in literature. In each of them we explained the formulation with linear control and, using the forward backward sweep method, we optimized each functional. In the completion of this thesis we have detected other ways to optimize the underlying functional cost. One alternative that we have explored is the so-called differential evolution optimization method, [17].

All the above examples of the optimal control theory involves open-loop controls. These kind of models work under the following assumptions. i) The model is perfect, ii) there is no disturbance and, iii) the parameters and inputs are known accurately. However these assumptions are unrealistic. Some authors report that closed-loop controls would be more realistic. Although, it is difficult to obtain optimal closed-loop controls for nonlinear systems, even so there is a way to do it using the Bellman's Equation.

With this in mind, we note two main ways to extend this work. The first one is to study the closed-loop controls and its applications. The latter is to change the kind of

dynamics —discrete or continuous, deterministic or stochastic. Each dynamic has its own theory and provides tools for a great spectre of applications.

Appendices

Appendix A

Auxiliary results

Theorem A.1 (Taylor's Theorem, p. 359 [16]). *Let $f : A \to \mathbb{R}$ be of class C^r for $A \subseteq \mathbb{R}^n$ an open set. Let $x, y \in A$, ans suppose that the segment joining x and y lies in A. Then there is a point c on that segment such that*

$$f(y) - f(x) = \sum_{k=1}^{r-1} \frac{1}{k!} \mathbf{D}^k f(x)(y-x, \ldots, y-x) + \frac{1}{r!} \mathbf{D}^r f(c)(y-x, \ldots, y-x),$$

where $\mathbf{D}^k f(x)(y-x, \ldots, y-x)$ denotes $\mathbf{D}^k f(x)$ as a k-linear map applied to the k-tuple $(y-x, \ldots, y-x)$. In coordinates,

$$\mathbf{D}^k f(x)(y-x, \ldots, y-x) = \sum_{i_1, \ldots, i_k = 1} \left(\frac{\partial^k f}{\partial x_{i_1} \cdots \partial x_{i_k}} \right) (y_{i_1} - x_{i_1}) \cdots (y_{i_k} - x_{i_k}).$$

Setting $y = x + h$, we can write the Taylor formula as

$$f(x+h) = f(x) + \mathbf{D}f(x) \cdot h + \cdots + \frac{1}{(r-1)!} \mathbf{D}^{r-1} f(x) \cdot (h, \ldots, h) + R_{r-1}(x, h),$$

where $R_{r-1}(x, y)$ is the remainder. Furthermore,

$$\frac{R_{r-1}(x, h)}{\|h\|^{r-1}} \to 0 \text{ as } h \to 0.$$

Theorem A.2 (Arzela-Ascoli, [22, Thm.1.4.2]). *Let $\mathcal{Z} \subseteq C([t_0, T]; \mathbb{R}^n)$ be an infinite set which is uniformly bounded*

$$\sup_{\varphi(\cdot) \in \mathcal{Z}} \|\varphi(\cdot)\|_{C([t_0,T];\mathbb{R}^n)} < \infty,$$

and equi-continuous, i.e. for any $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$|\varphi(t) - \varphi(s)| < \varepsilon, \ \forall \ |t - s| < \delta, \ \forall \varphi(\cdot) \in \mathcal{Z}$$

Then there exists a sequence $\varphi_k(\cdot) \in \mathcal{Z}$ such that

$$\lim_{k\to\infty} \|\varphi_k(\cdot) - \bar{\varphi}(\cdot)\|_{C([t_0,T];\mathbb{R}^n)} = 0,$$

for some $\bar{\varphi}(\cdot) \in C([t_0, T]; \mathbb{R}^n)$.

