Czech Technical University in Prague

Faculty of Electrical Engineering
Department of Control Engineering

Nonsmooth Analysis and the Maximum Principle

Karolína Sehnalová

Master's thesis

Supervisor:
doc. RNDr. Martin Bohata, Ph.D.

Field of study:
Cybernetics and Robotics

May 2024

MASTER‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Sehnalová Karolína** Personal ID number: **491904**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Control Engineering**

Study program: **Cybernetics and Robotics**

II. Master’s thesis details

Master’s thesis title in English:

Nonsmooth Analysis and the Maximum Principle

Master’s thesis title in Czech:

Nehladká analýza a princip maxima

Guidelines:

The main goal of the thesis is to discuss the maximum principle in the framework of nonsmooth analysis. The first part of the work should contain an overview of essential parts of nonsmooth analysis used in necessary conditions for optimal control problems. This should be followed by a formulation of the extended maximum principle and a discussion of its consequences (more specifically, various “standard” versions of the maximum principle). The second part of the thesis should be focused on examples illustrating applications of the maximum principle. These examples should include fixed time problems as well as free end-time problems.

Bibliography / sources:

[1] F. Clarke: Functional Analysis, Calculus of Variations and Optimal Control, Springer, London, 2013.
[2] M. Mesterton-Gibbons: A Primer on the Calculus of Variations and Optimal Control Theory, American Mathematical Society, Providence, 2009.

Name and workplace of master’s thesis supervisor:

doc. RNDr. Martin Bohata, Ph.D. Department of Mathematics FEE

Name and workplace of second master’s thesis supervisor or consultant:

Date of master’s thesis assignment: **23.01.2024** Deadline for master's thesis submission: **24.05.2024**

Assignment valid until: **21.09.2025**

doc. RNDr. Martin Bohata, Ph.D.	prof. Ing. Michael Šebek, DrSc.	prof. Mgr. Petr Páta, Ph.D.
Supervisor’s signature	Head of department’s signature	Dean’s signature

III. Assignment receipt

The student acknowledges that the master’s thesis is an individual work. The student must produce her thesis without the assistance of others, with the exception of provided consultations. Within the master’s thesis, the author must state the names of consultants and include a list of references.

.	
Date of assignment receipt	Student’s signature

© ČVUT v Praze, Design: ČVUT v Praze, VIC

Acknowledgments

First of all, I wish to express genuine gratitude to my supervisor doc. RNDr. Martin Bohata, Ph.D. for his continuous support throughout the past two semesters. His remarks, feedback, guidance and most importantly, the time he dedicated to our consultations have been immensely valuable.

Furthermore, I would like to express special thanks to my family, particularly my father, mother, and brother, for their enduring support over the years. I am also deeply grateful to my dear J, whose encouragement has been a constant source of motivation. I also wish to thank to my friends from and outside of CTU, especially those, who helped me with polishing of the text.

Declaration

I declare that I have written and implemented this thesis by myself and that I have listed all used sources according to the Guideline no. 2/2024 for adhering to ethical principles when elaborating an academic final thesis.

This thesis has also not been submitted for any degree in any university previously.

In Prague,

date	Karolína Sehnalová

Abstract

This thesis elaborates on the problematics and applications of the nonsmooth version of Pontryagin maximum principle in control theory. Firstly, basic concepts of nonsmooth analysis are introduced. In particular, definitions of generalized gradients and generalized normal cones are given. This is followed with the formulation of the extended maximum principle and its consequences, including the hybrid principle, variable-time principle and fixed-time principle. The rest of the thesis is focused on application of these principles to selected problems. The first discussed problem is the steering boat in a flowing water to a specific target set, which covers dealing with pure state constraints and time-dependent target set. The second problem concerns the optimal treatment of HIV, where situations with time-dependent control set as well as optimization of treatment turn-off time are discussed. The main results of this thesis include found optimal trajectories for various versions of the boat steering problem and found optimal control for different cases of HIV treatment and a possible approach to STI (structured treatment interruptions) HIV treatment with the usage of the hybrid principle.

Supervisor

doc. RNDr. Martin Bohata, Ph.D.
Department of Mathematics, Faculty of Electrical Engineering, CTU in Prague

Keywords

Pontryagin maximum principle, nonsmooth analysis, generalized gradients, generalized normal cones, optimal boat steering problem, optimal HIV treatment problem, STI HIV treatment

Abstrakt

Tato práce se zabývá problematikou a aplikacemi nehladké verze Pontryaginova principu maxima v teorii řízení. Nejprve jsou rozebrány základní koncepty nehladké analýzy, jako jsou zobecněné gradienty a zobecněné normálové kužely. Dále následuje formulace rozšířeného principu maxima a jeho různé důsledky, mezi nimiž je hybridní princip, princip s proměnným časem a s fixním časem. Dále se práce věnuje aplikacím těchto principů na vybrané úlohy. První rozebíranou úlohou je řízení loďky pohybující se v proudící vodě do cílové množiny. Tento problém zahrnuje čistě stavová omezení a také časově závislou cílovou množinu. Druhá úloha rozebírá optimální léčbu HIV. Je diskutována situace s časově proměnnou množinou přípustných řízení a také optimalizace času vypnutí léčby. Hlavní výsledky této práce zahrnují nalezené optimální trajektorie pro různé verze úlohy řízení loďky a nalezené optimální řízení pro různé případy léčby HIV a návrh možného přístupu k řešení STI léčby HIV (strukturovaná přerušení léčby) s použitím hybridního principu.

Vedoucí

doc. RNDr. Martin Bohata, Ph.D.
katedra matematiky, Fakulta elektrotechnická, ČVUT v Praze

Klíčová slova

Pontryaginův princip maxima, nehladká analýza, zobecněné gradienty, zobecněné normálové kužely, úloha optimálního řízení loďky, úloha optimální léčby HIV, STI léčba HIV

Contents

List of Figures

1 Introduction

1.1 History of optimal control and calculus of variations

The origins of optimal control can be traced back to the birth of calculus of variations in the 17^{th} century. Some of the first problems (see [Bry96]) of the calculus of variations were inspired by physics — the Fermat's principle, the brachystochrone or the shape of a heavy chain hanging between two points. According to [SW97], it is the solution to the brachystone problem by Johan Bernoulli that marks the birth of optimal control. Later, Euler came with the necessary condition of optimality for the basic problem of calculus of variations, which is now known as the Euler-Lagrange equation. The development of this field continued and Legendre came with additional (second order) necessary condition. Hamilton offered a new perspective when he introduced a new formalism to tackle problems of the calculus of variations. Then, Weierstrass contributed to the theory with a new necessary condition, which resembles the first occurrence of the maximum principle, when written in the Hamiltonian formalism (which was done by Carathéodory according to [PP09]).

The classical theory works with differential equations without a control function, which is crucial in the control theory. It was sufficient for the problems that emerged in physics and mechanics, so there was no need to generalize it to incorporate the control. In [SW97] the generalization is said to be straightforward and not that complicated. However, the generalization came not before 1962, when Pontryagin presented the first rigorous version of the maximum principle in [Pon+87], although some versions of the maximum principle were proven by his research group in 1950s (see [Gam19]). The Pontryagin maximum principle (PMP) encapsulates necessary conditions of Euler, Lagrange, Legendre and Weierstrass in an elegant way using the generalized Hamiltonian formalism, which incorporates the control function.

After that, many extensions and modern versions of the principle were formulated and proven. In this thesis, we will devote ourselves to Clarke's work in [Cla13], where he presents the nonsmooth version of the maximum principle. There are numerous reasons to work in a framework of nonsmooth analysis even when all the data are smooth. The simplest example of the necessity of the nonsmooth

analysis is the occurrence of discontinuous control, which is frequently used in control theory. Others involve the Hamilton-Jacobi equation or Lyapunov analysis, as stated in [Cla01].

1.2 The aim and structure of this thesis

The main goal of this thesis is to discuss the maximum principle in the framework of nonsmooth analysis. Clarke's extended principle is formulated with its consequences in the form of various standard versions of the maximum principle (fixed-time principle, variable-time principle) and the hybrid principle. Then these principles are applied to selected problems (which include both fixed-time and free-end-time problems).

The first part of the thesis contains mathematical theory of optimal control. The second chapter is devoted to establishing basic concepts, definitions and theorems of nonsmooth analysis, which will later be observed in the extended principle. In particular, some generalizations of gradients and normal vectors are discussed. The third chapter contains the formulation of the extended principle, followed by various standard versions of the maximum principle, which are indeed specific instances of the extended principle. The second part of the thesis is dedicated to solving specific examples. The fourth chapter contains the application of the variable-time principle for a minimum time problem of steering a boat in a vector field. We consider diverse types of the problem including, among other things, steering the boat to a time-dependent set of final states and pure state constraints. The last chapter focuses on the optimal treatment of HIV. Besides a basic version of the problem, we discuss the situation with a time-dependent control set. In particular, an important case with treatment interruptions is investigated by means of the hybrid principle.

Part I

Mathematical theory

2 Nonsmooth analysis

Classical calculus cannot be applied when working with nonsmooth functions and sets that can appear in control theory. Generalizing classical calculus starts with generalizing gradients and based on that we establish the basic concepts of nonsmooth analysis. Everything will be developed on normed spaces $\mathbb{R}^n$ because our state variable is a member of $\mathbb{R}^n$. This theory is presented according to Clarke's books [Cla13] and [Cla+98].

2.1 Subdifferentials and generalized gradients

In this section, we introduce various subdifferentials and generalized gradient. Definitions and theorems are taken from [Cla13]. Nonsmooth functions do not have a differential at some point. This can happen when working for example with piecewise smooth functions, which often appear in control theory.

We shall use the following notation:

- $\mathbb{R}_\infty := \mathbb{R} \cup \{+\infty\}$;
- if $f : \mathbb{R}^n \to \mathbb{R}_\infty$, then we denote $\mathrm{dom} f := \{ x \in \mathbb{R}^n \mid f(x) < +\infty \}$;
- $\langle x, y \rangle := \sum_{i=1}^n x_i y_i$ is the (Euclidean) scalar product of $x = (x_1, ..., x_n) \in \mathbb{R}^n$, $y = (y_1, ..., y_n) \in \mathbb{R}^n$;
- $\mathbf{0}$ is the zero vector;
- $|x| := \sqrt{\langle x, x \rangle}$ is the (Euclidean) norm of $x \in \mathbb{R}^n$;
- $\mathrm{co}\{M\}$ is the convex hull of $M \subseteq \mathbb{R}^n$ (i.e., the intersection of all convex subsets of $\mathbb{R}^n$ containing M);
- $B(x_0, r) := \{ x \in \mathbb{R}^n \mid |x - x_0| \leq r \}$.

In this text, we treat row vectors and column vectors interchangeably.

► **Definition 2.1.**
Subgradient of a convex function $f : \mathbb{R}^n \to \mathbb{R}_\infty$ at $x \in \mathrm{dom} f$ is any vector $\zeta \in \mathbb{R}^n$ satisfying

$$f(y) - f(x) \geq \langle \zeta \, , \, y - x \rangle$$

for all y $\in \mathbb{R}^n$.
Furthermore, the set

$$\partial f(x) := \{ \, \zeta \mid f(y) - f(x) \geq \langle \zeta \, , \, y - x \rangle \, \}$$

is called the *subdifferential* of f at x. ◄

We can rewrite subgradient inequality as follows:

$$f(y) \geq f(x) + \langle \zeta \, , \, y - x \rangle \; = (f(x) - \langle \zeta \, , \, x \rangle) + \langle \zeta \, , \, y \rangle,$$

which implies that the graph of f lies above the graph of an affine function

$$h : y \mapsto (f(x) - \langle \zeta \, , \, x \rangle) + \langle \zeta \, , \, y \rangle.$$

The graph of h is called a supporting hyperplane of f at x because the inequality holds with equality at x. We can notice that if a (convex) function f is continuously differentiable at x, the concept of subgradient reduces to classical gradient and subdifferential is a singleton containing the gradient, i.e., $\partial f(x) = \{\nabla f(x)\}$. We illustrate this concept in the following example.

► **Example 2.2.**
Let $f(x) = |x|$, $x \in \mathbb{R}$. This function is indeed convex and it is not differentiable at $x = 0$. Let us compute its subdifferential at 0. It follows from Definition 2.1 that $\zeta \in \partial f(0)$ if and only if

$$\zeta y \leq |y| \tag{2.1}$$

for all $y \in \mathbb{R}$. If $\zeta \in \langle -1, 1 \rangle$, then

$$\zeta y \leq |\zeta y| = |\zeta||y| \leq |y|.$$

Thus $\langle -1, 1\rangle \subseteq \partial f(0)$. Now assume that $|\zeta| > 1$. Take $y = \frac{\zeta}{|\zeta|}$. Then,

$$\zeta y = |\zeta| > 1 = |y|.$$

Therefore, equation (2.1) is not satisfied in this case. This shows that

$$\partial f(0) = \langle -1, 1\rangle.$$

◀

The use of convex functions can be quite restrictive. In order to generalize the subgradient of convex functions, we consider the following characterization.

► **Theorem 2.3.**
Let $f : \mathbb{R}^n \to \mathbb{R}_\infty$ be a convex function, $x \in \mathrm{dom} f$, $\zeta \in \mathbb{R}^n$. Then $\zeta \in \partial f(x)$ if and only if there exist a neighbourhood $U(x)$ of x and $\sigma \geq 0$ such that

$$f(y) - f(x) + \sigma|y - x|^2 \geq \langle \zeta, y - x\rangle \quad \text{for all } y \in U(x).$$

◀

Proof. $\Rightarrow$ Follows immediately from Definition 2.1.

$\Leftarrow$ Suppose that $\zeta \in \mathbb{R}^n$ is such that there are a neigbourhood $U(x)$ of x and $\sigma \geq 0$ safisfying $f(y) - f(x) + \sigma|y - x|^2 \geq \langle \zeta, y - x\rangle$ for all $y \in U(x)$. Given $y \in \mathbb{R}^n$, $ty \in U(x)$ for each $t \in (0, 1\rangle$ sufficiently small. Using the convexity of f and linearity of the scalar product, we conclude

$$\begin{aligned} f(y) - f(x) &= \frac{1}{t}[tf(y) + (1-t)f(x) - f(x)] \geq \frac{1}{t}[f(ty + (1-t)x) - f(x)] \\ &\geq \frac{1}{t}\big[\langle \zeta, ty + (1-t)x - x\rangle - \sigma|ty + (1-t)x - x|^2\big] \\ &\geq \frac{1}{t}\big[t\langle \zeta, y - x\rangle - t^2\sigma|y - x|^2\big] \geq \langle \zeta, y - x\rangle - t\sigma|y - x|^2. \end{aligned}$$

By taking the limit as $t \to 0^+$, we obtain

$$f(y) - f(x) \geq \langle \zeta, y - x\rangle,$$

which is the desirable subgradient inequality from Definition 2.1.

■

This characterization enables us to generalize subgradient by adding the term $\sigma|y-x|^2$ to the subgradient inequality.

► **Definition 2.4.**
Proximal subgradient of $f : \mathbb{R}^n \to \mathbb{R}_\infty$ at $x \in \mathrm{dom} f$ is any vector $\zeta \in \mathbb{R}^n$ such that there exist $\sigma \geq 0$ and a neighbourhood $U(x)$ of x satisfying

$$f(y) - f(x) + \sigma|y-x|^2 \geq \langle \zeta \ , \ y-x \rangle$$

for all $y \in U(x)$.
Furthermore, the set

$$\partial_P f(x) := \left\{ \zeta \ \middle| \ f(y) - f(x) + \sigma|y-x|^2 \geq \langle \zeta \ , \ y-x \rangle \right\}$$

is called a *proximal subdifferential* of f at x. ◄

We can see that there is an additional non-negative term on the left-hand side of the proximal subgradient inequality which weakens the previously mentioned subgradient inequality. The geometrical interpretation of proximal subgradient in a scalar case is that we look for locally supporting parabolas instead of supporting hyperplanes. This is clear when we rewrite the inequality in the same manner as we did with subgradient inequality

$$f(y) \geq f(x) + \langle \zeta \ , \ y-x \rangle \ - \sigma|y-x|^2 = (f(x) - \langle \zeta \ , \ x \rangle) + \langle \zeta \ , \ y \rangle - \sigma|y-x|^2,$$

where the quadratic term (defining a downward opening parabola for scalar case) appears. Important property of locally supporting parabolas of a function f at x is that they lie below the graph of f and touch the graph at f at $(x, f(x))$. A continuously differentiable scalar function f has only one locally supporting parabola and its slope at x is the same as $f'(x)$. Geometrical interpretation for functions of more than one variable is analogical - we have locally supporting graphs of polynomials of degree at most two instead of locally supporting parabolas.

The following example shows the computation of proximal subgradient.

► **Example 2.5.**
Let

$$f(x) = \begin{cases} -|x|^{\frac{3}{2}} & \text{if } x < 0, \\ x & \text{if } x \geq 0. \end{cases}$$

The function f is nonconvex, as can be seen in the Figure 2.1.

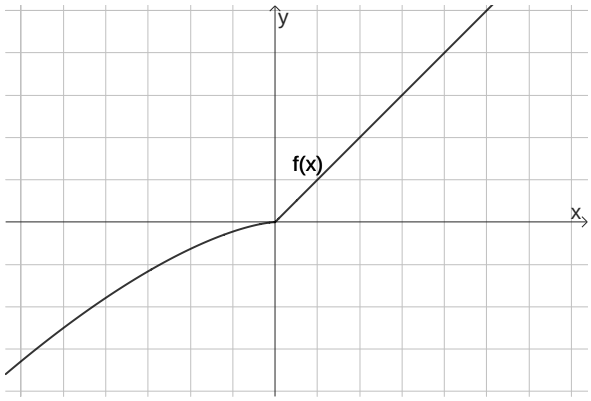


Figure 2.1: The graph of the function f from Example 2.5

We compute the proximal subgradient of f at $x = 0$, where the function f is not differentiable.

Let $y > 0$. Then, for sufficiently small y, we have

$$y + \sigma y^2 \geq \zeta y \iff \sigma y \geq \zeta - 1 \iff \zeta \leq 1.$$

Let $y < 0$. Then, we have the proximal subgradient inequality

$$-|y|^{\frac{3}{2}} + \sigma y^2 \geq \zeta y \iff |y|^{\frac{1}{2}} - \sigma|y| \leq \zeta \iff |y|^{\frac{1}{2}}(1 - \sigma|y|^{\frac{1}{2}}) \leq \zeta.$$

Thus, $\zeta > 0$. Note that ζ cannot be 0, because whenever $\sigma \geq 0$ and $y < 0$ is sufficiently small (recall that we need one constant σ for every y in a neighbourhood of 0), we have $1 - \sigma|y|^{\frac{1}{2}} > 0$, which means $|y|^{\frac{1}{2}}(1 - \sigma|y|^{\frac{1}{2}}) > 0$ and hence, $0 \notin \partial_P f(0)$. Finally, we conclude that

$$\partial_P f(0) = (0, 1\rangle.$$

◀

Note that $\partial_P f(x)$ may not be closed, as in Example 2.5, and for that reason, we now proceed to the definition of limiting subdifferential.

► **Definition 2.6.**
The *limiting subdifferential* of f at x is defined as

$$\partial_L f(x) = \left\{ \zeta = \lim_{i\to\infty} \zeta_i \;\middle|\; \zeta_i \in \partial_P f(x_i), x_i \to x, f(x_i) \to f(x) \right\}.$$

◄

From the construction of limiting subdifferential, we can easily see that whenever $\zeta \in \partial_P f(x)$, it holds that $\zeta \in \partial_L f(x)$ and therefore, it follows that $\partial_P f(x) \subseteq \partial_L f(x)$. Computation of limiting subdifferential is illustrated in the following example.

► **Example 2.7.**
We take the same function as in Example 2.5, that is,

$$f(x) = \begin{cases} -|x|^{\frac{3}{2}} & \text{if } x < 0, \\ x & \text{if } x \geq 0, \end{cases}$$

and compute $\partial_L f(0)$. We have seen that $\partial_P f(0) = (0, 1\rangle$. As f is continuously differentiable at every point $x \in \mathbb{R} \setminus \{0\}$, we have

$$\partial_P f(x) = \begin{cases} \left\{ \frac{3}{2}|x|^{\frac{1}{2}} \right\} & \text{if } x < 0, \\ \{1\} & \text{if } x \geq 0. \end{cases}$$

Hence, when we add the limits from Definition 2.6, we conclude that

$$\partial_L f(0) = \langle 0, 1\rangle.$$

◄

To proceed further, we need to define Lipschitz property.

► **Definition 2.8.**
We say the function $f : \mathbb{R}^n \to \mathbb{R}_\infty$ is

(i) *Lipschitz on* $M \subseteq \mathrm{dom} f$, if there exists a constant $K > 0$ such that

$$|f(x) - f(y)| \leq K|x - y| \quad \text{for all } x, y \in M,$$

(ii) *Lipschitz*, if it is Lipschitz on $\mathbb{R}^n$;

(iii) *Lipschitz near x*, if there exist a neigbourhood $U(x)$ of x, such that f is Lipschitz on a set $U(x) \cap \mathrm{dom} f$;

(iv) *locally Lipschitz on an open set $M \subseteq \mathbb{R}^n$*, if it is Lipschitz near x for each $x \in M$;

(v) *locally Lipschitz*, if it is locally Lipschitz on $\mathbb{R}^n$.