Proposition A.1. *Given $f \in L^p$, $1 < p \leq \infty$ and $\varepsilon > 0$, there is a step function φ and a continuous function ψ such that $\|f - \varphi\|_p < \varepsilon$ and $\|f - \psi\|_p < \varepsilon$.*

Theorem A.3 (Danskin's theorem, p.20, [11])**.** *Let $X \subseteq \mathbb{R}^n$ open and Y a compact set. Suppose that $f : X \times Y \to \mathbb{R}$ is continuous and $\nabla_x f(x, y)$ exists and is continuous. Define*

$$\varphi(x) := \min_{y\in Y}\{f(x, y)\}.$$

Then φ is continuous and the directional derivative of ϕ exists and is given by

$$D_v^+\varphi(x) = \min_{y\in Y(x)}\{\langle\nabla_x f(x, y), v\rangle\},$$

where $Y(x) = \{y \in Y \ : \ \varphi(x) = f(x, y)\}$ is the set of minimizers. If the set of minimizers has only one element, that is, $Y(x) = \{y_0\}$ then

$$D_v^+\varphi(x) = \{\langle\nabla_x f(x, y_0), v\rangle\},$$

Theorem A.4. *Consider the following initial value problem*

$$\begin{aligned}
\dot{Y}^\varepsilon(s) &= f_x(s, X^\varepsilon(s), u^\varepsilon(s))Y^\varepsilon(s) \\
&\quad + f(s, X^\varepsilon(s), u(s)) - f(s, X^\varepsilon(s), u^\varepsilon(s)), \quad s \in [0, T], \\
Y^\varepsilon(0) &= 0.
\end{aligned}$$

Then

$$\lim_{\delta\to\infty} \|Y_\delta^\varepsilon - Y^\varepsilon\| = 0.$$

Proof.

$$
\begin{aligned}
|Y_\delta^\varepsilon(s) - Y^\varepsilon(s)| &= \Big| \int_t^s \left[\int_0^1 f_x(\tau, X^\varepsilon(\tau) + \theta(X_\delta^\varepsilon(\tau) - X^\varepsilon(\tau)), u_\delta^\varepsilon(\tau)) \right] d\theta Y_\delta^\varepsilon(\tau) d\tau \\
&\quad + \int_t^s [f(\tau, X^\varepsilon(\tau), u(\tau)) - f(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau))] d\tau - \frac{r_\delta^\varepsilon(s)}{\delta} \\
&\quad - \int_t^s f_x(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau)) Y^\varepsilon(\tau) d\tau \\
&\quad - \int_t^s [f(\tau, X^\varepsilon(\tau), u(\tau)) - f(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau))] d\tau \Big| \\
&\leq \int_t^s \Big| \int_0^1 f_x(\tau, X^\varepsilon(\tau) + \theta(X_\delta^\varepsilon(\tau) - X^\varepsilon(\tau)), u_\delta^\varepsilon(\tau)) d\theta Y_\delta^\varepsilon(\tau) \\
&\quad - f_x(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau)) Y^\varepsilon(\tau) \Big| d\tau + \frac{|r_\delta^\varepsilon(s)|}{\delta}.
\end{aligned}
$$

Define $F_\delta^\varepsilon(\tau) := \int_0^1 f_x(\tau, X^\varepsilon(\tau) + \theta(X_\delta^\varepsilon(\tau) - X^\varepsilon(\tau)), u_\delta^\varepsilon(\tau)) d\theta$. Then

$$
\begin{aligned}
&\int_t^s |F_\delta^\varepsilon(\tau) Y_\delta^\varepsilon(\tau) - f_x(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau)) Y^\varepsilon(\tau)| \, d\tau \\
&\qquad \leq \int_t^s |[F_\delta^\varepsilon(\tau) - f_x(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau))] Y_\delta^\varepsilon(\tau)| \, d\tau \\
&\qquad + \int_t^s \Big| f_x(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau)) Y_\delta^\varepsilon(\tau) - f_x(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau)) Y^\varepsilon(\tau) \Big| d\tau \\
&\qquad = \int_t^s |F_\delta^\varepsilon(\tau) - f_x(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau))| \, |Y_\delta^\varepsilon(\tau)| \, d\tau \\
&\qquad + \int_t^s |f_x(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau))| \, |Y_\delta^\varepsilon(\tau) - Y^\varepsilon(\tau)| \, d\tau
\end{aligned}
$$

Now, note that

$$\begin{aligned}
&\int_t^s |F_\delta^\varepsilon(\tau) - f_x(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau))|\,|Y_\delta^\varepsilon(\tau)|\,d\tau \\
&\quad= \int_{[t,s]\cap E_\delta^\varepsilon} |F_\delta^\varepsilon(\tau) - f_x(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau))|\,|Y_\delta^\varepsilon(\tau)|\,d\tau \\
&\quad+ \int_{[t,s]\setminus E_\delta^\varepsilon} |F_\delta^\varepsilon(\tau) - f_x(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau))|\,|Y_\delta^\varepsilon(\tau)|\,d\tau \\
&\quad= \int\limits_{[t,s]\cap E_\delta^\varepsilon} \left| \int_0^1 f_x(\tau, X^\varepsilon(\tau) + \theta(X_\delta^\varepsilon(\tau) - X^\varepsilon(\tau)), u(\tau))d\theta - f_x(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau)) \right| |Y_\delta^\varepsilon(\tau)|\,d\tau \\
&\quad+ \int\limits_{[t,s]\setminus E_\delta^\varepsilon} \left| \int_0^1 f_x(\tau, X^\varepsilon(\tau) + \theta(X_\delta^\varepsilon(\tau) - X^\varepsilon(\tau)), u^\varepsilon(\tau))d\theta - f_x(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau)) \right| |Y_\delta^\varepsilon(\tau)|\,d\tau \\
&\quad\le \int_{[t,s]\cap E_\delta^\varepsilon} 2M\,|Y_\delta^\varepsilon(\tau)|\,d\tau + \int_{[t,s]\setminus E_\delta^\varepsilon} \int_0^1 L\,|\theta(X_\delta^\varepsilon(\tau) - X^\varepsilon(\tau))|\,d\theta\,|Y_\delta^\varepsilon(\tau)|\,d\tau \\
&\quad\le 2Mk(T-t)\lambda(E_\delta^\varepsilon) + \int_{[t,s]\setminus E_\delta^\varepsilon} L\,|(X_\delta^\varepsilon(\tau) - X^\varepsilon(\tau))|\,k(T-t)d\tau \\
&\quad\le K\delta
\end{aligned}$$

Then,

$$\int_t^s |F_\delta^\varepsilon(\tau)Y_\delta^\varepsilon(\tau) - f_x(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau))Y^\varepsilon(\tau)|\,d\tau \le K\delta + \int_t^s M\,|Y_\delta^\varepsilon(\tau) - Y^\varepsilon(\tau)|\,d\tau,$$

and

$$|Y_\delta^\varepsilon(\tau) - Y^\varepsilon(\tau)| \le K\delta + M\int_t^s |Y_\delta^\varepsilon(\tau) - Y^\varepsilon(\tau)|\,d\tau + \delta.$$

By Gronwall's inequality (A.2)

$$|Y_\delta^\varepsilon(\tau) - Y^\varepsilon(\tau)| \le (K+1)\delta e^{M(T-t)}, \qquad \forall\, s \in [t,T],$$

and so $\|Y_\delta^\varepsilon(s) - Y^\varepsilon(s)\| \le K\delta$, where K is a generic constant. Hence, letting $\delta \to 0$ we have that $\|Y_\delta^\varepsilon(s) - Y^\varepsilon(s)\| \to 0$.