◀

Observe that if f is (locally) Lipschitz, then $\mathrm{dom} f = \mathbb{R}^n$.

Now we introduce a generalized gradient. Let us restrict our attention to locally Lipschitz functions $f : \mathbb{R}^n \to \mathbb{R}_\infty$. According to a corollary of the famous Rademacher's theorem presented in [CMN19, p. 196], any locally Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}_\infty$ is differentiable almost everywhere. The restriction to locally Lipschitz functions allows us to use an alternative definition of generalized gradient compared to what Clarke presents in [Cla13, p. 196]. In our special case, the derivative of f is used to generate generalized gradient, whereas Clarke uses generalized directional derivative.

▶ **Definition 2.9.**
Let $f : \mathbb{R}^n \to \mathbb{R}_\infty$ be a locally Lipschitz function and let E_f be a set of all points in $\mathbb{R}^n$, where f is not differentiable. The *generalized gradient* of f at x is a set

$$\partial_C f(x) = \mathrm{co}\left\{ \lim_{i \to \infty} \nabla f(x_i) \;\middle|\; x_i \to x, x_i \in \mathbb{R}^n \setminus E_f \right\}.$$

◀

This definition is nothing but a Theorem 10.27 in [Cla13, p. 208] (Gradient formula) for the specific case, where the set of points at which f fails to be differentiable has zero measure. Consequently, if f is continuously differentiable at x, it is evident that $\partial_C f(x) = \{\nabla f(x)\}$, because all sequences in the definition converge to $\nabla f(x)$. Moreover, if the function f is continuously differentiable at x, we have

$$\partial_P f(x) = \partial_L f(x) = \partial_C f(x) = \{\nabla f(x)\}.$$

It can be proved that if f is locally Lipschitz, then it holds that

$$\partial_C f(x) = \mathrm{co} \partial_L f(x). \tag{2.2}$$

A construction of generalized gradient is shown in the following example.

► **Example 2.10.**
We consider the same function as in Examples 2.5 and 2.7, that is,

$$f(x) = \begin{cases} -|x|^{\frac{3}{2}} & \text{if } x < 0, \\ x & \text{if } x \geq 0, \end{cases}$$

which, indeed, is locally Lipschitz. Clearly, the set of points, where f fails to be differentiable, is $E_f = \{0\}$. From Definition 2.9, we have

$$\partial_C f(0) = \operatorname{co}\{0, 1\} = \langle 0, 1 \rangle.$$

Note that in this case, $\partial_L f(0) = \partial_C f(0)$, which corresponds to equation (2.2), because $\partial_L f(0)$ is a convex set (in this particular example).

◄

A good and brief overview of the terms established in this section is presented in [Cla09].

2.2 Nonsmooth geometry

We would like to study geometrical aspects of sets with tools of nonsmooth analysis taken from [Cla+98].

► **Definition 2.11.**
Let $M \subseteq \mathbb{R}^n$ be a nonempty set and $x \in \mathbb{R}^n$. The *distance of the point x from the set M* is defined as

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

The *set of closest points in M to x* is defined as

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

◄

The distance function d_M plays a crucial role when developing nonsmooth geometry.

► **Remark 2.12.**
If a nonempty set $M \subseteq \mathbb{R}^n$ is closed, then $P_M(x) \neq \emptyset$. ◄

Proof. We distinguish two cases. If $x \in M$, then trivially $x \in P_M(x)$. Let us discuss the case when $x \notin M$. Since $\{\, |x - y| \mid y \in M \,\}$ is nonempty and bounded below, $d_M(x) \in \mathbb{R}$. By definition of infimum, there exists a sequence $(z_i)_{i=1}^{\infty}$, $z_i \in M$, such that $|x - z_n| < d_M(x) + \frac{1}{n}$, where $n \in \mathbb{N}$. The sequence $(z_i)_{i=1}^{\infty}$ is bounded, because we have

$$|z_n| = |z_n - x + x| \leq |z_n - x| + |x| < d_M(x) + \frac{1}{n} + |x| \leq d_M(x) + 1 + |x|.$$

Therefore, the sequence $(z_i)_{i=1}^{\infty}$ has a convergent subsequence and hence, without loss of generality, we assume $(z_i)_{i=1}^{\infty}$ is convergent. Let $\hat{z} = \lim_{i\to\infty} z_i$. Therefore, we have

$$d_M(x) = \inf_{z\in M} |x - z| = \lim_{i\to\infty} |x - z_i| = |x - \lim_{i\to\infty} z_i| = |x - \hat{z}|.$$

From closeness of M we know that $\hat{z} \in M$ and therefore, $\hat{z}$ is an element of $P_M(x)$. ■

The aim of this section is to define various normal cones for sets with nonsmooth boundary. Firstly, we define proximal normal cone.

► **Definition 2.13.**
Proximal normal of a nonempty set $M \subseteq \mathbb{R}^n$ at $x \in M$ is any vector $\zeta \in \mathbb{R}^n$ satisfying

$$d_M(x + t\zeta) = t|\zeta|$$

for some $t > 0$.
Furthermore, the set

$$N_M^P(x) := \{\, \zeta \in \mathbb{R}^n \mid \zeta \text{ is proximal normal of } M \text{ at } x \,\}$$

is called a *proximal normal cone* of set M at x. ◄

Proximal normals represent directions from $x \in M$ to $x + t\zeta \notin M$, where x is the closest point in M to $x + t\zeta$. If M has nonsmooth boundary at x, it can happen that there exist two linearly independent vectors ζ_1 and ζ_2, which both belong to the proximal normal cone of M at x.

We can describe $N_M^P(x)$, using an inequality similar to proximal normal inequality

in the previous section in the following proposition. This underlines the correspondence of the proximal subdifferential and proximal normal cone, which will be discussed later in this section.

▶ **Proposition 2.14.**
Let $x \in M$. A vector $\zeta \in N_M^P(x)$ if and only if there exists some $\sigma = \sigma(x, \zeta) \geq 0$ such that

$$\langle \zeta, u - x \rangle \leq \sigma |u - x|^2 \quad \text{for all } u \in M. \tag{2.3}$$

◀

Proof. $\Rightarrow$ We know that $\zeta \in N_M^P(x)$ satisfies $\mathrm{d}_M(x + t\zeta) = t|\zeta|$ for some $t > 0$. Thus, $t|\zeta| = \inf_{u \in M} |x + t\zeta - u|$. By the definition of infimum, the inequality

$$t|\zeta| \leq |x + t\zeta - u|$$

holds for all $u \in M$. Therefore, for all $u \in M$, we have

$$\begin{aligned} t^2 \langle \zeta, \zeta \rangle &\leq \langle x + t\zeta - u, x + t\zeta - u \rangle \\ &= \langle x - u, x - u \rangle + 2\langle t\zeta, x - u \rangle + \langle t\zeta, t\zeta \rangle. \end{aligned}$$

From this inequality, we conclude

$$\langle \zeta, u - x \rangle \leq \frac{1}{2t} |u - x|^2$$

for all $u \in M$. This shows that the inequality (2.3) is satisfied with $\sigma := \frac{1}{2t}$.

$\Leftarrow$ We know that there is a $\sigma \geq 0$ such that

$$\langle \zeta, u - x \rangle \leq \sigma |u - x|^2 \quad \forall u \in M.$$

We compute the (squared) distance of the point $x + t\zeta$ from the set M as follows

$$\mathrm{d}_M(x + t\zeta)^2 = \inf_{u \in M} |x + t\zeta - u|^2 = \inf_{u \in M} \{\langle t\zeta, t\zeta \rangle + 2t\langle \zeta, x - u \rangle + \langle x - u, x - u \rangle\}.$$

If $\sigma = 0$, then we have

$$\mathrm{d}_M(x + t\zeta)^2 \geq \inf_{u \in M} \{t^2 |\zeta|^2 + |x - u|^2\} = t^2 |\zeta|^2,$$

because $u \mapsto |x-u|^2$ attains minimum at $u = x$ (recall that $x \in M$). If $\sigma > 0$, we have

$$\mathrm{d}_M(x+t\zeta)^2 \geq \inf_{u\in M}\{t^2|\zeta|^2 - 2t\sigma|x-u|^2 + |x-u|^2\}.$$

We choose $t := \frac{1}{2\sigma}$, which gives us

$$\mathrm{d}_M(x+t\zeta)^2 \geq \inf_{u\in M} t^2|\zeta|^2 = t^2|\zeta|^2.$$

We show the opposite inequality from the definition of infimum as follows

$$\mathrm{d}_M(x+t\zeta)^2 = \inf_{u\in M}|x+t\zeta-u|^2 \leq |x+t\zeta-u|^2 \quad \forall u \in M.$$

Since $x \in M$, we have

$$\mathrm{d}_M(x+t\zeta)^2 \leq t^2|\zeta|^2.$$

Thus, we have shown that

$$t^2|\zeta|^2 \leq \mathrm{d}_M(x+t\zeta)^2 \leq t^2|\zeta|^2$$

and consequently,

$$\mathrm{d}_M(x+t\zeta) = t|\zeta|.$$

■

► **Remark 2.15.**
From the proof it is clear, that the inequality (2.3) holds with $\sigma = 0$ if and only if the equation $\mathrm{d}_M(x+t\zeta) = t|\zeta|$ holds for all $t > 0$. ◄

We proceed to a definition of limiting normal cone. Observe the correspondence of definitions of limiting subdifferential and limiting normal cone.

► **Definition 2.16.**
Limiting normal cone of a nonempty set M at x is defined as

$$N_M^L(x) = \left\{ \zeta = \lim_{i\to\infty}\zeta_i \;\middle|\; \zeta_i \in N_M^P(x_i), x_i \to x, x_i \in M \right\}.$$

◄

It is useful to notice the inclusion $N_M^P(x) \subseteq N_M^L(x)$, which is not surprising, because of the corresponding inclusion holds between proximal and limiting subdifferential. Indeed, we shall see later, that $N_M^P(x)$ and $N_M^L(x)$ are, respectively, the proximal subdifferential and the limiting subdifferential of an appropriate function. From the definition of $N_M^L(x)$, it can be seen, that $N_M^L(x)$ is closed and may not be convex and from Proposition 2.14 it is evident that $N_M^P(x)$ is convex. Consider $\zeta_1, \zeta_2 \in N_M^P(x)$. Then, we have $\langle \zeta_1, u - x\rangle \leq \sigma_1 |u - x|^2$ and $\langle \zeta_2, u - x\rangle \leq \sigma_2 |u - x|^2$ for all $u \in M$. Thus, for $t \in (0, 1)$, we have

$$\langle t\zeta_1 + (1-t)\zeta_2, u - x\rangle = t\langle \zeta_1, u - x\rangle + (1-t)\langle \zeta_2, u - x\rangle \leq \left(\frac{\sigma_1}{t} + \frac{\sigma_2}{1-t}\right)|u - x|^2$$

and hence, $t\zeta_1 + (1-t)\zeta_2 \in N_M^P(x)$. However, $N_M^P(x)$ may not be closed, as we will see later.

Now we would like to show a connection between limiting (resp. proximal) normal cones and limiting (resp. proximal) subdifferentials, but firstly it is necessary to define the so-called indicator function of a set.

► **Definition 2.17.**
The *Indicator function* $I_M : \mathbb{R}^n \to \mathbb{R}_\infty$ of a nonempty set $M \subseteq \mathbb{R}^n$ is defined as

$$I_M(x) = \begin{cases} 0 & \text{if } x \in M, \\ \infty & \text{if } x \notin M. \end{cases}$$

◄

► **Proposition 2.18.**
Let $x \in M$. It holds that

$$N_M^L(x) = \{\, t\zeta \mid t \geq 0, \zeta \in \partial_L d_M(x) \,\} = \partial_L I_M(x),$$
$$N_M^P(x) = \{\, t\zeta \mid t \geq 0, \zeta \in \partial_P d_M(x) \,\} = \partial_P I_M(x).$$

◄

For the proof, we refer the reader to [Cla13]. Simply said, for a set M, limiting (resp. proximal) normal cone of $x \in M$ is equal to limiting (resp. proximal) subdifferential of its indicator function $I_M(x)$ at x. This relates cones to previously defined subdifferential in an elegant way. Observe that the limiting (resp. proximal) normal cone consists of limiting (resp. proximal) subdifferential of the distance function. This proposition has an intuitive geometrical interpretation. Indeed, the

limiting (resp. proximal) subdifferential of the distance function at $x \in M$ are all directions, where the distance from M grows the most. These directions then form the limiting (resp. proximal) normal cone. Let us define a lower semicontinuous function.

► **Definition 2.19.**
Let $M \subseteq \mathbb{R}^n$, $f : M \to \mathbb{R}_\infty$. We say the function f is *lower semicontinuous* if, for all $c \in \mathbb{R}$, the set $\{ x \in M \mid f(x) \leq c \}$ is closed. ◄

Proposition 2.18 allows us to formulate optimization problem of minimizing a function f on a set M as minimizing function $h(x) := f(x) + I_M(x)$ with tools similar to classical calculus. This is shown in the following theorem.

► **Theorem 2.20.**
Let $\hat{x}$ be a local minimizer of the function $f : \mathbb{R}^n \to \mathbb{R}_\infty$ on a nonempty closed set $M \subset \mathbb{R}^n$, where f is Lipschitz near $\hat{x}$. Then it holds that

$$\mathbf{0} \in \partial_L f(\hat{x}) + N_M^L(\hat{x}).$$

◄

Proof. We will use a proximal sum rule proved in [Cla13, p. 234], which says that

$$\partial_L(f_1 + f_2)(x) \subset \partial_L f_1(x) + \partial_L f_2(x), \tag{2.4}$$

whenever f_1 and f_2 are lower semicontinuous, $x \in \mathrm{dom} f_1 \cap \mathrm{dom} f_2$ and at least one of the functions f_1 and f_2 is Lipschitz near x.

We reformulate the optimization problem as $\min f(x) + I_M(x)$. We know that limiting subdifferential exists and hence, the proximal subdifferential exists (recall that $\partial_P f(x) \subseteq \partial_L f(x)$). The definition of a local minimum says that there exits a neighbourhood $U(\hat{x})$ of $\hat{x}$, such that

$$(f(x) + I_M(x))(x) - (f(x) + I_M(x))(\hat{x}) \geq 0 = \langle \mathbf{0}, x - \hat{x} \rangle \quad \text{for all } x \in U(\hat{x})$$

and hence, from Definition 2.4, we immediately see that

$$\mathbf{0} \in \partial_P(f + I_M)(\hat{x}), \quad \mathbf{0} \in \partial_L(f + I_M)(\hat{x}).$$

We can use (2.4) and Proposition 2.18 to derive

$$\mathbf{0} \in \partial_L(f + I_M)(\hat{x}) \subset \partial_L f(\hat{x}) + \partial_L I_M(\hat{x}) = \partial_L f(\hat{x}) + N_M^L(\hat{x}).$$

■

This theorem allows us to combine the geometry of a set of feasible solutions with the objective function and use proximal calculus to solve these problems. It can be easily seen that when $M = \mathbb{R}^n$, then $N_{\mathbb{R}^n}^L = \{\mathbf{0}\}$ and hence, necessary conditions of optimality for minimizing $f(x)$ are reduced to $\mathbf{0} \in \partial_L f(x)$, which corresponds to knowledge from classical calculus in the case of continuously differentiable functions. This theorem allows us to put both constrained and unconstrained optimization problems into one formalism.

We have already encountered a cone $N_M^P(x)$, which is convex and cone $N_M^L(x)$, which is closed. We follow with the definition of the generalized normal cone, which is both closed and convex (which can be seen directly from the definition).

► **Definition 2.21.**
Let $x \in M$. Then, the *generalized normal cone* of M at x is defined as

$$N_M^C(x) = \overline{\text{co}} N_M^L(x).$$

◄

We state a lemma, which can be used to compute proximal normal cones.

► **Lemma 2.22.**
Let $x \in M \subseteq \mathbb{R}^n$ and $\zeta \in \mathbb{R}^n$. If there is $v \in \mathbb{R}^n$ such that $x + \alpha v \in M$ for all $\alpha > 0$ suffuciently small and $\langle \zeta, v \rangle > 0$, then $\zeta \notin N_M^P(x)$. ◄

Proof. Let $\zeta \in N_M^P(x)$. It follows from Proposition 2.14, that there is $\sigma \geq 0$ such that

$$\langle \zeta, v \rangle \leq \alpha\sigma|v|^2$$

for all $\alpha > 0$ sufficiently small. This is a contradiction with the assumption that $\langle \zeta, v \rangle > 0$ and hence, $\zeta \notin N_M^P(x)$. ■

The construction of all the sets is illustrated in a following example.

▶ **Example 2.23.**
We consider a set

$$M = \{ (x, y) \mid x \leq 0 \} \cup \{ (x, y) \mid y \leq 0 \}$$

and our task is to compute $N_M^P(\mathbf{0}), N_M^L(\mathbf{0})$ and $N_M^C(\mathbf{0})$. At first, we take a look at the proximal normal cone. From the definition of proximal normal cone we know, that $\zeta = (\zeta_x, \zeta_y) \in \mathbb{R}^2$ is a proximal normal to M at $\mathbf{0}$ if and only if it holds that

$$\mathrm{d}_M(t\zeta) = t|\zeta|$$

for some $t > 0$. We compute

$$\mathrm{d}_M(t\zeta) = \inf_{y \in M} |t\zeta - y| = \begin{cases} 0 & \text{if } t\zeta \in M, \\ t \min\{\zeta_x, \zeta_y\} & \text{if } t\zeta \notin M. \end{cases}$$

Hence, when $t\zeta \in M$, it holds that $0 = t|\zeta|$ and so $\zeta = \mathbf{0}$. Now assume $t\zeta \notin M$. Then $t \min\{\zeta_x, \zeta_y\} = t|\zeta|$, which can hold only for $\zeta = (\zeta_x, 0)$ or $\zeta = (0, \zeta_y)$. Both of these elements belong to M. Since M is a cone (i.e., $t > 0$ and $\zeta \in M$ imply $t\zeta \in M$), we obtain a contradiction with our assumption $t\zeta \notin M$. Hence, there is no vector $t\zeta \notin M$ such that $\mathrm{d}_M(t\zeta) = t|\zeta|$. Consequently,

$$N_M^P(\mathbf{0}) = \{\mathbf{0}\}.$$

To construct the imiting normal cone $N_M^L(\mathbf{0})$, we need to know $N_M^P(x)$ for all $x \in M$. Observe that it follows from Lemma 2.22, that $N_M^P(x) = \{\mathbf{0}\}$ whenever x belongs to the interior of M (set $v = \zeta$).

Let $x \in \{ (0, a) \mid a > 0 \}$. From Lemma 2.22, we know that $\zeta \notin N_M^P(x)$ whenever $\zeta = (\zeta_x, \zeta_y)$ be such that $\zeta_x < 0$. Consider $\zeta_x \geq 0$. If $\zeta_y > 0$, then Lemma 2.22 ensures that $\zeta \notin N_M^P(x)$ (take $v = (0, 1)$). Similarly, $\zeta \notin N_M^P(x)$ if $\zeta_y < 0$ (take $v = (0, 1)$). If $\zeta_y = 0$, then we have $\langle \zeta, u - x \rangle \leq 0$ for all $u \in M$. Thus, $N_M^P(x) = \{ (t, 0) \mid t \geq 0 \}$.

Analogically, we can compute that for $x \in \{ (a, 0) \mid a > 0 \}$, the proximal normal cone is $N_M^P(x) = \{ (0, t) \mid t \geq 0 \}$.

Let us divert our attention to $N_M^L(\mathbf{0})$. Let

$$C = \{ (t, 0) \mid t \geq 0 \} \cup \{ (0, t) \mid t \geq 0 \}.$$

It follows from the definition of limiting normal cone and the discussion given above, that $C \subseteq N_M^L(\mathbf{0})$. Let $\zeta \in \mathbb{R}^n \setminus C$. Then $\zeta \notin N_M^L(\mathbf{0})$, because for all $x \in M$,

$N_M^P(x)$ contains only vectors with at least one zero component. Therefore,

$$N_M^L(\mathbf{0}) = \{\,(t,0) \mid t \geq 0\,\} \cup \{\,(0,t) \mid t \geq 0\,\}.$$

Finally, we compute generalized normal cone from Definition 2.21 as

$$N_M^C(\mathbf{0}) = \overline{\text{co}} N_M^L(\mathbf{0}) = \{\,(t,s) \mid t,s \geq 0\,\}.$$

The situation is illustrated in the Figure 2.2, where we can see the vector $\zeta = (\zeta_x, \zeta_y) \notin M$ (in this situation we assume $\zeta_x < \zeta_y$) and the closest point $(0, \zeta_y)$ in M to ζ. Proximal normal cone is blue, limiting normal cone is red and generalized normal cone is yellow.