■

Similarly

Theorem A.5. *Consider the initial value problem*

$$\begin{aligned}
&\dot{Y}(s) = f_x(s, \bar{X}(s), \bar{u}(s))^\top Y(s) + f(s, \bar{X}(s), u(s)) - f(s, \bar{X}(s), \bar{u}(s)), \\
&Y(0) = 0,
\end{aligned} \tag{A.1}$$

with $s \in [0,T]$.

$$\lim_{\varepsilon\to 0} \|Y^\varepsilon(\cdot) - Y(\cdot)\|_{C([0,T];\mathbb{R}^n)} = 0,$$

Proof. From the proof of the theorem 4.2

$$Y^\varepsilon(s) = \int_0^s f_x(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau))Y^\varepsilon(\tau)d\tau + \int_0^s \big[f(\tau, X^\varepsilon(\tau), u(\tau)) - f(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau))\big]d\tau.$$

On the other hand the solution of the I.V.P. (A.1) is

$$Y(s) = \int_0^s [f_x(\tau, \bar{X}(\tau), \bar{u}(\tau))Y(\tau) + f(\tau, \bar{X}(\tau), u(\tau)) - f(\tau, \bar{X}(\tau), \bar{u}(\tau))]d\tau.$$

Then

$$\begin{aligned}|Y^\varepsilon(s) - Y(s)| \le \int_0^s \big| & f_x(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau))Y^\varepsilon(\tau) - f_x(\tau, \bar{X}(\tau), \bar{u}(\tau))Y(\tau) \\ & + f(\tau, X^\varepsilon(\tau), u(\tau)) - f(\tau, X^\varepsilon(\tau), u^\varepsilon(\tau)) - f(\tau, \bar{X}(\tau), u(\tau)) \\ & + f(\tau, \bar{X}(\tau), \bar{u}(\tau))\big|d\tau\end{aligned}$$

■

Proposition A.2 ([22, prop.1.4.7, p. 32], Gronwall's Inequality)**.** *Let* $\theta : [a,b] \to \mathbb{R}_+$ *be continuous and satisfy*

$$\theta(s) \le \alpha(s) + \int_a^s \beta(r)\theta(r)dr, \ \ s \in [a,b],$$

for some $\alpha(\cdot), \beta(\cdot) \in L^1(a,b;\mathbb{R}_+)$*. Then*

$$\theta(s) \le \alpha(s) + \int_a^s \alpha(\tau)\beta(\tau)e^{\int_\tau^s \beta(r)dr}d\tau, \ \ s \in [a,b]. \tag{A.2}$$

In particular, if $\alpha(\cdot) = \alpha$ *is a constant, then*

$$\theta(s) \le \alpha e^{\int_a^s \beta(r)dr}, \ \ s \in [a,b]. \tag{A.3}$$

Proof. Let $\varphi(s) = \int_a^s \beta(r)\theta(r)dr.$, by the fundamental theorem of calculus, we have

$$\dot{\varphi}(s) = \beta(s)\theta(s) \le \beta(s)[\alpha(s) + \varphi(s)].$$

This leads to

$$[\varphi(s)e^{-\int_a^s \beta(r)dr}]' \le \alpha(s)\beta(s)e^{-\int_a^s \beta(r)dr}.$$

Consequently,

$$\varphi(s)e^{-\int_a^s \beta(r)dr} \leq \int_a^s \alpha(\tau)\beta(\tau)e^{-\int_a^\tau \beta(r)dr}d\tau,$$

then

$$\varphi(s) \leq \int_a^s \alpha(\tau)\beta(\tau)e^{-\int_a^\tau \beta(r)dr}e^{\int_a^s \beta(r)dr}d\tau,$$

rewriting

$$\varphi(s) \leq \int_a^s \alpha(\tau)\beta(\tau)e^{\int_\tau^s \beta(r)dr}d\tau.$$

Hence,

$$\theta(s) \leq \alpha(s) + \int_a^s \alpha(\tau)\beta(\tau)e^{\int_\tau^s \beta(r)dr}d\tau.$$

and (A.2) holds. Now, consider α constant, then

$$\theta(s) \leq \alpha + \int_a^s \alpha\beta(\tau)e^{\int_\tau^s \beta(r)dr}d\tau.$$

By integration rules, if $u = \alpha, dv = \beta(\tau)e^{\int_\tau^s \beta(r)dr}d\tau$. In the other hand,

$$\frac{d}{d\tau}e^{\int_\tau^s \beta(r)dr} = \frac{d}{d\tau}e^{-\int_s^\tau \beta(r)dr} = -\beta(\tau)e^{-\int_s^\tau \beta(r)dr} = -\beta(\tau)e^{\int_\tau^s \beta(r)dr}$$

Then

$$\theta(s) \leq \alpha - \alpha\int_a^s \frac{d}{d\tau}e^{\int_\tau^s \beta(r)dr}d\tau = \alpha - \alpha[e^0 - e^{\int_a^s \beta(r)dr}].$$

Therefore,

$$\theta(s) \leq \alpha e^{\int_a^s \beta(r)dr},$$

and (A.3) holds. ∎

Theorem A.6 ([9]). *If $f \in L^1_{loc}$, that is, f i locally integrable then*

$$\lim_{r\to 0}\frac{1}{\lambda(B(x,r))}\int_{B(x,r)} f(y)dy = f(x), \quad a.e.\ x \in \mathbb{R}^n.$$

Theorem A.7 (Interior Maximum Theorem, [2, Thm. 19.4, p. 209]). *Let c be an interior point of the domain of f, at which f has a relative maximum. If the derivative of f at c exists, then it must be equal to zero.*

Theorem A.8 (Rolle's Theorem,[2, Thm. 19.5, p. 209]). *Suppose that f is continuous on a closed interval $J = [a, b]$, that the derivative f' exists in the open interval (a, b), and that $f(a) = f(b) = 0$. Then there exists a point $c \in (a, b)$ such that $f'(c) = 0$.*

Proof. ∎

Theorem A.9 (Mean Value Theorem, [2, Thm. 19.6, p. 210]). *Suppose that f is continuous on a closed interval $J = [a, b]$ and differentiable on the open interval (a, b). Then there exists $c \in (a, b)$ such that*

$$f(b) - f(a) = f'(c)(b - a).$$

Proof. Supose that f is continuous on a closed interval $J = [a, b]$ ∎

Lemma A.1 (Fatou's Lemma, [3, Thm. 4.8, p. 33]). *If (f_n) belongs to $M^+(X, X)$, then*

$$\int \liminf f_n d\mu \leq \liminf \int f_n d\mu$$

Corollary A.1 ([3, Thm. 5.4, p. 43]). *If f is measurable, g is integrable and $|f| \leq |g|$, then f is integrable and*

$$\int |f| d\mu \leq \int |g| d\mu$$

Theorem A.10 ([3, Thm. 5.5, p. 43]). *A constant multiply αf and a sum $f + g$ of functions in L belongs to L and*

$$\int \alpha f d\mu = \alpha \int f d\mu$$
$$\int (f + g) d\mu = \int f d\mu + \int g d\mu$$

Theorem A.11 (Lebesgue Dominated Convergence Theorem, [3, Thm. 5.6, p. 44]). *Let (f_n) be a sequence of integrable functions which converges almost everywhere to a real-valued measurable function f. If the exists an integrable function g such that $|f_n| < g$ for all n, then f is integrable and*

$$\int f d\mu = \lim \int f_n d\mu$$

Proof. ∎

Corollary A.2 ([3, Thm. 5.9, p. 46]). *Suppose that for some $t_0 \in [a, b]$, the function $x \to f(x, t_0)$ is integrable on X, that $\partial f / \partial t$ exists on $X \times [a, b]$, and that there exists an integrable function g on X such that*

$$\left| \frac{\partial f}{\partial t}(x, t) \right| \leq g(x).$$

Then the function $F(t) = \int f(x,t)d\mu(x)$ *is differentiable on* $[a,b]$ *and*

$$\frac{dF}{dt} = \frac{d}{dt}\int f(x,t)d\mu(x) = \int \frac{\partial f}{\partial t} f(x,t)d\mu(x)$$