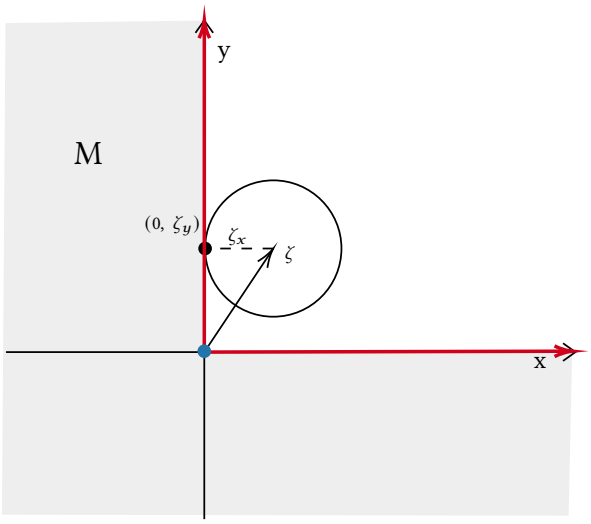


Figure 2.2: Illustration of $N_M^P(\mathbf{0})$, $N_M^L(\mathbf{0})$ and $N_M^C(\mathbf{0})$ from Example 2.23

We can clearly see that all three cones are different and that the limiting normal cone is not convex. ◀

Relations between all defined sets are described in the following theorem.

▶ **Theorem 2.24.**
Let $x \in M$. Then

$$N_M^P(x) \subseteq N_M^L(x) \subseteq N_M^C(x). \tag{2.5}$$

Moreover, if M is closed and convex, all the sets $N_M^P(x)$, $N_M^L(x)$ and $N_M^C(x)$ coincide,

they are convex and closed and can be expressed as

$$N_M^P(x) = N_M^L(x) = N_M^C(x) = \{\, \zeta \in \mathbb{R}^n \mid \langle \zeta, u - x \rangle \leq 0 \quad \forall u \in M \,\}. \tag{2.6}$$

◀

Proof. The inclusion $N_M^P(x) \subseteq N_M^L(x)$ has already been discussed. The inclusion $N_M^L(x) \subseteq N_M^C(x)$ is evident from the definitions of sets $N_M^L(x)$ and $N_M^C(x)$. We prove that equation (2.6) is satisfied whenever M is convex and closed.

Firstly, we show that $\zeta \in N_M^P(x) \iff \langle \zeta, u - x \rangle \leq 0 \quad \forall u \in M$.

$\Leftarrow$ If $\langle \zeta, u - x \rangle \leq 0$ for all $u \in M$, then we immediately see from Proposition 2.14 that $\zeta \in N_M^P(x)$.

$\Rightarrow$ Let $x \in M$. Let $\zeta \in N_M^P(x)$. From Proposition 2.14, there is $\sigma \geq 0$ such that

$$\langle \zeta, u - x \rangle \leq \sigma |u - x|^2 \quad \text{for all } u \in M.$$

Hence, for all $u \in M$ and for all $t \in (0, 1\rangle$, we have

$$\begin{aligned} \langle \zeta, u - x \rangle &= \frac{1}{t} \langle \zeta, t(u - x) \rangle = \frac{1}{t} \langle \zeta, \underbrace{tu + (1 - t)x}_{:= v \in M} - x \rangle \\ &\leq \frac{1}{t} \sigma |v - x|^2 = \frac{1}{t} \sigma |tu + (1 - t)x - x|^2 \\ &= \frac{1}{t} \sigma |t(u - x)|^2 = t\sigma |u - x|^2. \end{aligned}$$

We take the limit $t \to 0^+$, which gives us

$$\langle \zeta, u - x \rangle \leq 0 \quad \text{for all } u \in M.$$

Therefore,

$$N_M^P(x) = \{\, \zeta \in \mathbb{R}^n \mid \langle \zeta, u - x \rangle \leq 0 \quad \forall u \in M \,\}.$$

Now we would like to show that $N_M^L(x) \subseteq N_M^P(x)$. Let $\zeta \in N_M^L(x)$. By Definition 2.16, there exist sequences $(x_i)_{i=1}^\infty$ and $(\zeta_i)_{i=1}^\infty$ such that $x_i \in M$, $x_i \to x$, $\zeta_i \in N_M^P(x_i)$, $\zeta_i \to \zeta$. Using what we have proved above, we know that

$$\langle \zeta_i, u - x_i \rangle \leq 0 \quad \text{for all } u \in M.$$

By taking the limit $i \to \infty$, we obtain

$$\langle \zeta, u - x \rangle \leq 0$$

and so $\zeta \in N_M^P(x)$ and thus, $N_M^L(x) \subseteq N_M^P(x)$. Recall that $N_M^P(x)$ is convex, $N_M^L(x)$ is closed and we have just shown that $N_M^P(x) = N_M^L(x)$. Thus,

$$N_M^C(x) = \overline{\text{co}} N_M^L(x) = N_M^L(x) = N_M^P(x) = \{ \zeta \in \mathbb{R}^n \mid \langle \zeta, u - x \rangle \leq 0 \quad \forall u \in M \},$$

which is our desired conclusion. ■

Finally, we state one more proposition, that helps with the computation of proximal normal cones in some particular situations.

► **Proposition 2.25.**
Let $g : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable convex function. Assume that

$$M = \{ x \in \mathbb{R}^n \mid g(x) \leq 0 \},$$

$y \in \mathbb{R}^n$ belongs to the boundary of M and $\nabla g(y) \neq \mathbf{0}$. Then,

$$N_M^P(y) = \{ \alpha \nabla g(y) \mid \alpha \geq 0 \}.$$

◄

Proof. Firstly, we show that M is closed. If we have a sequence $(x_i)_{i=1}^{\infty}$, $x_i \in M$, $x_i \to x \in \mathbb{R}^n$, then $g(x) = \lim_{i \to \infty} g(x_i) \leq 0$ (by continuity of g) and therefore, $x \in M$ and hence, M is closed. Moreover, M is convex, because

$$g(tx + (1-t)y) \leq tg(x) + (1-t)g(y) \leq 0,$$

whenever $x, y \in M$ and $t \in \langle 0, 1 \rangle$.

Let y be a boundary point of M. We know that $y \in M$. It follows from Theorem 2.24, that $\zeta \in N_M^P(y)$ if and only if

$$\langle \zeta, x - y \rangle = \langle \zeta, x \rangle - \langle \zeta, y \rangle \leq 0 \quad \text{for all } x \in M.$$

In other words, $\zeta \in N_M^P(y)$ if and only if y is a minimizer of the function $f_\zeta(x) = \langle \zeta, x \rangle$ on the set M.

By KKT conditions (see [AEP20]), y minimizes f_ζ on M if and only if there exists

$\alpha \geq 0$ such that

$$-\nabla f_\zeta(y) + \alpha \nabla g(y) = 0.$$

Note that $\nabla f_\zeta(y) = \zeta$, which gives us $\zeta \in N_M^P(y)$ if and only if $\zeta = \alpha \nabla g(y)$ for some $\alpha \geq 0$. That is,

$$N_M^P(y) = \{\, \alpha \nabla g(y) \mid \alpha \geq 0 \,\}.$$

■

3 Pontryagin maximum principle

In this chapter, we formulate an optimal control problem and then follow with Clarke's formulation of Pontryagin maximum principle (the extended principle) stated in [Cla13, p. 465], which provides necessary conditions of optimality for optimal control problem. Then, we explain various standard versions of maximum principles used in optimal control theory (standard fixed-time principle, standard variable-time principle and hybrid principle). Finally, we briefly elaborate on the problematics of mixed constraints in optimal control.

3.1 Basic concepts

We begin with the introduction of some well-known concepts. Firstly, we make a few definitions (taken from [Cla13] and [Cla90]), which are necessary in the formulation of the maximum principle and we introduce the reader to optimal control problems.

► **Definition 3.1.**
Let $\langle a, b\rangle \subset \mathbb{R}$ be an interval. We say a function $f : \langle a, b\rangle \to \mathbb{R}$ is *absolutely continuous*, if there exists a (Lebesgue) measurable function $g : \langle a, b\rangle \to \mathbb{R}$ such that

$$f(x) = f(a) + \int_a^x g(t)dt$$

for all $x \in \langle a, b\rangle$.

Additionally, we say the vector function $f : \langle a, b\rangle \to \mathbb{R}^n$ is an *arc* if all its components are absolutely continuous functions. ◄

It follows that if $f : \langle a, b\rangle \to \mathbb{R}$ is absolutely continuous, it is continuous on $\langle a, b\rangle$ and differentiable for almost all $t \in \langle a, b\rangle$. Furthermore, we need to define a supremum norm of a function.

► **Definition 3.2.**
We define the *supremum norm* of a bounded function $f : \langle a, b\rangle \to \mathbb{R}^n$ as

$$\|f\|_\infty := \sup_{t\in\langle a,b\rangle} \{|f(t)|\}.$$

◄

The *controlled differential equation* is

$$x'(t) = f(t, x(t), u(t)), \quad u(t) \in U \quad \text{for all } t \in \langle a, b\rangle, \tag{3.1}$$

where $f : \langle a, b\rangle \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is a *dynamics function*, $U \subset \mathbb{R}^m$ is the *control set* (later, we use time dependent control set), $u : \langle a, b\rangle \to \mathbb{R}^m$ is a measurable function, the solution $x \in \mathbb{R}^n$ of (3.1) for a given initial condition $x(a) = x_0 \in \mathbb{R}^n$ is the *state*, $t \in \mathbb{R}$ is the *time variable*, the couple (x, f, U) is refered to as a *control system* and n is the *order of the system*. The equation (3.1) is also commonly referred to as the *state equation*. If the function f is independent of time, we say the system is *autonomous*. Let us clarify what we mean by the solution of (3.1). In an usual way (see [LR14]), we consider x is a solution of (3.1) with an initial condition $x(a) = x_0$, if it satisfies

$$x(t) = x_0 + \int_a^t f(\tau, x(\tau), u(\tau))d\tau \quad \text{for all } t \in \langle a, b\rangle. \tag{3.2}$$

From the definition of an arc we immediately see that x is an arc.

We typically want to enforce some specific behaviour of the control system (x, f, U) and we achieve it by choosing an appropriate control $u(t)$. The procedure of choosing the appropriate $u(t)$ can be referred to as a control problem. Control problem can have many different solutions. If we want to select only "the best" solution for our problem, we encounter an optimal control problem, where we want to find control u (and the state x) such that it minimizes a cost functional

$$J(x, u) := l(x(a), x(b)) + \int_a^b L(t, x(t), u(t))dt, \tag{3.3}$$

where $L : \langle a, b\rangle \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a *running cost* and $l : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a *boundary cost*. Running cost penalizes the system's behaviour during time interval $\langle a, b\rangle$. Boundary cost penalizes state at initial time and final time specifically. The function L is also commonly referred to as a *Lagrangian*. The couple (x, u) is referred to as a

process of the control system (x, f, U). When both the dynamics function and the Lagrangian are independent of t (and later also the set U is independent of t), we say the problem is autonomous. To proceed, it is necessary to define Hamiltonian.

► **Definition 3.3.**
The *Hamiltonian* function $H^\eta : \langle a, b\rangle \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ associated to the optimal control problem is

$$H^\eta(t, x, p, u) = \langle p, f(t, x, u)\rangle - \eta L(t, x, u),$$

where $\eta \in \{0, 1\}$ is a parameter, based on which we distinguish between normal ($\eta = 1$) and abnormal ($\eta = 0$) cases and $p : \langle a, b\rangle \to \mathbb{R}^n$ is a function. ◄

The function p is commonly referred to as a *costate*. It appears in the optimal control problem due to the fact that we want to minimize (3.3) while the state equation (3.1) holds. Hence, we encounter constrained optimization problem, where p plays a role of a multiplier enforcing the process satisfies the state equation. We can notice it holds that:

$$\nabla_p H^\eta = f(t, x, u) = x'.$$

When we divert our attention to examples, we will use symbol H for the Hamiltonian. In the theory, we use the symbol H^η to emphasise the dependence on η, even though we do not put η into the arguments of H^η. We will later encounter another type of Hamiltonian, when we study the presence of mixed constraints in optimal control, but it will be clear from the context, which Hamiltonian is referred to by H.

3.2 Extended maximum principle

The extended maximum principle is the most general maximum principle we provide in this chapter, because it allows more general problem formulation compared to standard versions of maximum principles presented in literature (which will be discussed later in this chapter). It is formulated for a fixed-time optimal control

problem (i.e., we consider a time interval of fixed length $\langle a, b\rangle$)

$$\begin{aligned} \min J(x,u) &= l(x(a),x(b)) + \int_a^b L(t,x(t),u(t))dt \\ \text{s.t.: } & x'(t) = f(t,x(t),u(t)), \quad \text{for almost all } t \in \langle a,b\rangle, \\ & u(t) \in U(t), \quad \text{for almost all } t \in \langle a,b\rangle, \\ & (x(a),x(b)) \in E, \end{aligned} \tag{EP}$$

where E is the set of desired states at initial and final time, $J(x,u)$ is the cost functional, which measures the quality of our solution. Note that instead of a traditional approach of defined initial condition, we assume an initial condition belongs to a set. Thereafter, when the initial condition belongs to a set, we refer to this as a *generalized initial condition*. This may be helpful for example in trajectory planning, when we may want to find an optimal initial condition from a set of all possible initial conditions.

We assume the following:

- The functions $f(t,x,u)$ and $L(t,x,u)$ are f, L are lower semicontinuous in (t,u) for each x;
- the graph $\operatorname{gr} U = \{\, (t,u) \in \langle a,b\rangle \times \mathbb{R}^m \mid u \in U(t) \,\}$ is a Borel set;
- The set $E \subset \mathbb{R}^n \times \mathbb{R}^n$ is closed, and l is locally Lipschitz.

Note that every closed set is Borel. For background information on Borel sets, we refer the interested reader to [Axl20, p. 136]. Our assumptions are sterner compared to what Clarke assumes in [Cla13].

A process (x,u) is *admissible* for (EP) if all the constraints are satisfied and $J(x,u)$ is well-defined and finite. We must state what we refer to by minimizing $J(x,u)$. We say a process (x^*,u^*) is a local minimizer if and only if for some $\varepsilon > 0$ and for every admissible process (x,u) satisfying $\|x - x^*\|_\infty < \varepsilon$, we have $J(x^*,u^*) \le J(x,u)$.

As an additional assumption for the extended principle, we consider a hypothesis, which is basically local Lipschitz property in state, admittting different constant for each time and control.

▶ **Hypothesis 3.4.**
There exists an lower semicontinuous function $k(t,u) : \operatorname{gr} U \to \mathbb{R}$, $(t,u) \mapsto k(t,u)$ such that, for almost every $t \in \langle a,b\rangle$, we have

$$x, y \in B(x^*(t),\varepsilon), \quad u \in U(t) \implies$$

$$| f(t,x,u) - f(t,y,u) | + | L(t,x,u) - L(t,y,u) | \leq k(t,u) | x - y |$$

and such that $t \mapsto k(t, u^*(t))$ is integrable. ◀

Now we formulate the extended principle, proved in [Cla13, p. 514].

► **Theorem 3.5 (Clarke).**
Let (x^*, u^*) be a local minimizer for (EP), where Hypothesis 3.4 holds. Then there exist an arc $p : \langle a, b\rangle \to \mathbb{R}^n$ together with a scalar η equal to 0 or 1 satisfying the **nontriviality condition**

$$(\eta, p(t)) \neq \mathbf{0} \quad \text{for all } t \in \langle a, b\rangle, \tag{3.4}$$

the **transversality condition**

$$(p(a), -p(b)) \in \eta \partial_L l(x^*(a), x^*(b)) + N_E^L(x^*(a), x^*(b)), \tag{3.5}$$

the **adjoint inclusion** for almost every t:

$$-p'(t) \in \partial_C H^\eta(t, \bullet, p(t), u^*(t))(x^*(t)), \tag{3.6}$$

as well as the **maximum condition** for almost every t:

$$H^\eta(t, x^*(t), p(t), u^*(t)) = \sup_{u \in U(t)} H^\eta(t, x^*(t), p(t), u). \tag{3.7}$$

If the problem is autonomous, then one may add to these conclusions the **constancy of the Hamiltonian**: for some constant h, we have

$$H^\eta(t, x^*(t), p(t), u^*(t)) = h \quad \text{a.e.}$$

◀

3.2.1 A commentary on the extended principle

Note that the maximum principle is a very complex multiplier rule. Indeed, (η, p) are multipliers. Multiplier η distinguishes between two cases. When $\eta = 0$, we encounter a degenerate case, when the constraints itself determine the optimal solution regardless of the cost functional. In this case, we can see that the necessary conditions provided by the extended principle do not take the cost functional into

account. We say the problem is normal, whenever $\eta = 1$. Then, the minimization of the cost functional verily happens.

Multiplier p ensures, that the state equation is fulfilled at almost every time, i.e., the state trajectory is admissible for the system. The state equation is a differential equation and therefore, it is logical that the maximum principle provides us with a differetial equation, which the costate p must satisfy. However, the expression (3.6) is not a differential equation, it is called a *differential inclusion*. The generalized gradient $\partial_C H^\eta(t, \bullet, p(t), u^*(t))(x^*(t))$ is a set, which, in general, can contain multiple elements. Therefore, we say, that the time derivative of p lies in this set for almost every t. If the set $\partial_C H^\eta(t, \bullet, p(t), u^*(t))(x^*(t))$ is a singleton for all t, the differential inclusion is reduced to a standard differential equation, as stated in [Cla+98] or [Cla90]. Anything less complex than differential inclusion (equation) would not be sufficient to enforce the state equation is satisfied.

Furthermore, we can see that the transversality condition is linking together the knowledge of the state and costate at the final (resp. initial) time. When the final (resp. initial) condition is prescribed (for the state), the limiting normal cone of the set E does not give any restriction on the final (resp. initial) value of the costate (i.e., this value can be arbitrary). This is caused by the fact that when we need to reach a pre-defined value of the state at the final (resp. initial) time, we need the freedom in costates to do so. Conversely, when we have freedom in choosing the final (resp. initial) destination of state, we restrict the final (resp. initial) set of admissible costates based on the boundary cost l, which leads to choosing the appropriate final (resp. initial) value of the state to minimize the cost functional. When the set E is defined by functional relations (equalities and inequalities), the transversality condition can be expressed by additional multipliers (for further information, see [Cla13]), which corresponds to the explanation above.

Observe that for the Hamiltonian function it holds that

$$\begin{aligned} -p'(t) &\in \partial_C H^\eta(t, \bullet, p(t), u^*(t))(x^*(t)), \\ x' &= \nabla_p H^\eta, \end{aligned}$$

we call them the *Hamiltonian canonical equations*. When H^η is differentiable with respect to x, it reduces to

$$\begin{aligned} -p' &= \nabla_x H^\eta, \\ x' &= \nabla_p H^\eta, \end{aligned}$$

which is a well-known concept.

3.2.2 Standard fixed-time maximum principle as a special case of the extended principle

In this section we consider a simpler (fixed-time) optimal control problem with sterner assumptions (compared to the assumptions given for the extended principle), which are often sufficient in applications. A standard form of maximum principle, which is indeed a specific instance of the extended principle, will be presented. We consider a problem

$$\begin{aligned} \min J(x,u) &= l(x(b)) + \int_a^b L(t,x(t),u(t))dt \\ \text{s.t.: } & x'(t) = f(t,x(t),u(t)), \quad \text{for almost all } t \in \langle a,b\rangle, \\ & u(t) \in U, \quad \text{for almost all } t \in \langle a,b\rangle, \\ & x(a) = x_0, \quad x(b) \in E, \end{aligned} \tag{SP}$$

where the meaning of symbols stays the same as in previous section with some exceptions, which we emphasise. A control set U is not time dependent and U is assumed to be compact. We consider a standard initial condition instead of the generalized initial condition and consequently, the boundary cost l becomes function of the final state solely, i.e., $l : \mathbb{R}^n \to \mathbb{R}$ and we call it a *terminal cost*. We limit ourselves to differentiable functions.

We consider the following hypothesis.

► **Hypothesis 3.6.**
The function l is continuously differentiable. Functions f and L are continuous, and admit derivatives $\nabla_x f(t,x,u)$ and $\nabla_x L(t,x,u)$ relative to x which are themselves continuous in all variables (t,x,u). ◄

Since this is a special case of the extended principle from the previous section, we assume the local minimizer for (SP) is defined in the same way as the local minimizer for (EP).

Now we formulate a standard version of the maximum principle.

► **Theorem 3.7.**
Let the process (x^*,u^*) be a local minimizer for the problem (SP) under Hypothesis 3.6, and where U is bounded. Then there exists an arc $p : \langle a,b\rangle \to \mathbb{R}^n$ and a scalar η equal to 0 or 1 satisfying the **nontriviality condition**

$$(\eta, p(t)) \neq \mathbf{0} \quad \text{for all } t \in \langle a,b\rangle,$$

the **transversality condition**

$$-p(b) \in \eta \nabla l(x^*(b)) + N_E^L(x^*(b)),$$

the **adjoint equation** for almost every t:

$$-p'(t) = \nabla_x H^\eta(t, x^*(t), p(t), u^*(t)),$$

as well as the **maximum condition** for almost every t:

$$H^\eta(t, x^*(t), p(t), u^*(t)) = \sup_{u \in U} H^\eta(t, x^*(t), p(t), u).$$

If the problem is autonomous, then one may add to these conclusions the **constancy of the Hamiltonian**: for some constant h, we have

$$H^\eta(t, x^*(t), p(t), u^*(t)) = h \quad \text{a.e.}$$

◀

Note the similarity between this theorem and the extended principle. This theorem is evidently a specific case of the extended principle. For detailed discussion see [Cla13].