Theorem A.12 (Taylor's Theorem, [19, Thm.4, p. 391])**.** *Suppose that* $f', \dots f^{(n+1)}$*, are defined on* $[a,x]$ *and that* $R_{n,a}(x)$ *is defined by*

$$f(x) = f(a) + f'(a)(x-a) + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + R_{n,a}(x).$$

Then

(i) $R_{n,a}(x) = \dfrac{f^{(n+1)}(t)}{(n)!}(x-t)^n(x-a)$ *for some* $t \in (a,x)$.

(ii) $R_{n,a}(x) = \dfrac{f^{(n+1)}(t)}{(n+1)!}(x-a)^{n+1}$ *for some* $t \in (a,x)$.

(iii) Moreover, if $f^{(n+1)}$ *is integrable on* $[a,x]$*, then*

$$R_{n,a}(x) = \int_0^x \frac{f^{(n+1)}(t)}{(n)!}(x-t)^n dt.$$

The lagrange problem:
A general optimization problem with equality constraints if of the form

$$\max\ (\min) \quad f(x_1,\dots,x_n) \quad \text{subject to} \begin{cases} g_1(x_1,\dots,x_n) = b_1 \\ \vdots \\ g_m(x_1,\dots,x_n) = b_m \end{cases} (m < n)\,. \tag{A.4}$$

We assume that $m < n$ because otherwise there are usually no degrees of freedom. In vector formulation, the problem is

$$\max\ (\min) \quad f(x) \quad \text{subject to } g_j(x) = b_j$$

Theorem A.13 (Lagrange Theorem, [15, Thm. 3.3.1, p. 118])**.** *Suppose that the function* f *and* $g_1,\dots,g_m$ *are defined on a set* S *in* $\mathbf{R}^n$*, and that* $x^* = (x_1^*,\dots,x_n^*)$ *that solves problem*

Corollary A.3. *Suppose that for some $t_0 \in [a,b]$, $f(x,t_0) = \lim_{t\to t_0} f(x,t)$ for each $x \in X$, and that there exists an integrable function g on X such that $|f(t,x)| \leq g(x)$ for all $t \in [a,b]$. Then*

$$\int f(x,t_0)d\mu = \lim_{t\to t_0} \int f(x,t)d\mu.$$

Corollary A.4. *If the function $t \to f(t,x)$ is continuous on $[a,b]$ for each fixed $x \in X$ and exists $g \in \mathscr{L}$ such that*

$$|f(x,t)| \leq g(x).$$

Then the function

$$F(t) = \int f(x,t)d\mu(x)$$

is continuous on $[a,b]$.

A.1 Optimal Control

Definition A.1. *Let $I \subset \mathbb{R}$ be an interval. We say a finite-valued function $u : I \leftarrow \mathbb{R}$ is piecewise continuous if it continuous if it is continuous at each $t \in I$, with possible exception of at most a finite number of t, and if u is equal to either its left or right limit at every $t \in I$.*

Definition A.2. *Let $x : I \to \mathbb{R}$ be continuous on I, differentiable at all but finitely points of I. Further, suppose that x' is continuous wherever it is defined. Then, we say x is piecewise differentiable.*

Definition A.3. *Let $k : I \to \mathbb{R}$. We say k is continuously differentiable if k' exists and is continuous on I.*

Definition A.4. *A function $k(t)$ is said to be concave on $[a,b]$ if*

$$\alpha k(t_1) + (1-\alpha)k(t_2) \leq k(\alpha t_1 + (1-\alpha)t_2)$$

for all $0 \leq \alpha \leq 1$ and for any $a \leq t_1, t_2 \leq b$.

A function k is said to be convex on $[a,b]$ if it satisfies the reverse inequality , or equivalently, if $-k$ is concave. The second derivative of a twice differentiable concave function is non-positive; in the case of a convex function, is non-negative. If k is concave and differentiable, then we have a tangent line property

$$k(t_2) - k(t_1) \geq (t_2 - t_1)k'(t_2)$$

for all $a \leq t_1, t_2 \leq b$. In the case where k is a function in two variables, we have the analogue to the tangent line property as follows

$$k(x_1, y_1) - k(x_2, y_2) \geq (x_1 - x_2)k_x(x_1, y_1) + (y_1 - y_2)k_y(x_1, y_1)$$

for all points $(x_1, y_1), (x_2, y_2)$ in the domain of k.

Definition A.5. *A function k is called Lipchitz if there exists a constant c (particular to k) such that $|k(t_1) - k(t_2)| \leq c|t_1 - t_2|$ for all points t_1, t_2 in the domain of k. The constant c is called the Lipchitz constant of k.*

Note that a Lipschitz function is uniformly continuous

Theorem A.14. *If a function $k : I \to \mathbb{R}$ is piecewise differentiable on a bounded interval I, then K is Lipschitz*

Theorem A.15 (Existence Theorem)**.** *Consider the standard optimal control problem*

Pontryagins theorems

Theorem A.16 ([13, Thm.*])**.** *Consider*

$$J(u) = \int_{t_0}^{t_1} f(t, x(t), u(t))dt$$
$$\text{subject to } x'(t) = g(t, x(t), u(t)),\ x(t_0) = x_0$$

Suppose that $f(t, x(t), u(t))$ and $g(t, x(t), u(t))$ are both continuously differentiable functions in their three arguments and concave in x and u. Suppose u^* is a control, with associated state x^*, and λ a piecewise differentiable function, such that u^*, x^*, and λ together satisfy on $t_0 \leq t \leq t_1$:

$$f_u + \lambda g_u = 0,$$
$$\lambda' = f_u + \lambda g_u,$$
$$\lambda(t_1) = 0,$$
$$\lambda(t) \geq 0.$$

Then for all controls u, we have

$$J(u^*) \geq J(u)$$

Theorem A.17. *Let the set of controls for problem (aqui va una referencia) be Lebesgue integrable functions (instead of just piecewise continuous functions) on* $t_0 \leq t \leq t_1$ *with values in* $\mathbb{R}$ *Suppose that* $f(t, x(t), u(t))$ *is convex in* u*, and there exist constants* C_4 *and* $C_1, C_2, C_3 > 0$ *and* $\beta > 1$ *such that*

i. $g(t, x, u) = \alpha(t, x) + \beta(t, x)u$

ii. $|g(t, x, u)| \leq C_1|1 + |x| + |u||$

iii. $|g(t, x_1, u) - g(t, x, u)| \leq C_2|x_1 - x|(1 + |u|)$

iv. $f(t, x, u) \geq C_3|u|^\beta - C_4$

for all t *with* $t_0 \leq t \leq t_1$*,* x*,* x_1*,* u *in* $\mathbb{R}$*. Then there exists an optimal control* u^* *maximizing* $J(u)$*, with* $J(u^*)$ *finite.*

Bibliography

Bibliography

[1] Salinas René A., Lenhart Suzanne, and Gross Louis J. Control of a metapopulation harvesting model for black bears. *Natural Resource Modeling*, 18(3):307–321, 2005. doi: 10.1111/j.1939-7445.2005.tb00160.x. 1

[2] R.G. Bartle. *Elements of real analysis*. A Wiley Arabook. John Wiley & Sons Incorporated, 1982. ISBN 9780471063919. 90, 91

[3] R.G. Bartle. *The elements of integration and Lebesgue measure*. Wiley Classics Library. Wiley, 2014. ISBN 9781118626122. 91

[4] J.C. Butcher. *Numerical Methods for Ordinary Differential Equations*. Wiley, 2004. ISBN 9780470868263. 61