3.3 Hybrid maximum principle

In this section, we focus on an application of the maximum principle for *hybrid systems*. A hybrid system is a system, which experiences different behaviour in various so-called *modes*. We assume a hybrid system of a specific form

$$\begin{aligned} x'(t) &= f(x(t), u(t)), \quad u(t) \in U, \quad \text{for almost all } t \in \langle 0, \tau \rangle, \\ y'(t) &= g(y(t), v(t)), \quad v(t) \in V, \quad \text{for almost all } t \in \langle \tau, T \rangle, \end{aligned}$$

where (x, f, U) and (y, g, V) are control systems of orders n_1 and n_2 respectively. We can see that this is a very special form of a hybrid system — it has only two modes and it experiences a switch between modes only at $t = \tau$. We formulate a

hybrid optimal control problem

$$\begin{aligned}
\min J(x,u) &= l_1(x(0)) + l_2(y(T)) + l_0(\tau, x(\tau), y(\tau)) \\
&\quad + \int_0^\tau L_1(x(t),u(t))dt + \int_\tau^T L_2(y(t),v(t))dt \\
\text{s.t.: } &\tau \in \langle 0,T\rangle, \\
&x'(t) = f(x(t),u(t)), \quad u(t) \in U, \quad \text{for almost all } t \in \langle 0,\tau\rangle, \\
&y'(t) = g(y(t),v(t)), \quad v(t) \in V, \quad \text{for almost all } t \in \langle \tau,T\rangle, \\
&x(0) \in E_1, \quad (\tau, x(\tau), y(\tau)) \in S, \quad y(T) \in E_2,
\end{aligned} \tag{HP}$$

where E_1 is the initial set, E_2 is the final set (note that initial and final conditions are imposed on the first and second systems respectively), $S \subseteq \langle 0,T\rangle \times \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$, $E_1 \subseteq \mathbb{R}^{n_1}, E_2 \subseteq \mathbb{R}^{n_2}$. We refer to $(\tau, x(\tau), y(\tau)) \in S$ as a linking condition, which may be used to impose constraints on the switching time τ, e.g., we may want to impose constraints that the switch occurs when the state x acquires some specific values and we may want to choose an "initial" condition for the state y at the switching time τ. We assume that $l_1, l_2, l_0, L_1, L_2, f, g$ are locally Lipschitz, the control sets U, V are bounded and the sets E_1, E_2, S are closed. Note that the hybrid problem (HP) is autonomous.

We consider a hybrid process (x,u), (y,v) with a switching time τ. We consider that x is defined beyond τ by constancy and y is defined prior to τ by constancy. Then, we can clarify what we understand by minimizing the cost functional. We say a hybrid process (x^*,u^*), (y^*,v^*) with a switching time τ^* admissible for (HP) is a local minimizer for (HP) provided that, for some $\varepsilon > 0$ and for any admissible hybrid process (x,u), (y,v) with switching time τ satisfying the following: $\|x - x^*\|_\infty < \varepsilon$, $\|y - y^*\|_\infty < \varepsilon$, $|\ \tau - \tau^*\ | < \varepsilon$, we have $J(x^*,y^*,\tau^*) \le J(x,y,\tau)$.

We define two Hamiltonians according to Definition 3.3: H_1^η is the Hamiltonian for the first system with the Lagrangian L_1 and costates p and H_2^η is the Hamiltonian for the second system with the Lagrangian L_2 and costates q. Now we proceed to the formulation of the hybrid principle.

► **Theorem 3.8.**
Let (x^*,u^*) and (y^*,v^*), with switching time $\tau^* \in (0,T)$, be a local minimizer for the problem (HP). Then there exist arcs p on $\langle 0,\tau^*\rangle$ and q on $\langle \tau^*,T\rangle$ and a scalar η equal to 0 or 1 satisfying the **nontriviality condition**

$$(\eta, p(\tau^*), q(\tau^*)) \neq \mathbf{0},$$

the **transversality conditions**

$$\begin{aligned} p(0) &\in \eta\partial_L l_1(x^*(0)) + N^L_{E_1}(x^*(0)),\\ -q(T) &\in \eta\partial_L l_2(y^*(T)) + N^L_{E_2}(y^*(T)), \end{aligned}$$

the **adjoint inclusions**

$$\begin{aligned} -p'(t) &\in \partial_C H^\eta_1(\bullet, p(t), u^*(t))(x^*(t)), \quad \text{for almost all } t \in \langle 0, \tau^*\rangle,\\ -q'(t) &\in \partial_C H^\eta_2(\bullet, q(t), v^*(t))(y^*(t)), \quad \text{for almost all } t \in \langle \tau^*, T\rangle, \end{aligned}$$

as well as the **maximum conditions**: for almost every t,

$$\begin{aligned} t \in \langle 0, \tau^*\rangle &\implies H^\eta_1(x^*(t), p(t), u^*(t)) = \sup_{u\in U} H^\eta_1(x^*(t), p(t), u),\\ t \in \langle \tau^*, T\rangle &\implies H^\eta_2(y^*(t), q(t), v^*(t)) = \sup_{v\in V} H^\eta_2(y^*(t), q(t), v). \end{aligned}$$

In addition, there exist constants h_1, h_2 such that the **constancy conditions** hold:

$$\begin{aligned} H^\eta_1(x^*(t), p(t), u^*(t)) &= h_1, \quad \text{for almost all } t \in \langle 0, \tau^*\rangle,\\ H^\eta_2(y^*(t), q(t), v^*(t)) &= h_2, \quad \text{for almost all } t \in \langle \tau^*, T\rangle, \end{aligned}$$

together with the **switching condition**:

$$(h_1 - h_2, -p(\tau^*), q(\tau^*)) \in \eta\partial_L l_0(\tau^*, x^*(\tau^*), y^*(\tau^*)) + N^L_S(\tau^*, x^*(\tau^*), y^*(\tau^*)).$$

◀

Note that as complicated as it may seem, this principle contains two maximum principles (each for every system on its respective time interval), transversality conditions for the initial condition of costate involves the first system and the transversality condition regarding the final costates value involves the second system. The switching condition resembles a transversality condition as well and it indeed plays a role of the transversality condition. This is caused by the fact that as we formulated the hybrid problem (HP), we imposed a linking condition on the values of states at the (optimal) switching time τ^*. For the first system, the linking condition plays a role of the (generalized) final condition and for the second system, it plays a role of the (generalized) initial condition, which defines the signs of h_1 and h_2 in the switching condition according to the extended maximum principle. Values h_1 and h_2 do not appear in the transversality conditions we have seen so far, but

they appear in this switching condition due to the fact that this principle is a first step to variable-time problem. The (optimal) switching time τ^* is chosen during the optimization, which is why the function l_0 in the linking condition may depend on τ as well as the set S has one time dimension and naturally, this additional variable occurs in the transversality condition (here switching condition). Note that it is sufficient that the nontriviality condition holds only at time $t = \tau^*$, which is a consequence of Gronwall's lemma, presented in [Cla13, p. 130] in the derivation of the hybrid principle.

The hybrid principle is derived from the extended principle in [Cla13, p. 468]. The main idea of the derivation is to rewrite a hybrid problem as an augmented problem admissible for the extended principle and derive necessary conditions for this augmented problem.

3.4 Standard variable-time maximum principle as a special case of the hybrid principle

In optimal control theory, the fixed-time problem is often not sufficient. We may encounter situations, when we want to optimize the length of the time interval as well. We formulate a variable-time problem

$$\begin{aligned} \min J(\tau, x, u) &= l(\tau, x(\tau)) + \int_0^\tau L(x(t), u(t))dt \\ \text{s.t.: } & \tau \geq 0, \\ & x'(t) = f(t, x(t), u(t)), \quad \text{for almost all } t \in \langle 0, \tau \rangle, \\ & u(t) \in U, \quad \text{for almost all } t \in \langle 0, \tau \rangle, \\ & x(a) = x_0, \quad (\tau, x(\tau)) \in S, \end{aligned} \tag{VP}$$

where we consider a variable length of the interval τ (often called a horizon). We assume the initial condition is given and the generalized final condition also includes the variable time τ. Notice that the cost functional as well as the terminal cost l now in general depend on τ. Note that the problem (VP) is autonomous.

3.4.1 Derivation of the variable-time principle

We use Theorem 3.8 (hybrid principle) to derive the standard variable-time maximum principle. Let $T > \tau^*$, $g = 0$, $L_2 = 0$, $l_2 = 0$, $E_2 = \mathbb{R}^n$, $L_1 = L$, $E_1 = \{x_0\}$, $l_1 = 0$, $l_0 = l$, $S = S$. We invoke necessary conditions of optimality. The **nontriviality**

condition says that $(\eta, p(\tau^*), q(\tau^*)) \neq \mathbf{0}$, but according to the derivation of the hybrid principle in [Cla13, p. 471], we can take the equivalent nontriviality condition in a form

$$(\eta, p(t), q(t), h_2 - h_1, -h_2) \neq 0 \quad \text{for all } t \in \langle 0, T\rangle. \tag{3.8}$$

Then, we have a **transversality conditions**

$$\begin{aligned} p(0) &\in \{\mathbf{0}\} + \mathbb{R}^n = \mathbb{R}^n, \\ -q(T) &\in \{\mathbf{0}\} + \{\mathbf{0}\} = \{\mathbf{0}\}, \end{aligned}$$

which gives us no information about the initial condition for p, but we know that $q(T) = \mathbf{0}$. The **adjoint inclusions** give us

$$\begin{aligned} -p'(t) &\in \partial_C H_1^\eta(\bullet, p(t), u^*(t))(x^*(t)), \quad t \in \langle 0, \tau^*\rangle \text{ a.e.}, \\ -q'(t) &\in \partial_C H_2^\eta(\bullet, q(t), v^*(t))(y^*(t)) = \{\mathbf{0}\}, \quad t \in \langle 0, \tau^*\rangle \text{ a.e.}, \end{aligned}$$

because $H_2^\eta = \langle q, g\rangle - \eta L_2 = 0 - \eta \cdot 0 = 0$. Together with the final condition $q(T) = \mathbf{0}$, we know that $q = \mathbf{0}$ almost everywhere on $\langle \tau^*, T\rangle$. The **maximum conditions** are reduced to maximum condition only for the first system

$$t \in \langle 0, \tau^*\rangle \implies H_1^\eta(x^*(t), p(t), u^*(t)) = \sup_{u \in U} H_1^\eta(x^*(t), p(t), u).$$

Because the second Hamiltonian $H_2^\eta = 0$, the maximum condition concerning the second Hamiltonian does not provide us with any new information. **Constancy conditions** give us $h_2 = 0$ and a condition for the first system

$$H_1^\eta(x^*(t), p(t), u^*(t)) = h_1, \quad \text{for almost all } t \in \langle 0, \tau^*\rangle.$$

Note that l_0 does not depend on y and $S \subset \mathbb{R} \times \mathbb{R}^n$, that is, S does not provide us with any information about $y^*(\tau^*)$ and therefore, the switching condition does not contain any information about $q(\tau^*)$. Recall that $h_2 = 0$. Due to that, the **switching condition** is reduced to:

$$(h_1, -p(\tau^*)) \in \eta \partial_L l_0(\tau^*, x^*(\tau^*)) + N_S^L(\tau^*, x^*(\tau^*)).$$

With knowledge provided by additional conditions of the hybrid principle, we

return back to the nontriviality condition (3.8):

$$\begin{aligned}(\eta, p(t), q(t), h_2 - h_1, -h_2) \neq 0 \quad &\text{for all } t \in \langle 0, T\rangle \implies\\ (\eta, p(t), -h_1) \neq 0 \quad &\text{for all } t \in \langle 0, T\rangle \implies\\ (\eta, p(t)) \neq 0 \quad &\text{for all } t \in \langle 0, \tau^*\rangle.\end{aligned}$$

Let us explain the second implication (the first one follows from the fact that $h_2 = 0$). When $\eta = 1$, the nontriviality condition is automatically satisfied. Let $\eta = 0$. The Hamiltonian of the first system $H_1^\eta = \langle p, f\rangle - \eta L = \langle p, f\rangle = h_1$ is constant. When $p = 0$, it must hold that $h_1 = 0$ and therefore, we may omit h_1 from the nontriviality condition. Finally, we know that $\tau^* < T$ and therefore, the condition holds on $\langle 0, \tau^*\rangle$. These derived relations for the first system represent necessary conditions for problem (VP) and will be summarized in the following section in a standard variable-time maximum principle.

3.4.2 Formulation of the variable-time principle

By the results of the previous section, we can formulate the maximum principle for general nonsmooth data. This nonsmooth version can be found in [Cla13, p, 460]. However, for our purposes, a smooth version is sufficient. Therefore, we restrict our attention to this case. We assume Hypothesis 3.6 holds and we assume the control set U is bounded.

Now we clarify what we understand by minimizing the cost functional. Let (x, u) be a process admissible for problem (VP) defined on $\langle 0, \tau\rangle$. We extend an arc x beyond τ by constancy. We say a process (x^*, u^*), defined on $\langle 0, \tau^*\rangle$ and admissible for problem (VP) is a local minimizer for (VP) if, for some $\varepsilon > 0$ and for all admissible processes (x, u) defined on $\langle 0, \tau\rangle$ satisfying $|\tau - \tau^*| < \varepsilon$ and $\|x - x^*\|_\infty < \varepsilon$, we have $J(\tau^*, x^*, u^*) \leq J(\tau, x, u)$.

► **Theorem 3.9.**
Let the process (x^*, u^*) defined on the interval $\langle 0, \tau^*\rangle$ $(\tau^* > 0)$ be a local minimizer for the problem (VP) under the hypotheses above. Then there exists an arc $p : \langle 0, \tau^*\rangle \to \mathbb{R}^n$ and a scalar η equal to 0 or 1 satisfying the **nontriviality condition**

$$(\eta, p(t)) \neq 0 \quad \text{for all } t \in \langle 0, \tau^*\rangle,$$

the **adjoint equation** for almost every $t \in \langle 0, \tau^*\rangle$

$$-p'(t) = \nabla_x H^\eta(x^*(t), p(t), u^*(t)),$$

as well as the **maximum condition**: for almost every $t \in \langle 0, \tau^* \rangle$

$$H^\eta(x^*(t), p(t), u^*(t)) = \sup_{u \in U} H^\eta(x^*(t), p(t), u),$$

and such that, for some constant h, we have **constancy of the Hamiltonian**:

$$H^\eta(x^*(t), p(t), u^*(t)) = h \quad \text{a.e.},$$

and the **transversality condition**

$$(h, -p(\tau^*)) \in \eta \nabla l(\tau^*, x^*(\tau^*)) + N_S^L(\tau^*, x^*(\tau^*)).$$

◀

Let $S = \{T\} \times E$. Then, there is no freedom in choosing the horizon τ, but we may still apply the variable-time principle. When the horizon is prescribed, the transversality condition provides us with no information about h and therefore, reduces to transversality condition of Theorem 3.7. Let us emphasise that Theorem 3.9 (variable-time principle) is not a generalization of Theorem 3.7, because Theorem 3.9 holds only for autonomus problems.

Consider a special case, when the choice of the horizon is arbitrary, that is, the set $S = \mathbb{R}_+ \times E$ and the terminal cost l is independent of the horizon. Then, the transversality condition gives us $h = 0$. Therefore, the Hamiltonian is necessarily equal to zero along the optimal trajectory.

3.5 Constancy of the Hamiltonian for standard fixed-time and variable-time principles

In this section, we present a proof of the constancy of the Hamiltonian for Theorem 3.7 and Theorem 3.9 for the case, when the problem is autonomous. Recall that we assume the validity of Hypothesis 3.6 for both of these theorems. Furthermore, we assume the control set U is compact.

The proof is inspired by Liberzon, as he suggests the possible approach to proving these theorems in [Lib12, p. 124]. In the proof, we consider variable-time principle, but it is evident, that we can use a fixed interval $\langle a, b \rangle$ instead of a variable interval $\langle 0, \tau^* \rangle$ and prove the constancy condition of the autonomous case of standard fixed-time principle. Recall that we discussed the possibility to use variable-time principle for autonomous fixed-time problems in section 3.4.2, which also justifies

our choice to prove the constancy of the Hamiltonian for the standard variable-time problem.

Proof. Let

$$\Phi_{\hat{t}}(t) = H^{\eta}(x^*(t), p(t), u^*(\hat{t})), \quad \text{where } \hat{t} \in \langle 0, \tau^* \rangle, \tag{3.9}$$

$$\Psi(t) = \max_{u \in U} H^{\eta}(x^*(t), p(t), u). \tag{3.10}$$

Recall that in the autonomous problem, the Hamiltonian does not depend on t. From Hypothesis 3.6, we know that $f, \nabla_x f, L, \nabla_x L$ are continuous and hence, the function H^{η} is continuous in all variables (x, p, u) and the function $\Phi_{\hat{t}}(t)$ is continuously differentiable for each $\hat{t} \in \langle 0, \tau^* \rangle$. From continuity of H^{η} and the fact that U is compact, the maximum condition of Theorem 3.7 and Theorem 3.9 holds everywhere and the supremum is reduced to maximum. The derivative of function $\Phi_{\hat{t}}(t)$ (for a fixed $\hat{t}$) at time t is

$$\Phi'_{\hat{t}}(t) = \langle \nabla_x H^{\eta}(x^*(t), p(t), u^*(\hat{t})), x^{*\prime}(t) \rangle + \langle \nabla_p H^{\eta}(x^*(t), p(t), u^*(\hat{t})), p'(t) \rangle.$$

We might be tempted to proclaim the derivative is equal to 0 using Hamiltonian canonical equations, which, as a matter of fact, would be incorrect, because recall we consider $u(\hat{t})$ instead of $u(t)$. This also allows us to compute the time derivative of $\Phi_{\hat{t}}$ without a time derivative of u, which is an essential part of the proof.

We show that the the absolute value of the derivative of $\Phi_{\hat{t}}(t)$ is bounded. We can proceed in a following way, using the triangle inequality and Cauchy-Schwarz inequality

$$\begin{aligned} \mid \Phi'_{\hat{t}}(t) \mid \leq & \mid \langle \nabla_x H^{\eta}(x^*(t), p(t), u^*(\hat{t})), x^{*\prime}(t) \rangle \mid \\ & + \mid \langle \nabla_p H^{\eta}(x^*(t), p(t), u^*(\hat{t})), p'(t) \rangle \mid \\ \leq & |\nabla_x H^{\eta}(x^*(t), p(t), u^*(\hat{t}))||x^{*\prime}(t)| + |\nabla_p H^{\eta}(x^*(t), p(t), u^*(\hat{t}))||p'(t)|. \end{aligned}$$

Let

$$\mathcal{X} = \{ x^*(t) \mid t \in \langle 0, \tau^* \rangle \},$$
$$\mathcal{P} = \{ p(t) \mid t \in \langle 0, \tau^* \rangle \}.$$

Both $x^*(t)$ and $p(t)$ are arcs, i.e., they are necessarily continuous. Thus, the sets $\mathcal{X}, \mathcal{P}$ are compact. We can recall that the Hamiltonian is also dependent on η. We know that $\eta \in \{0, 1\}$, which is indeed a compact set as well. Ultimately, recall that

we assumed U is compact. Therefore, the function $\nabla_x H$ is a continuous function on a compact set and therefore, we can bound its norm by a constant. We can bound it by its maximum, which is attained, because its domain is a compact set. Analogically, we can bound the norm of $\nabla_p H$. We label the bounds K_1 and K_2 respectively. States x and costates p are arcs defined on a compact set $\langle 0, \tau^* \rangle$ and therefore, we may bound their norms by constants C_1, C_2 as well. Thus, it holds that

$$| \Phi'_{\hat{t}}(t) | \leq K_1 C_1 + K_2 C_2$$

and hence, the derivative of $\Phi_{\hat{t}}(t)$ for a fixed $\hat{t}$ is bounded with a global bound independent of both t and $\hat{t}$ and consequently, the function $t \mapsto \Phi_{\hat{t}}(t)$ is Lipschitz for every $\hat{t}$.

Subsequently, we can write following inequalities

$$\Phi_t(\hat{t}) - \Phi_t(t) \leq \Psi(\hat{t}) - \Psi(t) \leq \Phi_{\hat{t}}(\hat{t}) - \Phi_{\hat{t}}(t), \tag{3.11}$$

which are an immediate consequence of definitions of $\Phi_{\hat{t}}$ and Ψ in equations (3.9) and (3.10), because we know that $\Phi_{\hat{t}}(t) \leq \Psi(t)$. It follows from inequalities (3.11) that Ψ is Lipschitz. Therefore, Ψ is differentiable almost everywhere.

Let $t \in \langle 0, \tau^* \rangle$ be such that the derivative $\Psi'(t)$ exists. Therefore, both one-sided derivatives exist and they are equal. We can bound the derivative $\Psi'(t)$ in a following way

$$\Psi'(t) = \lim_{\hat{t} \to t^+} \frac{\Psi(\hat{t}) - \Psi(t)}{\hat{t} - t} \geq \lim_{\hat{t} \to t^+} \frac{\Phi_t(\hat{t}) - \Phi_t(t)}{\hat{t} - t} = \Phi'_t(t) =$$
$$\langle \nabla_x H^\eta(x^*(t), p(t), u^*(t)), x^{*\prime}(t) \rangle + \langle \nabla_p H^\eta(x^*(t), p(t), u^*(t)), p'(t) \rangle = 0,$$

where we used the first inequality from (3.11) and Hamiltonian canonical equations. Analogically, we can compute the limit from the left side, which gives us

$$\Psi'(t) = \lim_{\hat{t} \to t^-} \frac{\Psi(\hat{t}) - \Psi(t)}{\hat{t} - t} \leq \lim_{\hat{t} \to t^-} \frac{\Phi_t(\hat{t}) - \Phi_t(t)}{\hat{t} - t} = \Phi'_t(t) =$$
$$\langle \nabla_x H^\eta(x^*(t), p(t), u^*(t)), x^{*\prime}(t) \rangle + \langle \nabla_p H^\eta(x^*(t), p(t), u^*(t)), p'(t) \rangle = 0.$$

Therefore, it must hold that

$$\Psi'(t) = \frac{d}{dt} \left(\max_{u \in U} H^\eta(x^*(t), p(t), u) \right) = 0$$

for almost all $t \in \langle 0, \tau^* \rangle$. As a result, we conclude that the constancy condition holds for almost all $t \in \langle 0, \tau^* \rangle$.

■

3.6 Optimal control problems with mixed constraints

In this section, we elaborate on the problematics of mixed constraints in optimal control. That is, the presence of constraints concerning both states and controls at the same time. We assume the geometrical form of mixed constraints expressed in the following way:

$$\varphi(t, x(t), u(t)) \in \Phi \quad \text{for all } t \in \langle a, b \rangle,$$

where $\Phi \subset \mathbb{R}^k$ for some $k \in \mathbb{N}$. We consider the following constrained optimal control problem

$$\begin{aligned} \min J(x,u) &= l(x(a), x(b)) + \int_a^b L(t, x(t), u(t))dt \\ \text{s.t.: } & x'(t) = f(t, x(t), u(t)), \quad \text{for almost all } t \in \langle a, b \rangle, \\ & u(t) \in U, \quad \text{for almost all } t \in \langle a, b \rangle, \\ & \varphi(t, x(t), u(t)) \in \Phi, \quad t \in \langle a, b \rangle, \\ & (x(a), x(b)) \in E. \end{aligned} \tag{CP}$$

We assume the functions l, f, L and φ satisfy Hypothesis 3.6 and the sets U, E and Φ are closed. This is a specific form of the mixed constrained optimal control problem, for which the necessary conditions can be expressed in the form of multipliers. More general constrained optimal control problem and theorem providing necessary for the problem is presented in [CP10], where necessary conditions are purely geometrical and they do not contain multipliers (additional multipliers compared to the problems without constraints). The special case is also outlined in [CP10, p. 8] and the theorem (necessary conditions) for a smooth case of the special case is presented in [Cla13, p. 536].

Before we proceed to the theorem, it is necessary to introduce constraint qualifications. We require the following hypothesis to hold.

► **Hypothesis 3.10.**
Let (x^*, u^*) be a local minimizer for problem (CP) and we assume u^* is bounded. Then, for every $t \in \langle a, b\rangle$, we have

$$u \in U, \quad \varphi(t, x^*(t), u) \in \Phi, \quad \lambda \in N_\Phi^L(t, x^*(t), u),$$
$$\mathbf{0} \in \nabla_u \langle \lambda, \varphi\rangle(t, x^*(t), u) + N_U^L(u) \implies \lambda = \mathbf{0}.$$

◄

Note that a new variable λ appeared and later we see that this will be the multiplier corresponding to mixed constraints. It is necessary to define an augmented Hamiltonian as

$$H_\varphi^\eta(t, x, p, u, \lambda) = H^\eta(t, x, p, u) - \langle \lambda, \varphi\rangle = \langle p, f(t, x, u)\rangle - \eta L(t, x, u) - \langle \lambda, \varphi\rangle$$

Now we proceed with theorem providing necessary conditions for problem (CP).

► **Theorem 3.11.**
Under the hypotheses above, there exists an arc $p : \langle a, b\rangle \to \mathbb{R}^n$, a scalar η equal to 0 or 1 and a bounded measurable function $\lambda : \langle a, b\rangle \to \mathbb{R}^k$ with

$$\lambda(t) \in N_\Phi^L(\varphi(t, x^*(t), u^*)) \quad \text{a.e.}$$

satisfying the **nontriviality condition**

$$(\eta, p(t)) \neq \mathbf{0} \quad \text{for all } t \in \langle a, b\rangle,$$

the **transversality condition**

$$(p(a), -p(b)) \in \eta \nabla l(x^*(a), x^*(b)) + N_E^L(x^*(a), x^*(b)),$$

the **adjoint equation** for almost every t:

$$-p'(t) = \nabla_x H_\varphi^\eta(t, x^*(t), u^*(t), \lambda)$$

as well as, for almost every t, the **maximum condition**

$$u \in U, \varphi(t, x^*(t), u) \in \Phi \implies H^\eta(t, x^*(t), p(t), u) \leq H^\eta(t, x^*(t), p(t), u^*(t))$$

and the **stationarity condition**

$$\mathbf{0} \in \nabla_u H_\varphi^\eta(t, x^*(t), p(t), u^*(t), \lambda(t)) - N_U^C(u^*(t)).$$

◀

One could easily think that the stationarity condition could be omitted, since it is the necessary condition corresponding to the maximum condition. However, it is necessary to notice that each of these conditions concerns a different Hamiltonian and hence, the stationarity condition provides us with additional information to the maximum condition.

Note that the constraint qualification (Hypothesis 3.10) automatically rules out pure state constraints. As it was mentioned at the beginning of this section, this is a special case of a more general theorem presented in [CP10] and one might wonder, if the general theorem allows pure state constraints. The answer is negative, pure state constraints are ruled out by the bounded slope condition in [CP10, p. 4].

Theorem 3.11 is formulated for the fixed-time problem (CP). If we encounter a variable-time problem and fulfill all assumptions of Theorem 3.11, we can use it as well. Suppose τ^* is the optimal time for our variable-time problem. Then, we can reformulate the problem to a fixed time problem with prescribed time interval $\langle 0, \tau^* \rangle$, for which our optimal solution is also an optimal solution. Then, we can use Theorem 3.11 for our augmented problem and it provides us with necessary conditions of optimality. Note that if we consider a minimal time problem, the augmented problem lacks the optimization, i.e., we are looking for any feasible solution of our problem.

Part II

Theory applied to examples

4 Navigational problem

Navigational problem (taken and modified from [Cla13, p. 546]) is a famous optimal control problem consisting of finding time-optimal trajectories to a target for a boat moving in (x, y)-plane through flowing water. The flow of water is represented by a two-dimensional vector field. The boat can be steered in any direction, but its speed relative to water cannot exceed given value.

4.1 Problem formulation

The problem can be formulated as follows

$$\begin{aligned} \min \quad & \tau = \int_0^\tau 1 dt \\ \text{s.t.:} \quad & \tau \geq 0, \\ & \begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} v \\ w \end{pmatrix} + \begin{pmatrix} c_x(x,y) \\ c_y(x,y) \end{pmatrix}, \\ & \left| \begin{pmatrix} v \\ w \end{pmatrix} \right| \leq V, \\ & \begin{pmatrix} x(0) \\ y(0) \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}, \quad \begin{pmatrix} x(\tau) \\ y(\tau) \end{pmatrix} \in E(\tau), \end{aligned} \tag{4.1}$$

where $V > 0$ is the maximum speed of a boat, $x_0, y_0 \in \mathbb{R}$ are the initial conditions, v, w are input variables (components of boat velocity relative to water), $c = (c_x, c_y)$ is continuously differentiable vector field describing water flow, $E(\tau)$ is the target set (notice that it is in general time dependent). We assume $(x_0, y_0) \in \mathbb{R}^2 \setminus E$, which ensures the problem is nontrivial. The problem will be solved for selected special cases further in this chapter.

4.1.1 Problem analysis

We can take functions L and l according to notation in chapter 3 as either $L = 1$ and $l = 0$ or $L = 0$ and $l = \tau$. Both of these approaches yield the same result. Let

$L = 1$ and $l = 0$. Then the Hamiltonian of the problem is:

$$H = -\eta L + \langle p, f \rangle = -\eta + p_1(v + c_x(x, y)) + p_2(w + c_y(x, y)), \tag{4.2}$$

where $\eta \in \{0, 1\}$. From the Hamiltonian we can write down costate equations as follows:

$$p' = -\nabla_{(x,y)} H(x, y, p_1, p_2, v, w) = -\begin{pmatrix} p_1 \frac{\partial c_x}{\partial x} + p_2 \frac{\partial c_y}{\partial x} \\ p_1 \frac{\partial c_x}{\partial y} + p_2 \frac{\partial c_y}{\partial y} \end{pmatrix} \quad \text{for almost all t } \in \langle 0, \tau^* \rangle. \tag{4.3}$$

Now we will invoke the maximum condition of Theorem 3.9 (variable-time principle). We want to maximize (4.2) with respect to (v, w). More precisely, we would like to find $(v^*(t), w^*(t))$ such that it is an element of the set

$$\underset{|(v,w)| \le V}{\operatorname{argmax}} H(x^*(t), y^*(t), p_1(t), p_2(t), v, w) = \underset{|(v,w)| \le V}{\operatorname{argmax}} \{p_1(t)v + p_2(t)w\},$$

for almost all $\mathrm{t} \in \langle 0, \tau^* \rangle$. We also have the nontriviality condition

$$(\eta, p_1(t), p_2(t)) \neq \mathbf{0} \quad \text{for all t } \in \langle 0, \tau^* \rangle \tag{4.4}$$

and transversality condition

$$\begin{aligned} (h, -p_1(\tau^*), -p_2(\tau^*)) &\in \nabla l(\tau^*, x(\tau^*), y(\tau^*)) + N_S^L(\tau^*, x(\tau^*), y(\tau^*)) \\ &= \{\mathbf{0}\} + \{0\} \times N_E^L(\tau^*, x(\tau^*), y(\tau^*)) = \{0\} \times N_E^L(\tau^*, x(\tau^*), y(\tau^*)), \end{aligned}$$

where $S := \mathbb{R}_+ \times E$. We can see that

$$H(x^*, u^*, p) = h = 0 \quad a.e. \tag{4.5}$$

The function $(v, w) \mapsto p_1 v + p_2 w$ is convex on the disk

$$M = \{(v, w) \in \mathbb{R}^2 | |(v, w)| \le V\}.$$

Consequently, the maximum is achieved on the boundary

$$\partial M = \{(v, w) \in \mathbb{R}^2 | |(v, w)| = V\}$$

of M. It can be easily seen that the maximizer belongs to the intersection of the circle

∂M with the halfline $\{s(p_1, p_2)|s \geq 0\}$. The situation is illustrated by the Figure 4.1. Dashed lines represent level sets of the objective function $(v, w) \mapsto p_1 v + p_2 w$.

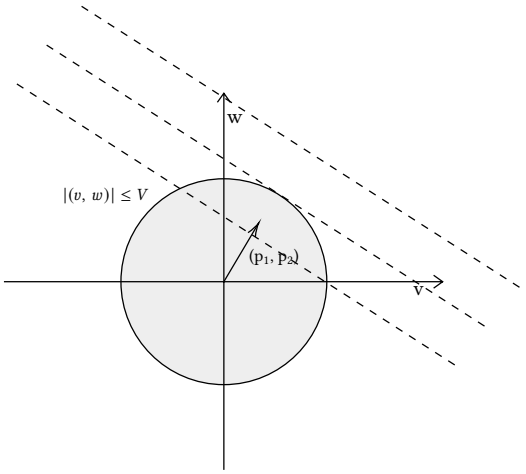


Figure 4.1: Maximization of Hamiltonian for navigational problem

This solution has only one degree of freedom and can be parametrized via a steering angle θ:

$$v = V \cos\theta, \quad w = V \sin\theta. \tag{4.6}$$

Thus

$$\tan\theta = \frac{p_2}{p_1}, \tag{4.7}$$

whenever $p_1 \neq 0$. Alternatively, we can use Theorem 2.20 to get this result as well. We can write our problem as

$$\max p_1 v + p_2 w - I_M,$$

where I_M is the indicator function of M. Then we can use Theorem 2.20 to derive necessary conditions of optimality:

$$\mathbf{0} \in \left\{ \begin{pmatrix} -p_1 \\ -p_2 \end{pmatrix} \right\} + N_M^L(\tilde{u}, \tilde{v}) = \left\{ \begin{pmatrix} -p_1 \\ -p_2 \end{pmatrix} + K \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \middle| K > 0, \theta \in \left\langle -\frac{\pi}{2}, \frac{3\pi}{2} \right\rangle \right\},$$

where $(\tilde{u}, \tilde{v})$ is the boundary point of M. From necessary conditions above, we derive precisely equation (4.7) for $p_1 \neq 0$.

When $p_1 = 0$, we have

$$\theta = \begin{cases} \pi/2 & \text{if } p_2 > 0 \text{ and } p_1 = 0, \\ -\pi/2 & \text{if } p_2 < 0 \text{ and } p_1 = 0, \end{cases}$$

which is optimal control for certain initial conditions (depending on the vector field and target set).

Note that $(p_1, p_2) \neq \mathbf{0}$ for all $t \in \langle 0, \tau^* \rangle$. Indeed, we know that there is $K > 0$ such that $p_1(t) = K \cos\theta(t)$ and $p_2(t) = K \cos\theta(t)$ almost everywhere on $\langle 0, \tau^* \rangle$. Thus $|(p_1(t), p_2(t))| = K \neq 0$ almost everywhere on $\langle 0, \tau^* \rangle$. Moreover, by continuity of p_1 and p_2, $|(p_1(t), p_2(t))|$ is a continuous function on $\langle 0, \tau^* \rangle$. This ensures that $|(p_1(t), p_2(t))| \neq 0$ everywhere on $\langle 0, \tau^* \rangle$ and so $(p_1(t), p_2(t)) \neq 0$ everywhere on $\langle 0, \tau^* \rangle$.

It follows from continuity of p_1 and p_2 that the function $\theta : \langle 0, \tau^* \rangle \to \mathbb{R}$ may be chosen such that it is continuous.

4.2 Simple linear vector field

In this section, we choose a simple vector field with components $c_x = -y$, $c_y = 0$. Firstly, we can write our state equation (substituting (4.6) and the form of vector field into state equation in (4.1)) as:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} -y + V\cos\theta \\ V\sin\theta \end{pmatrix}, \tag{4.8}$$

Now we write down the extract form of the costate equation from (4.3) as:

$$\begin{pmatrix} p_1' \\ p_2' \end{pmatrix} = \begin{pmatrix} 0 \\ p_1 \end{pmatrix},$$

which is a double integrator system. Therefore, its solution is

$$p_1(t) = A, \qquad p_2(t) = At + B,$$

where $A, B \in \mathbb{R}$.

Recall that $p_1(t) = K\cos\theta(t)$ and $p_2(t) = K\sin\theta(t)$ for some $K > 0$. Whenever $p_1(t) = A > 0$, $\cos\theta(t)$ has to be positive. Therefore, we may choose the function θ such that its range is contained in $(-\pi/2, \pi/2)$. More precisely, we may set $\theta(t) = \arctan(t + B/A)$ because $p_2(t)/p_1(t) = \tan\theta(t)$. This situation is shown in Figure 4.2. Similarly, if $A < 0$, then we take $\theta(t) = \arctan(t + B/A) + \pi$. If $A = 0$,

then $p_2(t) = B \neq 0$ and so θ may be chosen such that either $\theta(t) = \pi/2$ (when $B > 0$) or $\theta(t) = -\pi/2$ (when $B < 0$). It is worth to note that the function θ is continuously differentiable in all cases mentioned above. In the sequel, we restrict our attention to the nontrivial case (i.e., $A \neq 0$).

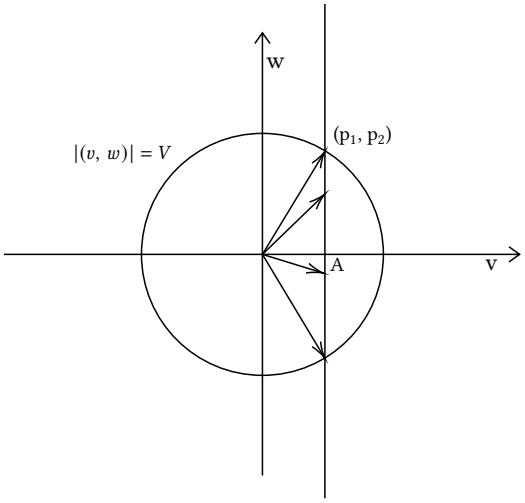


Figure 4.2: Vectors p_1, p_2 for boat steering when $A > 0$

Let $A \neq 0$. We have seen that θ is continuously differentiable and it holds that

$$\theta^*(t) = \begin{cases} \arctan\left(t + \frac{B}{A}\right) & \text{if } A > 0, \\ \arctan\left(t + \frac{B}{A}\right) + \pi & \text{if } A < 0. \end{cases} \tag{4.9}$$

and

$$\tan\theta = t + \frac{B}{A}. \tag{4.10}$$

Moreover, we can differentiate (4.10) to get

$$\theta' \frac{1}{\cos^2\theta} = 1$$

and from that we can express θ' as:

$$\theta' = \cos^2\theta. \tag{4.11}$$

Note that θ grows along the optimal trajectory.

Now we denote $\theta_{start} = \theta(0)$ and $\theta_{fin} = \theta(\tau^*)$. Inspired by [MG09], we compute

$$\frac{dy}{d\theta} = \frac{y'}{\theta'} = \frac{V \sin\theta}{\cos^2\theta}.$$

By integration, we obtain

$$y = \frac{V}{\cos\theta} + C_y = \frac{V}{\cos\theta} - \frac{V}{\cos\theta_{fin}} + y_{fin}, \tag{4.12}$$

where $y_{fin} := y(\tau^*)$. Now we substitute (4.12) into (4.8) to express x'. We may then apply the same procedure for state x and get

$$\frac{dx}{d\theta} = \frac{x'}{\theta'} = \frac{V\cos\theta - \frac{V}{\cos\theta} + \frac{V}{\cos\theta_{fin}} - y_{fin}}{\cos^2\theta}. \tag{4.13}$$

Hence,

$$x = \int \frac{V\cos\theta - \frac{V}{\cos\theta} + \frac{V}{\cos\theta_{fin}} - y_{fin}}{\cos^2\theta} d\theta + C_x. \tag{4.14}$$

The constant C_x is then chosen to fulfill the final condition $x_{fin} := x(\tau^*)$. We can substitute initial conditions x_0, y_0 together with θ_{start} into both equations (4.12) and (4.14) to get the relation of initial conditions and initial steering angle. If E is a singleton (i.e., there is no freedom in the choice of the final destination (x_{fin}, y_{fin}) of the boat), we have these two equations for two unknowns θ_{start} and θ_{fin}. If not, we will see later that the transversality condition of Pontryagin maximum principle provides us with additional equations to solve the problem even if we do not know x_{fin} and y_{fin} precisely (there is freedom in the choice of the final destination).

From (4.10) we can get $\frac{B}{A} = \tan\theta_{start}$ and therefore, it holds that

$$\tau^* = \tan\theta_{fin} - \tan\theta_{start} \tag{4.15}$$

and hence, the optimal control is

$$\theta^*(t) = \begin{cases} \arctan(t + \tan\theta_{start}) & \text{if } |\theta_{start}| < \pi/2, \\ \arctan(t + \tan\theta_{start}) + \pi & \text{if } |\theta_{start}| > \pi/2. \end{cases} \tag{4.16}$$

Naturally, a question, whether the set E is reachable from the given initial condition arises. In general, the problem may not have a solution. For more information on reachability, see [BP07].

4.2.1 Steering to origin

In this section, we will solve an instance of problem (4.1) with $E = \mathbb{R}_+ \times E$. The set $S := (\mathbb{R}_+, E)$ and its limiting normal cone is: $N_S^L((t,0,0)) = \{0\} \times \mathbb{R}^2$ for $t \in \langle 0, \tau^* \rangle$ and the transversality condition gives us:

$$\begin{aligned}(h, -p_1(\tau^*), -p_2(\tau^*)) &= \nabla l(\tau^*, x(\tau^*), y(\tau^*)) + N_S^L(\tau^*, x(\tau^*), y(\tau^*)) \\ &= \{\mathbf{0}\} + \{0\} \times \mathbb{R}^2 = \{0\} \times \mathbb{R}^2.\end{aligned}$$

Using $y_{fin} = 0$ in (4.12), we obtain

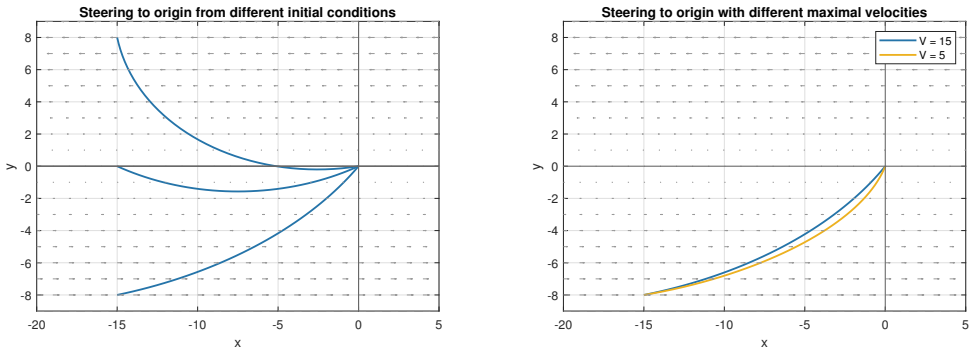
$$y = \frac{V}{\cos\theta} - \frac{V}{\cos\theta_{fin}}.$$

Now we can substitute $x_{fin} = 0$ into (4.14), which gives us

$$\begin{aligned}0 = C_x &+ 5\,\mathrm{atanh}\left(\tan\left(\frac{\theta_{fin}}{2}\right)\right) \\ &- \frac{\left(5\,\cos(\theta_{fin}) + 10\right)\tan\left(\frac{\theta_{fin}}{2}\right)^3 + \left(5\,\cos(\theta_{fin}) - 10\right)\tan\left(\frac{\theta_{fin}}{2}\right)}{\cos(\theta_{fin})\tan\left(\frac{\theta_{fin}}{2}\right)^4 - 2\,\cos(\theta_{fin})\tan\left(\frac{\theta_{fin}}{2}\right)^2 + \cos(\theta_{fin})},\end{aligned}$$

which determines the value of constant C_x of the expression (4.14) for x.