[5] Shelly Butler, Denise Kirschner, and Suzzane Lenhart. Optimal control of chemotherapy affecting the infectivity of hiv. *Ann Arbor*, 1001:48109–0620, 1997. 1, 65, 67

[6] John Carl Panetta and Katherine Fister. Optimal control applied to competing chemotherapeutic cell-kill strategies. *SIAM Journal of Applied Mathematics*, 63: 1954–1971, 08 2003. doi: 10.1137/S0036139902413489. 1

[7] Carlos Castillo-Chavez and Zhilan Feng. To treat or not to treat: the case of tuberculosis. *Journal of Mathematical Biology*, 35(6):629–656, Jun 1997. ISSN 1432-1416. doi: 10.1007/s002850050069. 69, 71

[8] I. Ekeland. On the variational principle. *Journal of Mathematical Analysis and Applications*, 47(2):324 – 353, 1974. ISSN 0022-247X. doi: https://doi.org/10.1016/0022-247X(74)90025-0. 3, 11, 15, 34, 35

[9] G.B. Folland. *Real analysis: modern techniques and their applications*. Pure and applied mathematics. Wiley, 1999. ISBN 9780471317166. URL `https://books.google.com.mx/books?id=uPkYAQAAIAAJ`. 90

[10] D.F. Griffiths and D.J. Higham. *Numerical methods for ordinary differential equations: Initial value problems*. Springer Undergraduate Mathematics Series. Springer London, 2010. ISBN 9780857291486. 57, 58

[11] O. Güler. *Foundations of optimization*. Graduate Texts in Mathematics. Springer New York, 2010. ISBN 9780387684079. 1, 3, 11, 13, 15, 86

[12] Eunok Jung, Suzanne Lenhart, and Zhilan Feng. Optimal control of treatments in a two-strain tuberculosis. *Discrete and Continuous Dynamical Systems*, 2, 11 2002. doi: 10.3934/dcdsb.2002.2.473. 1, 71

[13] S. Lenhart and J.T. Workman. *Optimal control applied to biological models*. Chapman & Hall/CRC Mathematical and Computational Biology. CRC Press, 2007. ISBN 9781420011418. 1, 62, 65, 82, 94

[14] Suzanne Lenhart and John Workman. Labs from mathematical and computational biology. `https://www.math.utk.edu/~lenhart/`, 2003. 65

[15] Daniel Leonard and Ngo van Long. *Optimal control theory and static optimization in economics*, volume 26. Cambridge University Press, 08 1993. 92

[16] J.E. Marsden, M.J. Hoffman, and U.M.J. Hoffman. *Elementary Classical Analysis*. W. H. Freeman, 1993. ISBN 9780716721055. 85

[17] Nohemy Palafox Lacarra and Saúl Díaz-Infante. Python implementation of the forward-backward-epidemic models. `https://github.com/nohemypalafox/Master-Thesis-Code`, 2018. 62, 65, 82

[18] H.L. Royden. *Real Analysis 3rd Ed.* Prentice-Hall Of India Pvt. Limited, 1988. ISBN 9788120309739. 33

[19] Michael Spivak. *Calculus*. Publish or Perish, 3 edition, 1994. ISBN 9780914098898,0914098896. 92

[20] Xiefei Yan and Yun Zou. Optimal and sub-optimal quarantine and isolation control in sars epidemics. *Mathematical and Computer Modelling*, 47(1):235 – 245, 2008. ISSN 0895-7177. doi: https://doi.org/10.1016/j.mcm.2007.04.003. 1

[21] Xiefei Yan and Yun Zou. Optimal and sub-optimal quarantine and isolation control in sars epidemics. *Mathematical and Computer Modelling*, 47(1):235 – 245, 2008. ISSN 0895-7177. doi: https://doi.org/10.1016/j.mcm.2007.04.003. 74

[22] Jiongmin Yong. *Differential games: A concise introduction*. World Scientific, 01 2014. doi: 10.1142/9789814596237. 1, 19, 21, 26, 29, 39, 43, 85, 89