With the knowledge of C_x, we have x and y both expressed as functions of θ and θ_{fin}. Consequently, if we evaluate these functions at $t = 0$, we obtain two nonlinear equations for two unknowns θ_{start} and θ_{fin}. We solve them numerically using *vpasolve* function of Symbolic Math Toolbox in Matlab to get θ_{start} and θ_{fin} for a given initial value. After that, the optimal time τ^* and the optimal control $\theta^*(t)$ are immediately determined by (4.15) and (4.16). Optimal trajectories for different initial conditions and maximal speed $V = 16$ are shown in the Figure 4.3 (a) and trajectories from $(x_0, y_0) = (-15, -8)$ for different maximal values of speed ($V = 15$ and $V = 5$) are shown in the Figure 4.3 (b). Trajectories for different values of maximal speed look similar, but the one with smaller maximal speed is more influenced by flowing water and the optimal times vary greatly. Optimal time for $V = 20$ is 0.72 and optimal time for $V = 5$ is 2.05 (expressed in suitable units).



(a) Steering boat to origin from different initial conditions

(b) Steering boat to origin with different maximal velocities

Figure 4.3: Simulations for steering boat to origin

4.2.2 Steering to right half-plane

Now we consider a different target set $E = \left\{ (x, y) \in \mathbb{R}^2 \,\middle|\, x \geq 0 \right\}$. At $t = 0$, we assume the boat is located at left half-plane. Hamiltonian, state and costate equations are all the same. The difference occurs in the transversality condition

$$\begin{aligned}(h, -p(\tau^*)) &\in \nabla l(\tau^*, x(\tau^*), y(\tau^*)) + N_S^L(\tau^*, x(\tau^*), y(\tau^*)) \\ &= \{\mathbf{0}\} + \{(0, -r, 0) | r > 0\} \\ &= \{(0, -r, 0) | r > 0\},\end{aligned}$$

which gives us final condition for p in the form

$$-\begin{pmatrix} A \\ A\tau^* + B \end{pmatrix} = \begin{pmatrix} -r \\ 0 \end{pmatrix} \tag{4.17}$$

for some $r > 0$. That automatically excludes the option of $\theta = \pi/2$ along the trajectory. It follows from (4.17) that $A = r > 0$ and $\tau^* = -B/A$. Using (4.9) with (4.17) and the fact that $A > 0 \implies \theta(t) \in (-\pi/2, \pi/2)$ for all $t \in \langle 0, \tau^* \rangle$, we see that

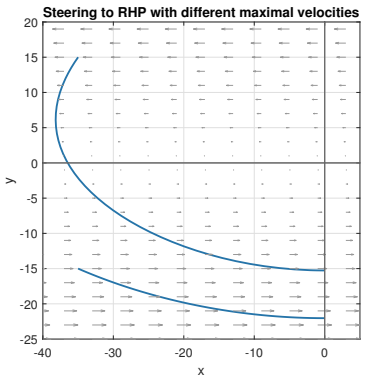
$$\theta_{fin} = 0.$$

We now follow the exact same procedure as in section 4.2. More concretely, we compute $\frac{dy}{d\theta}$ and then integrate to fulfill the final condition (in this case $\theta_{fin} = 0$) to

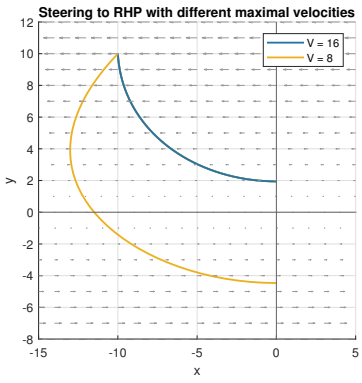
obtain

$$y = \frac{V}{\cos\theta} + y_{fin} - V, \tag{4.18}$$

where $y_{fin} \in \mathbb{R}$ is not determined in this case. Then we substitute (4.18) into the state equation (4.8) and compute $\frac{dx}{d\theta}$. Then we can integrate this expression to fulfill the final condition. After that, we can substitute initial conditions into expressions for x and y and solve this system of nonlinear equations numerically to get y_{fin}, θ_{fin} and θ_{start}. We can see that the transversality condition provided us with a third equation $\theta_{fin} = 0$, but we have one more unknown variable y_{fin}. Hence, we have a set of three nonlinear equations for three variables. Eventually, we get optimal time and optimal control from equations (4.15) and (4.16). Optimal trajectories for different initial conditions and maximal speed $V = 16$ are shown in the Figure 4.3 (a) and trajectories from $(x_0, y_0) = (-10, 10)$ for different maximal speeds $V = 8$ and $V = 16$ are shown in the Figure 4.3 (b). Smaller maximal velocity causes a longer trajectory which is more influenced by water flow. Compared to the problem of steering to origin, the difference between these two trajectories is more significant, because different maximal speed causes different end point of the trajectory (there is freedom in finding the optimal y_{fin}).



(a) Steering boat to right half-plane from different initial conditions



(b) Steering boat to right half-plane with different maximal velocities

Figure 4.4: Simulations for steering boat to right half-plane

4.2.3 Steering to an ellipse

Now we solve problem (4.1) with the same linear vector field as in previous sections, but with more complicated target set $E = \left\{ (x, y) \in \mathbb{R}^2 \;\middle|\; \frac{x^2}{a^2} + \frac{y^2}{b^2} - 1 \leq 0 \right\}$ for given $a, b > 0$, which is an ellipse. We need to compute $N_E^L(x)$, which is reduced to proximal normal cone according to Theorem 2.24. Then, we can use Theorem 2.25 to compute $N_E^P(x)$. We differentiate function $h(x, y) = \frac{x^2}{a^2} + \frac{y^2}{b^2} - 1$. By differentiation we get $\nabla h = \left(\frac{2x}{a^2}, \frac{2y}{b^2}\right)$. Note that a parametrization of E is

$$x(\alpha) = a \cos \alpha, \quad y(\alpha) = b \sin \alpha, \quad \alpha \in (-\pi, \pi\rangle.$$

The gradient of h can be then written as $\nabla h(x(\alpha), y(\alpha)) = \left(\frac{2 \cos \alpha}{a}, \frac{2 \sin \alpha}{b}\right)$. The cone $N_E^L(x(\alpha), y(\alpha))$ consists of all vectors of the form $\tilde{K} \nabla h(x(\alpha), y(\alpha))$ for some positive constant $\tilde{K}$. The transversality condition says that

$$\begin{aligned}(h, -p(\tau^*)) &\in \nabla l(\tau^*, x(\tau^*), y(\tau^*)) + N_S^L(\tau^*, x(\tau^*), y(\tau^*)) \\ &= \left\{ \left(0, \tilde{K} \frac{2 \cos \alpha}{a}, \tilde{K} \frac{2 \sin \alpha}{b}\right) \middle| \tilde{K} > 0 \right\} = \left\{ \left(0, K \frac{\cos \alpha}{a}, K \frac{\sin \alpha}{b}\right) \middle| K > 0 \right\},\end{aligned}$$

which gives us the final condition for p as:

$$-\begin{pmatrix} A \\ A\tau^* + B \end{pmatrix} = K \begin{pmatrix} \frac{\cos \alpha}{a} \\ \frac{\sin \alpha}{b} \end{pmatrix}. \tag{4.19}$$

Using (4.19) and (4.9), we get relation between α_{fin} (position on ellipse) and θ_{fin} (steering angle at $t = \tau^*$)

$$\tan \theta_{fin} = \frac{a}{b} \tan \alpha_{fin}, \tag{4.20}$$

where α_{fin} defines the point on ellipse where the boat reaches the ellipse. This equation itself has four solutions for values θ_{fin} and α_{fin} and we must choose the correct one. From the geometrical nature of the problem, the directional vector of the boat at the final time v_{boat} and the normal vector n of the ellipse at the point, where the boat reaches the ellipse form an obtuse angle, that is,

$$\langle n, v_{boat} \rangle = \left\langle \begin{pmatrix} \frac{\cos \alpha_{fin}}{a} \\ \frac{\sin \alpha_{fin}}{b} \end{pmatrix}, \begin{pmatrix} V \cos \theta_{fin} - y_{fin} \\ V \sin \theta_{fin} \end{pmatrix} \right\rangle < 0.$$

This condition eliminates two solutions. Then, we would be able to choose between the remaining two solutions had we known the value of A. However, we do not have the exact value of A (recall that we deliberately chose not to look for the exact time development of costates at the beginning of the chapter) and therefore, we may make an apriori assumption either $|\theta_{fin}| < \pi/2$ or $|\theta_{fin}| > \pi/2$, which distinguishes between the two branches of inverse tangent. If our apriori guess is incorrect (we do not get a solution), we try the second option. Needless to say, it could also happen that $|\theta_{fin}| = \pi/2$. This can indeed be a valid solution, which is constant. However, we solve this problem numerically and our solution is only an approximation of the real solution. With numerical methods, we may leave out the option of $|\theta_{fin}| = \pi/2$ and settle with a numerical approximation of this solution, which falls into one of the categories $|\theta_{fin}| < \pi/2$ or $|\theta_{fin}| > \pi/2$.

We can apply the same procedure once again to obtain an expression for y and x from equations (4.12) and (4.14), which contain unknown variables y_{fin} and x_{fin}. Evaluations of these expressions at $t = 0$ leads to the first two equations for five unknowns $\theta_{fin}, \alpha_{fin}, \theta_{start}, x_{fin}, y_{fin}$. The remaining three needed equations are $x_{fin} = a\cos\alpha_{fin}$, $y_{fin} = b\cos\alpha_{fin}$ and (4.20). We solve this system of equations numerically. We get optimal time and optimal control again from equations (4.15) and (4.16). Simulations for different initial conditions, maximal speed $V = 16$ and semi-axes $a = 1$, $b = 2$ are shown in Figure 4.5.

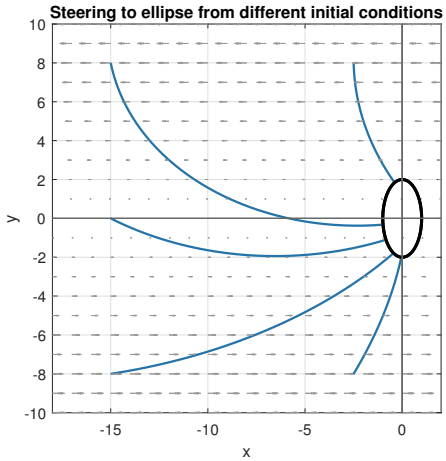


Figure 4.5: Steering boat to an ellipse from different initial conditions

4.2.4 Steering to a point drifted by currents

In this subtask, we solve the problem (4.1) with vector field (4.2) and a time dependent target set E. We assume we must steer a boat to a point drifted by currents. We assume that the coordinates x_p and y_p of that point at time $t = 0$ are known. Therefore, we know the exact position of the point at any time, which gives us the time dependent target set

$$E(\tau) = \left\{ \begin{pmatrix} x_p - \tau y_p \\ y_p \end{pmatrix} \right\},$$

where τ is the final time. Due to the fact that at optimal final time τ^*, the boat must reach E, we have time dependent final constraints

$$x_{fin} = x_p - \tau^* y_p, \qquad y_{fin} = y_p,$$

which can be substituted directly into (4.12) and used to obtain corresponding integration constant from (4.14). The transversality condition does not provide us with any relevant information (analogy with steering to origin, the target set is a singleton). Then we have x and y expressed as functions of $\theta_{start}, \theta_{fin}, t$ and τ^*. By evaluation of these equations at $t = 0$, we get two nonlinear equations for three unknown variables. When we add equation (4.15), we have three nonlinear equations for three unknowns $\theta_{start}, \theta_{fin}$ and τ^*, which can be solved numerically.

We solved the problem for different initial conditions. Simulations for a starting position of the moving point $(x_p, y_p) = (-2, 5)$ from various initial conditions of the boat and maximal boat speed $V = 15$ are shown in the Figure 4.6.

The trajectories of the boat are blue and the trajectory of the point is black. Note that when the boat starts in various initial positions, it reaches the point at different times and positions.

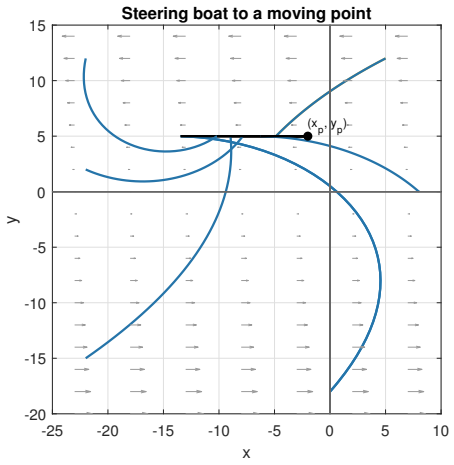


Figure 4.6: Steering boat to a point drifted by currents from different initial conditions

4.2.5 Steering to a moving ellipse

Now we consider a problem of steering boat to an ellipse moving with a constant speed U along a straight line. The ellipse is determined by its center (x_e, y_e), semi-axes a, b and angle φ between axis x and semi-axis a. Ellipse is moving along a line $y = x \tan\varphi + c$. Parametrization of an ellipse for $\varphi = 0$ and center at the origin is

$$x = a\cos\alpha, \quad y = b\sin\alpha, \quad \alpha \in (-\pi, \pi\rangle.$$

Limiting normal cone for $\varphi = 0$ is $N_E^L(x(\alpha), y(\alpha)) = \left\{\left(K\frac{\cos\alpha}{a}, K\frac{\sin\alpha}{b}\right)\middle| K > 0\right\}$ as it was shown earlier. To obtain limiting normal cone $N_{E_\varphi}^L(x(\alpha), y(\alpha))$ and parametrization of a rotated ellipse (with a center in the origin), one has to transform it by a rotation matrix

$$R_\varphi = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}$$

as

$$\begin{aligned} N_{E_\varphi}^L(x(\alpha), y(\alpha)) &= \left\{R_\varphi\zeta \middle| \zeta \in N_E^L(x(\alpha), y(\alpha))\right\} \\ &= \left\{K\begin{pmatrix} \frac{1}{a}\cos\alpha\cos\varphi - \frac{1}{b}\sin\alpha\sin\varphi \\ \frac{1}{a}\cos\alpha\sin\varphi + \frac{1}{b}\sin\alpha\cos\varphi \end{pmatrix}\middle| K > 0\right\}, \\ \begin{pmatrix} x(\alpha) \\ y(\alpha) \end{pmatrix} = R_\varphi\begin{pmatrix} a\cos\alpha \\ b\sin\alpha \end{pmatrix} &= \begin{pmatrix} a\cos\alpha\cos\varphi - b\sin\alpha\sin\varphi \\ a\cos\alpha\sin\varphi + b\sin\alpha\cos\varphi \end{pmatrix}. \end{aligned}$$

Now we proceed to the motion model of the center (x_e, y_e) of the ellipse

$$x_e' = U\cos\varphi, \quad y_e' = U\sin\varphi,$$

which gives us time development of the center of the ellipse

$$x_e(t) = tU\cos\varphi + x_{e0}, \quad y_e(t) = tU\sin\varphi + y_{e0},$$

where (x_{e0}, y_{e0}) is the known center of the ellipse at time $t = 0$. Therefore, for the optimal time $t = \tau^*$ it must hold that

$$x_{e1} = \tau^* U\cos\varphi + x_{e0}, \quad y_{e1} = \tau^* U\sin\varphi + y_{e0},$$

where (x_{e1}, y_{e1}) is the unknown center of an ellipse at time $t = \tau^*$. We know that at $t = \tau^*$, the boat must reach the ellipse, hence

$$x_{fin} = a\cos\alpha\cos\varphi - b\sin\alpha\sin\varphi + \tau^* U\cos\varphi + x_{e0}, \tag{4.21}$$
$$y_{fin} = a\cos\alpha\sin\varphi + b\sin\alpha\cos\varphi + \tau^* U\sin\varphi + y_{e0}. \tag{4.22}$$

From the transversality condition (using the rotated limiting normal cone) we get

$$-\begin{pmatrix} A \\ A\tau^* + B \end{pmatrix} = K\begin{pmatrix} \frac{1}{a}\cos\alpha\cos\varphi - \frac{1}{b}\sin\alpha\sin\varphi \\ \frac{1}{a}\cos\alpha\sin\varphi + \frac{1}{b}\sin\alpha\cos\varphi \end{pmatrix}.$$

We use transversality condition in (4.10) and evaluate at $t = \tau^*$ to obtain

$$\tan\theta_{fin} = \frac{\frac{1}{a}\cos\alpha_{fin}\sin\varphi + \frac{1}{b}\sin\alpha_{fin}\cos\varphi}{\frac{1}{a}\cos\alpha_{fin}\cos\varphi - \frac{1}{b}\sin\alpha_{fin}\sin\varphi},$$

where α_{fin} determines the position on the ellipse where the boat reaches the ellipse. This equation admits four solutions for values α_{fin} and θ_{fin} and therewith, we use the same approach as in equation (4.20) to select the correct one.

From (4.15) we get

$$\tau^* = \frac{\frac{1}{a}\cos\alpha_{fin}\sin\varphi + \frac{1}{b}\sin\alpha_{fin}\cos\varphi}{\frac{1}{a}\cos\alpha_{fin}\cos\varphi - \frac{1}{b}\sin\alpha_{fin}\sin\varphi} - \tan\theta_{start}. \tag{4.23}$$

We follow the already well known procedure to obtain expressions for both y and x from (4.12) and (4.14), which depend on unknowns y_{fin} and x_{fin}. Then we evaluate these expressions at time $t = 0$ and use them together with (4.21), (4.22) and (4.23)

to solve a system of five nonlinear equations for five unknowns x_{fin}, y_{fin}, θ_{start}, α_{fin} and τ^* numerically. Optimal time and control are determined by (4.15) and (4.16).

Simulations illustrating three scenarios are shown in the Figure 4.7 for the ellipse speed $U = 2$ and boat speed $V = 4$. Ellipses at their starting position are gray and ellipses as $t = \tau^*$ are black. Trajectories are blue and yellow. The yellow trajectory is especially interesting, because we can see the boat reaches the ellipse almost tangentially. Notice, that at the final time $t = \tau^*$, the ellipse is subject to very strong current relative to its speed. The steering angle at the final time $t = \tau^*$ is indeed a normal vector of the ellipse at the point of intersection. It is the strength of these currents that causes the final angle of the boat in $(x, y)-$plane and the steering angle at $t = \tau^*$ to differ significantly.

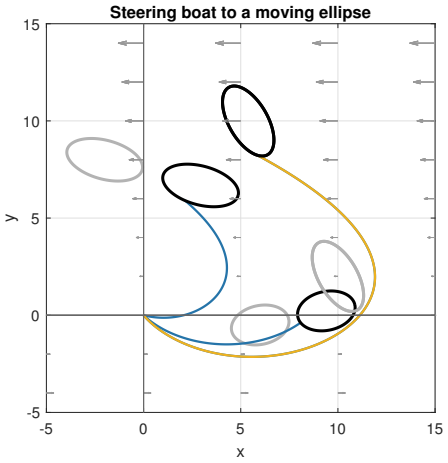


Figure 4.7: Steering boat to an ellipse moving with a constant speed along a straight line

4.3 Quadratic vector field

The vector field used in previous examples is rather simple. Let us analyze and solve problem (4.1) with the vector field with components $c_x = -y(y-1)$ and $c_y = 0$, which give us the state equations

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} -y(y-1) + V\cos\theta \\ V\sin\theta \end{pmatrix}. \tag{4.24}$$

The motivation for considering this particular vector field comes from a model of the water flow in a straight river with banks defined by lines $y = 0$ and $y = 1$. Naturally, this makes the problem more complex, because we impose a pure state constraint $0 \leq y \leq 1$. Let us ignore this state constraint for now. Later, we will say a few words about the case where the state constraint $0 \leq y \leq 1$ necessarily appears.

Costate equations for our problem are

$$\begin{pmatrix} p_1' \\ p_2' \end{pmatrix} = \begin{pmatrix} 0 \\ -p_1(1-2y) \end{pmatrix}.$$

We immediately see that the first costate is constant

$$p_1 = A, \tag{4.25}$$

where $A \in \mathbb{R}$. We again distinguish two cases. Let $A = 0$. Then we have constant optimal control $\theta^*(t) = \pm\pi/2$. If $A \neq 0$, then equation (4.7) holds. With justification from the beginning of this chapter, we can differentiate (4.7) to obtain:

$$\theta' = (2y-1)\cos^2\theta. \tag{4.26}$$

Note that this differential equation holds for the specific case when $\theta = \pm\pi/2$ as well. We could have proceeded the same way as in the case of the linear vector field to obtain equations for the development of x and y, but the integral for x coordinate development in time has no analytical formula. Therefore, we have to employ numerical integration to solve a specific instance of the problem.

4.3.1 Steering to an opposite river bank

Firstly, the problem will be solved for $x_0 \in \mathbb{R}$, $y_0 = 0$ (a defined point at the first bank of the river) and $E(\tau) = \left\{ (x_f, 1) \,\middle|\, x_f \in \mathbb{R} \right\}$ (we steer to any point at the second bank). The limiting normal cone of set E is $N_E^L(x,1) = \{\, (0,r) \mid r \in \mathbb{R} \,\}$, which makes the transversality condtion

$$(h, -p(\tau^*)) \in \nabla l(\tau^*, x(\tau^*), y(\tau^*)) + N_S^L(\tau^*, x(\tau^*), y(\tau^*)) = \{(0,0,r) | r \in \mathbb{R}\}.$$

This transversality condition itself provides us with the optimal solution, because from (4.25) we know that $p_1 = 0$ on $\langle 0, \tau^* \rangle$ and therefore, $\theta^* = \pm\pi/2$ on $\langle 0, \tau^* \rangle$. Note that we can rule out the solution $\theta^* = -\pi/2$ because the boat must reach the destination from the river and not from the shore.

We compute the point on the bank $y = 1$, where the boat reaches the bank. It is the solution of the initial value problem

$$\begin{aligned} x' &= -y(y-1), \\ y' &= V, \\ x(0) = x_0, y(0) &= 0. \end{aligned}$$

We can solve it analytically to obtain

$$\begin{aligned} y(t) &= Vt, \\ x(t) &= -\frac{V^2}{3}t^3 + \frac{V}{2}t^2 + x_0, \end{aligned}$$

from which we can compute the optimal time $\tau^* = 1/V$, which can be substituted into the expression for x time development to get the final value $x_{fin} = x_0 + \frac{1}{6V}$.

4.3.2 Steering to a specific point on the opposite bank

Now we consider the problem for $x_0 \in \mathbb{R}$, $y_0 = 0$, $E = \{(x_1, 1)\}$ (two prescribed points at opposite river banks). Its solution is not so simple because we need to incorporate state constraints, which is challenging with the usage of Pontryagin maximum principle. As it was stated in chapter 3, according to [Cla13], Theorem 3.11 is a multiplier rule for mixed constraints. However, our pure state constraints can never fulfill the required constraint qualifications (Hypothesis 3.10), because the derivative with respect to u is always zero for pure state constraints. This problematics is explained in [BPV16].

At first we can try to solve the problem relaxation without these constraints. Therefore, we can use derivation of steering angle dynamics (4.26), which was presented above. This is a boundary value problem, which can be solved using collocation method (explained in appendix B) or a Matlab function *bvp4c*. The results are shown in Figure 4.8.

The blue trajectory represents boat steering from starting point $(-5, 0)$ to end point $(0, 1)$ and the yellow trajectory represents boat steering from starting point $(0, 0)$ to end point $(-5, 1)$. We can see that the blue trajectory satisfies the constraints and therefore, we can state that the blue trajectory is indeed a solution of the original problem. However, the yellow trajectory evidently does not fulfill constraints and due to that, it cannot be a solution of the original problem. The yellow trajectory crosses the borders because in the river, the boat has to go against the stream and

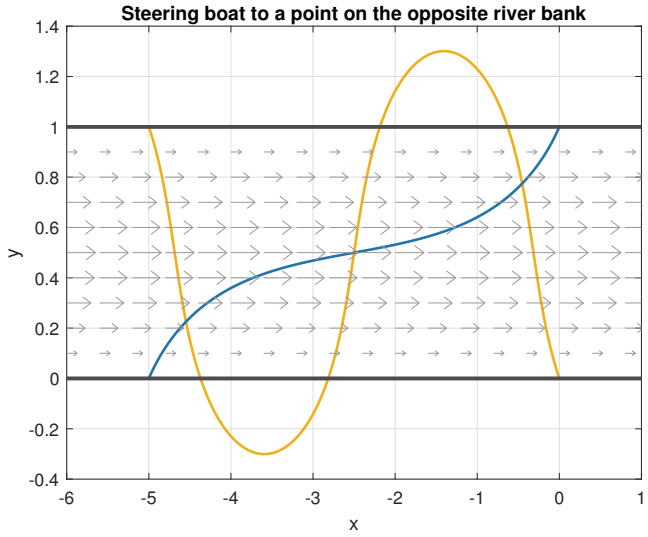


Figure 4.8: Steering boat to a point on the opposite river bank - relaxed problem with no constraints

the vector field outside of the river changes sign and therefore, it actually helps the boat move towards its target. The shape of the trajectory is determined by differential equation (4.26).

Now we focus only on the problem of steering against the stream. There are two possible ways to approach it. We can incorporate state constraints into our solution of the problem or we could use nonsmooth vector field with components

$$c_x = \begin{cases} -y(y-1) & \text{if } 0 \le y \le 1, \\ 0 & \text{otherwise,} \end{cases}$$
$$c_y = 0,$$

without constraints, which is an equivalent formulation of our problem. Suppose it is not equivalent, i.e., there exist an optimal solution $(\hat{x}, \hat{y})$ crossing river bank for the problem with nonsmooth field. Let the trajectory leave the river at $t = t_0$. Without loss of generality, suppose it crosses the bank $y = 1$. Therefore, from continuity of the solution and the fact that the final destination lies on the bank of the river, it is outside of the river for a nontrivial time interval $t \in (t_0, t_1)$, where t_1 is the time when the trajectory enters the river back. We can take a different solution and replace the trajectory on (t_0, t_1) with $(\tilde{x}, \tilde{y}) := (\hat{x}, 1)$. We can see that

trajectory $(\tilde{x}, \tilde{y})$ is faster and therefore, $(\hat{x}, \hat{y})$ is not an optimal solution of the problem with nonsmooth vector field.

Pure state constraints and PMP

There are many forms of theorems for various optimal control problems with pure state constraints. A theorem providing necessary conditions of optimality in the presence of a scalar state inequality $g(x(t), t) \leq 0$, where $g : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ is presented in [KP19] for a fixed time problem. Function g is assumed to be twice continuously differentiable. A new function is introduced:

$$\Gamma(x, u, t) := \langle \nabla_x g(x, t), f(x, u, t) \rangle + \nabla_t g(x, t).$$

If we consider a controlled dynamical system in a standard form:

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

and substitute it into Γ, we can notice that Γ is the derivative of g with respect to the system's trajectory. Then we use augmented Hamiltonian:

$$\bar{H}(x, p, u, \mu, \eta) := \langle p, f(x, u, t) \rangle - \mu\Gamma(x, u, t) - \eta L(x, u, t)$$

and necessary conditions are derived from this augmented Hamiltonian. There is a new multiplier $\mu(t)$, which has interesting properties:

- $\mu(t)$ is nonincreasing;
- $\mu(t)$ is constant whenever $g(x^*(t), t) < 0$, e.g., whenever the trajectory lies in the interior of a set $\{ (x, t) \mid g(x(t), t) \leq 0 \}$.

For a rigorous formulation we refer the interested reader to [KP19]. This theorem is used in a variable time navigational problem in [Che+18]. A useful insight into Γ function construction can be found in [MG09, p. 196].

The function Γ appears also in the transversality condition (together with the multiplier μ), as it is stated in [Aru12], where we can construct a function g to describe our state constraints such that it is regular (in the sense described in the article) and the theorem provided in this article can be used in our case. This theorem also allows vector function g.

Another extremely useful theorem, which is already formulated for a variable time navigational problem - our problem specifically, is presented in [CKP20]. The function Γ and augmented Hamiltonian $\bar{H}$ are defined in the same way. We return

to our notation and denote the state as (x, y). The state contraint is defined with a usage of absolute value as $|y - 1/2| \leq 0$ (in our case, state constraints in the article were $-1 \leq x \leq 1$ and the theorem is formulated according to that). The theorem then imposes multiplier $\mu(t)$, which has the following properties:

- $\mu(t)$ is constant whenever $|y - 1/2| < 0$;
- $\mu(t)$ is increasing on the time intervals $\{ t \in \langle 0, \tau^* \rangle \mid y^*(t) = 0 \}$;
- $\mu(t)$ is decreasing on the time intervals $\{ t \in \langle 0, \tau^* \rangle \mid y^*(t) = 1 \}$.

This theorem resembles the one in [KP19], but extended to incorporate two state constraints, where both are merged to only one multiplier μ. It is possible due to the specific form of our state constraints.

It is unimportant which theorem we use, the derivative of g with respect to the trajectory is zero at the boundary points of state constraints. That means $\Gamma(x^*, u^*, t) = 0$ whenever $g(x^*, t) = 0$.

We can choose for example $g := y(y - 1)$, then we have $\Gamma = (2y - 1)V \sin\theta$. Boundary points of our state constraints are $y = 1$ and $y = 0$. After we substitute them to Γ and set it equal to zero, we get two equations:

$$0 = -V \sin\theta^* \quad \text{for } y = 0,$$
$$0 = V \sin\theta^* \quad \text{for } y = 1.$$

We are interested only in situation when the boat has to go upstream and as a consequence of that, we choose $\theta^* = \pi$. When we are in the interior of

$$\{ (x, t) \mid g(x(t), t) \leq 0 \},$$

necessary conditions are not affected by the presence of state constraints.

Application of Clarke's multiplier rule for mixed constraints

Now we can invoke Clarke's necessary conditions for mixed constraints, i.e., Theorem 3.11, despite we know that constraint qualifications are not met. We will see later on that these provide us with optimal solution regardless of the fact that constraint qualifications are not met.

We have a geometrical constraint $\varphi(t, x, u) \in \Phi$ - in our case $\varphi(t, x, u) = y$ and $\Phi = \langle 0, 1 \rangle$. Then we invoke necessary conditions provided by Theorem 3.11. The

(augmented) Hamiltonian is:

$$H = -\eta + p_1(v - y(y-1)) + p_2 w - \lambda y,$$

where a measurable function $\lambda : \langle 0, \tau^* \rangle \to \mathbb{R}$ is a multiplier. Note, that the absence of control in function φ results in no need to distinguish between the augmented Hamiltonian and the standard Hamiltonian in the necessary conditions. The costate equations are then:

$$\begin{aligned} p_1' &= 0 \\ p_2' &= p_1(2y-1) + \lambda(t), \end{aligned}$$

$p_1 = A$ is a constant and we can differentiate (4.7) to get:

$$\theta' = \cos^2\theta\left(2y - 1 + \frac{\lambda(t)}{A}\right). \tag{4.27}$$

We know that:

$$\lambda(t) \in N_\Phi^C(\varphi(t, x^*(t), u^*(t))) = \begin{cases} r & \text{if } y = 1, \\ 0 & \text{if } y \in (0,1), \\ -s & \text{if } y = 0, \end{cases}$$

for some $r, s > 0$. From the theorem we know that this holds almost everywhere, but we can continuously extend the function λ. From the previous approach we know that for $y \in \{0, 1\}$, $\theta^* = \pi$ and therefore, $\theta' = 0$. We can calculate r and s from (4.27) as $r = -A$, $s = -A$. Since we are steering upstream, we know that $A < 0$. We do not need the exact value A for our solution, we may look for a multiplier $\tilde{\lambda}(t) = \frac{\lambda(t)}{A}$. Consequently, we can see that despite the fact that constraint qualifications are not satisfied, Clarke's theorem also provides us with the same solution as the methods above, for which all assumptions are satisfied.

The trajectories shown in Figure 4.9 were found using collocation method for two boat velocities.

The yellow trajectories are for $V = 1$ and the blue trajectories are for $V = 5$. The start of these trajectories is on the bank where $y = 0$ and the end of these trajectories is on the bank where $y = 1$. We can see that state constraints are satisfied.

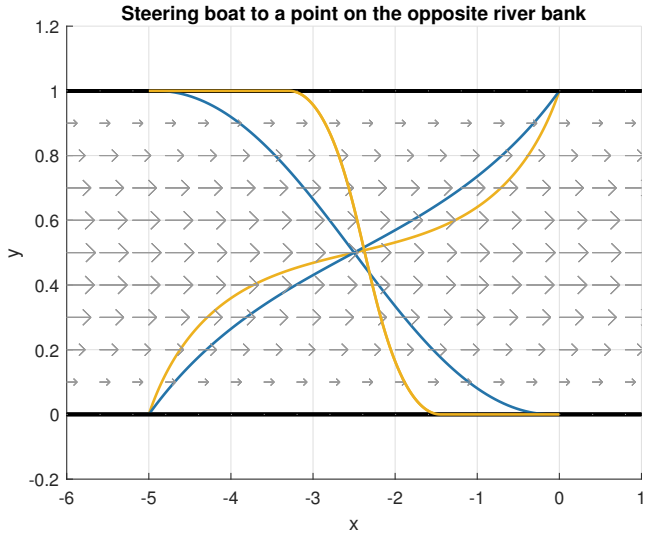


Figure 4.9: Steering boat to a point on the opposite river bank - incorporating state constraints

4.3.3 Steering to a line segment on the opposite bank

We consider a final set $E(\tau) = \{ (s, 1) \mid s \in \langle a, b \rangle \}$, which is a line segment on the bank $y = 1$. From the transversality condition we get

$$
\begin{aligned}
(h, -p(\tau^*)) \in & \nabla l(\tau^*, x(\tau^*), y(\tau^*)) + N_S^L(\tau^*, x(\tau^*), y(\tau^*)) \\
= & \begin{cases}
\{ (0, 0, r) \mid r \in \mathbb{R} \} & \text{for } x(\tau^*) \in (a, b), \\
\left\{ (0, r\cos\varphi, r\sin\varphi) \;\middle|\; \varphi \in \left\langle \frac{\pi}{2}, \frac{3\pi}{2} \right\rangle, r \geq 0 \right\} & \text{for } x(\tau^*) = a, \\
\left\{ (0, r\cos\varphi, r\sin\varphi) \;\middle|\; \varphi \in \left\langle -\frac{\pi}{2}, \frac{\pi}{2} \right\rangle, r \geq 0 \right\} & \text{for } x(\tau^*) = b.
\end{cases}
\end{aligned}
$$

We know that the optimal solution is $\theta^* = \pi/2$, whenever the final destination provided by this solution belongs to the set E, that is, whenever $x_0 + \frac{1}{6V} \in \langle a, b \rangle$ (recall the solution from section 4.3.1). If this is not the case, the optimal solution cannot be chosen $\theta^* = \pi/2$ and therefore, the boat must reach the set E in one of the points $(a, 1)$ or $(b, 1)$, because the transversality condition does not allow otherwise.

If this is the case, the solution is reduced to steering boat to a specific point on the opposite river bank, discussed in detail in section 4.3.2. We know that $x_0 + \frac{1}{6V} \notin \langle a, b \rangle$. Therefore, the line segment $\langle a, b \rangle$ lies on the left-hand side or the right-hand side from the point $x_0 + \frac{1}{6V}$. Therefore, we may choose the point a or b, which is closer to $x_0 + \frac{1}{6V}$. Alternatively, this follows from the transversality

condition as well. That is, we choose a point that belongs to the set

$$\operatorname*{argmin}_{z\in\{a,b\}}\left|z - x_0 - \frac{1}{6V}\right|.$$

Then we solve the problem of steering boat to a specific point on the opposite bank with a final value for the state $x_{fin} \in \operatorname{argmin}_{z\in\{a,b\}}\left|z - x_0 - \frac{1}{6V}\right|$.

4.4 Laminar flow around a circle

Now we take problem (4.1) with a uniform flow around a disk

$$D_R = \left\{ (x, y) \in \mathbb{R}^2 \;\middle|\; x^2 + y^2 \leq R^2 \right\}.$$

We assume an ideal liquid with no circulation. The potential is:

$$\varphi(x, y) = V_f x\left(1 + \frac{R^2}{x^2 + y^2}\right) \quad \text{for } (x, y) \in \mathbb{R}^2 \setminus D_R,$$

from which we can obtain the vector field as:

$$\begin{pmatrix} c_x \\ c_y \end{pmatrix} = \nabla\varphi = \begin{pmatrix} 1 + \frac{R^2(y^2-x^2)}{(x^2+y^2)^2} \\ -\frac{2xyV_fR^2}{(x^2+y^2)^2} \end{pmatrix}.$$

The target set is the circle $E = D_R$. The costate equations are

$$\begin{pmatrix} p_1' \\ p_2' \end{pmatrix} = \begin{pmatrix} p_1\left(\frac{6\,R^2\,V_f\,x}{(x^2+y^2)^2} - \frac{8\,R^2\,V_f\,x^3}{(x^2+y^2)^3}\right) + p_2\left(\frac{2\,R^2\,V_f\,y}{(x^2+y^2)^2} - \frac{8\,R^2\,V_f\,x^2\,y}{(x^2+y^2)^3}\right) \\ p_2\left(\frac{2\,R^2\,V_f\,x}{(x^2+y^2)^2} - \frac{8\,R^2\,V_f\,x\,y^2}{(x^2+y^2)^3}\right) + p_1\left(\frac{2\,R^2\,V_f\,y}{(x^2+y^2)^2} - \frac{8\,R^2\,V_f\,x^2\,y}{(x^2+y^2)^3}\right) \end{pmatrix}.$$

From now on we assume unit disk ($R = 1$). We can take a parametrization of ∂D_R as $x = \cos\alpha$, $y = \sin\alpha$, $\alpha \in (-\pi, \pi\rangle$. Then, from the transversality condition, we get

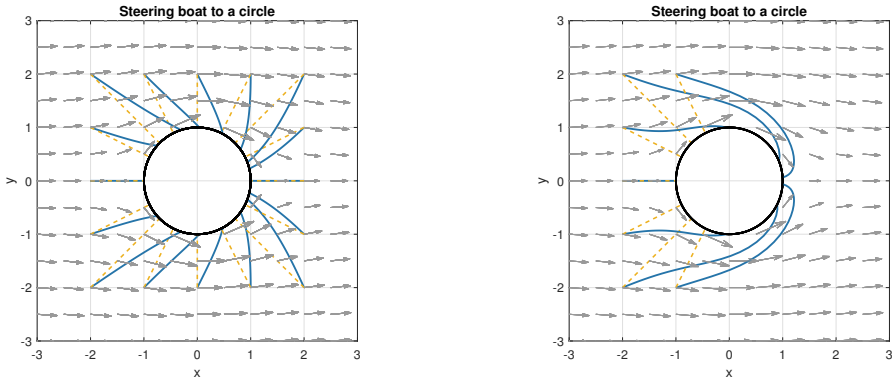
$$(h, -p_1(\tau^*), -p_2(\tau^*)) = (0, K\cos\alpha, K\sin\alpha) \quad \text{for some } K > 0.$$

For $\alpha \neq k\pi/2$, $k \in \mathbb{Z}$, we can rewrite the final condition imposed on costates as

$$\tan\alpha = \frac{p_2(\tau^*)}{p_1(\tau^*)}.$$

From the circle parametrization we also have final constraints on the states $x_{fin} = \cos\alpha$ and $y_{fin} = \sin\alpha$. We assume initial conditions $(x_0, y_0) \notin D_R$.

This problem was solved using collocation method for different initial conditions and two ratios of V and V_f. Trajectories for a slow flow $V = 2V_f$ and for a faster flow $V = V_f/2$ are shown in Figures 4.10 (a) and 4.10 (b) respectively. Blue lines represent trajectories and yellow dashed lines shows the shortest path to the disk in terms of Euclidean distance.



(a) Steering boat to a unit circle with $V = 2V_f$ **(b)** Steering boat to a unit circle with $V = V_f/2$

Figure 4.10: Simulations for steering boat to unit circle with laminar flow

5 Optimal HIV treatment

In this chapter, we encounter a problem of optimal HIV (human immunodeficiency virus) treatment proposed in [LW07, p. 124].

5.1 Insights into HIV treatment

The emergence of the first cases of HIV in 1981 marked the onset of the notorious HIV epidemic. Initial stages of HIV infection are characterized by a notable peak in viral load, as documented by [CR18]. The virus targets $CD4^+T$ cells, commonly referred to as T-helper cells, which play a pivotal role in the immune system. Originating in the bone marrow and maturing in the thymus, these cells are integral to activate human innate immune system, as elucidated by [Luc+12]. As outlined in [GG23], the virus infiltrates and replicates within these cells, resulting in a depletion of $CD4^+T$ cell counts, which weakens the body's immune system as the disease progresses. When untreated, the progression of the disease often culminates in acquired immune deficiency syndrome (AIDS), a condition invariably fatal. The concentration of $CD4^+T$ cells can be detected from blood and it is a significant indicator of the stage of the disease.

HIV can be treated with antiretroviral treatment, which can ensure a long and healthy life for an infected individual. When the medicine is taken correctly according to prescription, it is possible to supress the disease to have an undetectable load of virus, i.e., its concentration can not be detected by a standard blood test. According to [GG23], it is important that patients still undergo preventive blood tests. It can indicate when the treatment is ineffective. Individuals achieving an undetectable viral load are not HIV carriers.

Despite ongoing advancements in medicine, HIV treatment remains a lifelong commitment, which means patients need continual medication for the rest of their lives. Nevertheless, we consider a treatment lasting only 20 days, as the problem is formulated in [LW07]. This example is only illustrative. Longer treatment with two different medicines is presented in [Ada+04] or [ARK21]. However, there is another reason, why consider a short-term treatment, mentioned in [ARK21]. With ongoing treatment, the virus can develop drug resistance and in addition, the virus can mutate, which may alter the model or its parameters. Furthermore, according

to [Hug+14], individuals, who achieve low virus load are often advised by their doctors to stop or delay the treatmet to relieve the body of the side-effects of the treatment.

5.2 Optimal control problem formulation

We suppose an optimal control problem of a form:

$$\begin{aligned}
&\max \int_0^{t_{fin}} AT(t) - (1-u(t))^2 dt \\
s.t.: \quad & T'(t) = \frac{s}{1+V(t)} - m_1 T(t) + rT(t)\left[1 - \frac{T(t)+T_i(t)}{T_{max}}\right] - u(t)kV(t)T(t), \\
& T_i'(t) = u(t)kV(t)T(t) - m_2 T_i(t), \\
& V'(t) = Nm_2 T_i(t) - m_3 V(t), \\
& T(0) = T_0 > 0, \quad T_i(0) = T_{i0} > 0, \quad V(0) = V_0 > 0, \\
& 0 \le u(t) \le 1 \quad \text{for all } t \in \langle 0, t_{fin} \rangle,
\end{aligned} \tag{5.1}$$

where $A > 0$ is a constant, $V(t)$ is the concentration of free virus particles, $T(t)$ is the concentration of uninfected $CD4^+T$ cells, $T_i(t)$ is the concentration of infected $CD4^+T$ cells, r is the growth rate of T cells per day, T_{max} is assumed maximal concentration of $CD4^+T$ cells in a logistic growth model, m_1, m_2 and m_3 are the natural death rates of uninfected $CD4^+T$ cells, infected $CD4^+T$ cells and virus particles respectively, the rate that free virus particles infect $CD4^+T$ cells is proportional to constant k, N is an average number of free virus particles produced after an infection of T cell, s represents the creation of new $CD4^+T$ cells in thymus. The function $u(t)$ represents the strenght of chemotherapy, where $u = 1$ means zero treatment strength and $u = 0$ signifies the strongest treatment. The function $u \mapsto 1 - u$ seems more intuitive, because then the strongest treatment corresponds to function value 1 and no treatment corresponds to value 0. Note that the problem is autonomous. Constants corresponding to a good working model are provided in [LW07].

5.2.1 Necessary conditions

When we rewrite maximization as a minimization (by changing the sign of the Lagrangian), we can use Theorem 3.5 (the extended maximum principle). From (5.1) we have Hamiltonian in a form:

$$H = A\,T - p_3\,(m_3\,V - N\,m_2\,T_i) - (u-1)^2 - p_2\,(m_2\,T_i - k\,u\,T\,V) \tag{5.2}$$

$$- p_1 \left[m_1 T - \frac{s}{V+1} + r T \left(\frac{T+T_i}{T_{max}} - 1 \right) + k\, u\, T\, V \right],$$

which may look complicated, but it is a concave quadratic function in u, which attains global maximum at:

$$u^* = \frac{1}{2} kTV(p_2 - p_1) + 1$$

and a maximum on the set $U = \{ u \in \mathbb{R} \mid 0 \leq u \leq 1 \}$ is being attained at:

$$u^* = \begin{cases} 0 & \text{if } \frac{1}{2} kTV(p_2 - p_1) + 1 < 0, \\ 1 & \text{if } \frac{1}{2} kTV(p_2 - p_1) + 1 > 1, \\ \frac{1}{2} kTV(p_2 - p_1) + 1 & \text{otherwise.} \end{cases}$$

We can derive costate equations as:

$$\begin{aligned} p_1' &= p_1 \left[m_1 + r \left(\frac{T+T_i}{T_{max}} - 1 \right) + \frac{rT}{T_{max}} + kuV \right] - A - p_2 kuV, \\ p_2' &= m_2\, p_2 - N\, m_2\, p_3 + \frac{p_1\, r\, T}{T_{max}}, \\ p_3' &= m_3\, p_3 + p_1 \left(\frac{s}{(V+1)^2} + k\, u\, T \right) - p_2\, k\, u\, T. \end{aligned}$$

From the trasversality condition we get:

$$-p(t_{fin}) \in \nabla l(T^*(t_{fin}), T_i^*(t_{fin}), V^*(t_{fin})) + N_E^L(T^*(t_{fin}), T_i^*(t_{fin}), V^*(t_{fin})) = \{\mathbf{0}\},$$

which is essentially a final condition for costates.

5.2.2 Simulations

We can use forward-backward sweep method, described in appendix A, to solve this system of nonlinear differential equations. The optimal treatment strength 1 - u (yellow), the time development of virus particles and both infected and uninfected cells concentration (blue) are shown in Figure 5.1. Constants used for this simulation are taken from [LW07]:

$$\begin{gathered} s = 10, \quad m_1 = 0.02, \quad m_2 = 0.5, \quad m_3 = 4.4, \quad r = 0.03, \\ T_{max} = 1500, \quad k = 0.000024, \quad N = 300, \end{gathered}$$

units can be found in [BKL98], initial conditions are:

$$T_0 = 800, \quad T_{i0} = 0.4, \quad V_0 = 1.5,$$

the cost is $A = 0.05$ and the end time for optimization is $t_{fin} = 20$.

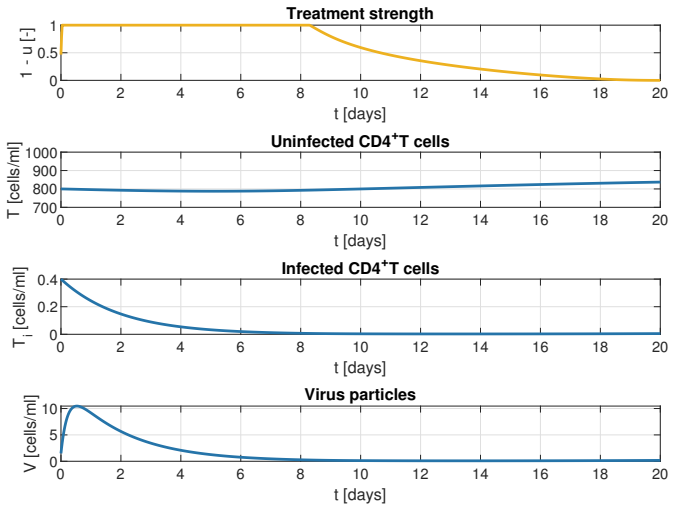


Figure 5.1: Optimal HIV treatment for $A = 0.05$

We can see that this treatment is very efficient - the concentration of uninfected $CD4^+T$ cells grows and the concentration of virus particles goes to zero. It is worth to mention that in this scenario, the presence of HIV was caught extremely early. The final cells concentrations are $T(t_{fin}) = 837.198, T_i(t_{fin}) = 0.005, V(t_{fin}) = 0.170$. In this scenario, we see that the virus is almost entirely eliminated within a short amount of time and therefore, we might alter our cost functional to put a higher weight on reducing the side effects of treatment, hence, use smaller constant $A = 0.018$. Optimal treatment strategy and time development of cells is shown in Figure 5.2. Final state values are $T(t_{fin}) = 823.317, T_i(t_{fin}) = 0.015, V(t_{fin}) = 0.479$. We can see that the virus elimination is still good and the strongest treatment is utilized for shorter time interval, which reduces its side effects.

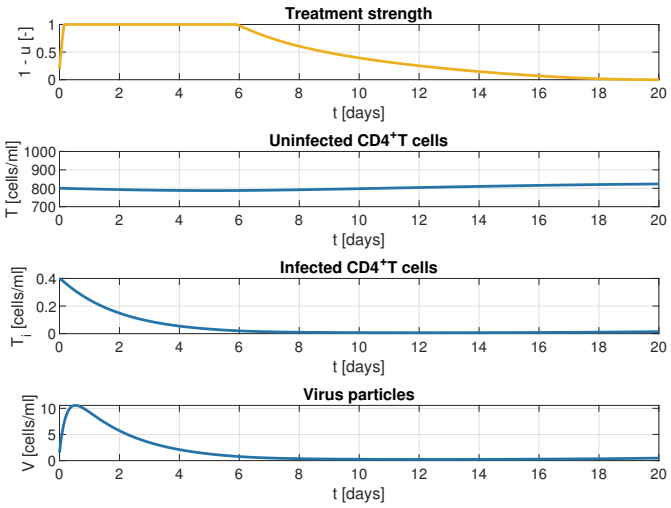


Figure 5.2: Optimal HIV treatment for $A = 0.018$

5.3 Time dependent constraints on the strength of treatment

In this section, we modify problem (5.1) to allow more complex constraints on u. Instead of requiring $0 \leq u(t) \leq 1$ on $\langle 0, t_{fin} \rangle$, we assume

$$0 \leq u_l(t) \leq u(t) \leq u_u(t) \leq 1 \quad \text{for all } t \in \langle 0, t_{fin} \rangle,$$

where $u_l : \langle 0, t_{fin} \rangle \to \mathbb{R}$ is the lower bound on treatment strength and $u_u : \langle 0, t_{fin} \rangle \to \mathbb{R}$ is the upper bound on treatment strength. Observe that the only difference in finding the optimal tretment compared to derivation in section 5.2.1 appears in the maximization of Hamiltonian (5.2). In this case, the Hamiltonian is maximized on a time varying set $M(t) := \{\tilde{u} \in \mathbb{R} | u_l(t) \leq \tilde{u} \leq u_u(t)\}$, which leads to maximization for every t on $\langle 0, t_{fin} \rangle$. This gives us optimal treatment

$$u^*(t) = \begin{cases} u_l(t) & \text{if } \tilde{u}_t < u_l(t), \\ u_u(t) & \text{if } \tilde{u}_t > u_u(t), \\ \tilde{u}_t & \text{otherwise,} \end{cases}$$

where

$$\tilde{u}_t := \frac{1}{2}kT(t)V(t)(p_2(t) - p_1(t)) + 1.$$

We can also see that the problem is not autonomous, because the set U is not independent of time. Hence, the Hamiltonian is not constant, although in the previous section, it was.

5.3.1 Simulations

We used forward-backward sweep method to solve these problems.

Delayed beginning of treatment

We set $u_u(t) = 1$ on $\langle 0, t_{fin}\rangle$ and

$$u_l(t) = \begin{cases} 1 & \text{if } t \in \langle 0, t_{fin}/4), \\ 0 & \text{if } t \in \langle t_{fin}/4, t_{fin}\rangle, \end{cases}$$

which corresponds to starting a treatment at $t = t_{fin}/4$. We used the same constants as in previous section. The simulation is shown in Figure 5.3. The constant A had to be significantly increased to $A = 5$ to achieve reasonable treatment. Delayed treatment alo results in much worse final cells concentrations $T(t_{fin}) = 776.881, T_i(t_{fin}) = 0.001, V(t_{fin}) = 0.032$, where we can see, despite increased weight A, the concentration of $CD4^+T$ cells at final time is lower than their concentration at the beginning.

Treatment stopped early

We set $u_u(t) = 1$ on $\langle 0, t_{fin}\rangle$ and

$$u_l(t) = \begin{cases} 0 & \text{if } t \in \langle 0, t_{fin}/2\rangle, \\ 1 & \text{if } t \in (t_{fin}/2, t_{fin}\rangle, \end{cases}$$

which corresponds to the situation that treatment stopped at $t = t_{fin}/2$. The simulation is shown in Figure 5.4. In this case, we can use $A = 0.05$ as before. Final values of states are $T(t_{fin}) = 830.227, T_i(t_{fin}) = 0.012, V(t_{fin}) = 0.405$. We can see a predictive behaviour when we compare this situation with basic situation in Figure 5.1. Because it was apriori known that the treatment stops at $t = t_{fin}/2$, the

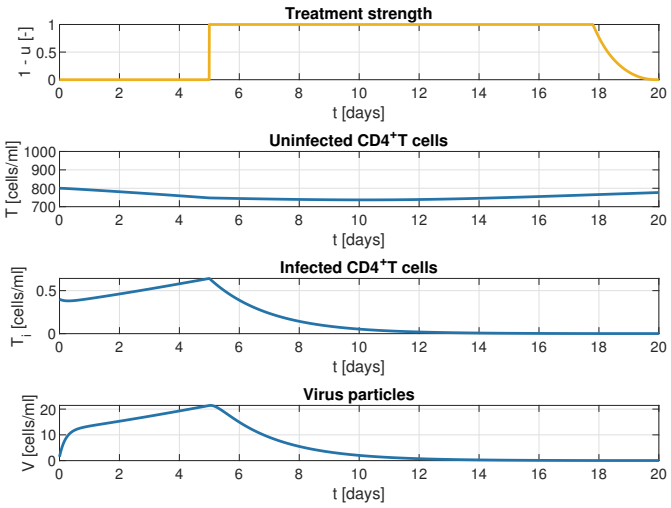


Figure 5.3: Delayed beginning of treatment

strongest treatment is being kept for a longer time interval. Also we can see that while there is no treatment, the virus concentration grows.

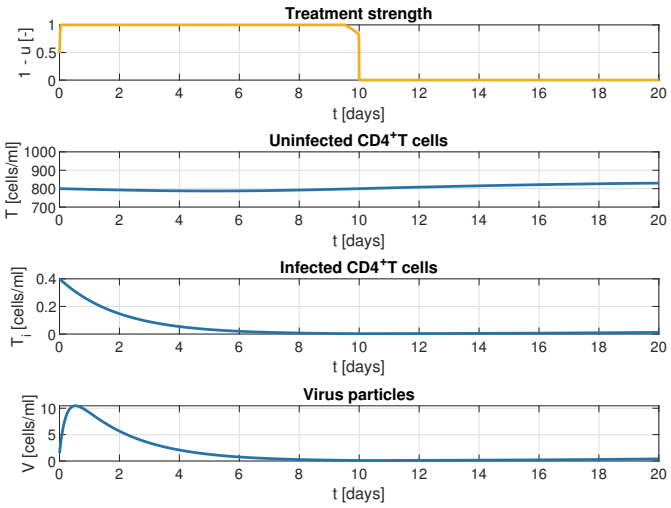


Figure 5.4: Treatment stopping early

Discontinuous treatment

Now we model a situation when the treatment is being received only for six days of a week. This situation may show what happens with the immune system, when the treatment is intermitted. Therefore, we set $u_u(t) = 1$ on $\langle 0, t_{fin} \rangle$ and

$$u_l(t) = \begin{cases} 1 & \text{if } t \in (6,7) \cup (13,14), \\ 0 & \text{otherwise.} \end{cases}$$

The simulation is shown in Figure 5.5 for $A = 0.05$. We can see that the treatment is still efficient. One might be misled by the shape of the graph and assume the difference compared to situation in Figure 5.1 is not significant, but there is difference in final state values, which in this case are $T(t_{fin}) = 831.451$, $T_i(t_{fin}) = 0.007$, $V(t_{fin}) = 0.220$. In the figure we can clearly see the impact of the first treatment turn-off.

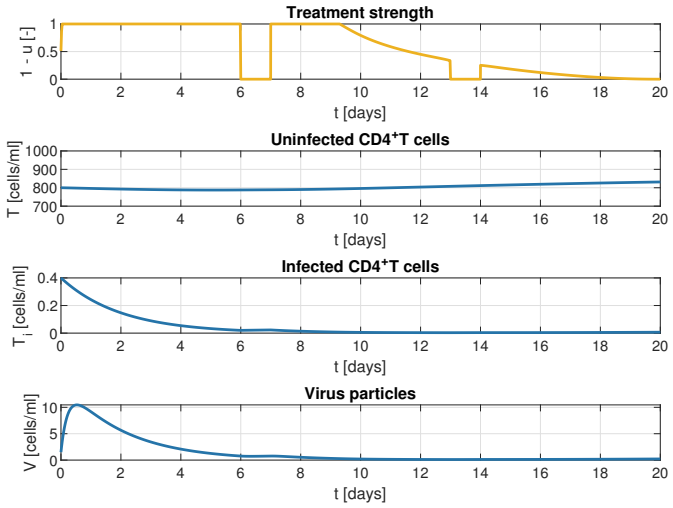


Figure 5.5: Treatment switching cycle starting on the seventh day

The fact that the first turn-off occured at seventh day from the beginning of the treatment is crucial. We can look at the situation, when the first turn-off happens on the second day of the week in the Figure 5.6. Final values of state variables are $T(t_{fin}) = 820.396$, $T_i(t_{fin}) = 0.008$, $V(t_{fin}) = 0.256$, which is apparently worse than in the previous case. One could argue that the change is caused by the fact that in this case, there is one more interval, where there is no treatment, but that is not entirely true. Clearly, the third interval worsens the situation a bit, but

what causes the most significant difference is the fact, that at early stage of the treatment (more precisely, on the second day of the treatment), the virus is given an opportunity to grow and its growth is faster, because its concentration has not been reduced enough during the first day of treatment. For completeness, if we omit the third interval, when there is no treatment, we get final values of states $T(t_{fin}) = 824.085, T_i(t_{fin}) = 0.007, V(t_{fin}) = 0.242$, which indicates that the cause of the difference between these two situations is indeed identified correctly.

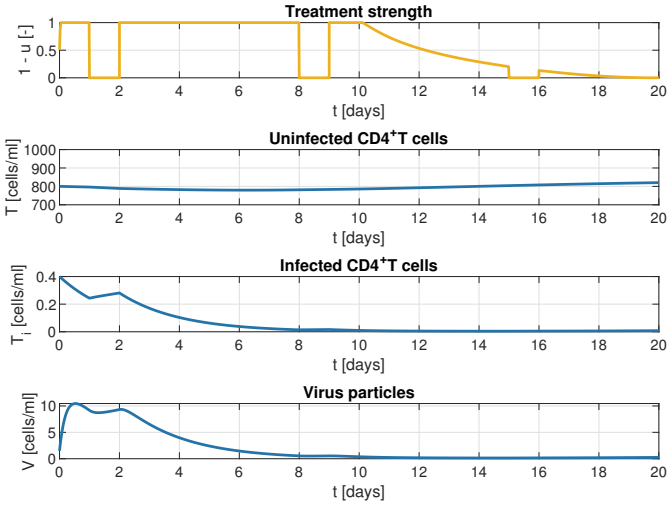


Figure 5.6: Treatment switching cycle starting on the second day

5.3.2 Discussion

In many cases, the patients must interrupt the treatment due to developed drug resistance, high toxicity of the medication or developed intercurrent conditions, which make the treatment continuation impossible, as stated in [MHH05]. This was the beginning of structured treatment interruptions, where the long-term treatment is divided into intervals of continued treatment and intervals, when the patient does not receive any treatment. However, this treatment is experimental, since it is still not proved nor disproved if it is beneficial. A study providing outcome (after 10 years) of 18 patients after several dozens of weeks (it varied from patient to patient) of structured treatment interruptions in one hospital is presented in [HPH20], which says that there were no significant differences between the $CD4^+T$ cells concentration before and after structured treatment interruptions.

According to a study [Hug+14], concerning HIV positive population of United States of America, 1 in every 20 infected individuals has experienced discontinued treatment. This study presents various reasons for intermitting the treatment and says that in 51% of all treatment discontinuations, the reason is the doctor's advice.

Simulations with discontinuous treatment can provide more information for making decision, whether treatment interruption is beneficial for the infected individual or not and it may help schedule the periods with and without treatment. Structural treatment interruptions solutions with optimization are presented in [Ada+04] or [AT17], where the control is assumed to be bang-bang and subject to the optimization are the intervals for switching the treatment. However, we had not found any allusions of usage of PMP in structured treatment interruptions.

Our approach allows us to optimize the control itself in a pre-defined interval scheme (thanks to Clarke's general formulation of PMP), which can be used in cases of enforced treatment interruption apart from the structured treatment interruption. The task could also be extended to optimize times of switches with the use of hybrid principle (we consider one mode with continuous control and second mode completely without control and the linking condition would be the continuity of states), where the optimal times of switches would be the optimal times to stop (resp. begin) the treatment. This situation is discussed in the following section, where we find optimal solution for one treatment turn-off with the usage of hybrid principle, although the principle could be used for more turn-off intervals.

5.4 Optimization of treatment turn-off time

We formulate our problem as a hybrid system with two modes. In the first mode, we assume there is a continuous control $0 \leq u \leq 1$ and in the second mode, we consider no control at all (no treatment). We label the state variables of the second mode R (concentration of healthy $CD4^+T$ cells), R_i (concentration of infected $CD4^+T$ cells) and W (virus particles). We formulate a hybrid optimal control problem as follows

$$\max \int_0^\tau AT(t) - (1-u(t))^2 dt + \int_\tau^{t_{fin}} AR(t)dt - C\tau$$

s. t.:

$$1^{\text{st}} \text{ mode} \begin{cases} T'(t) = \frac{s}{1+V(t)} - m_1 T(t) + rT(t)\left[1 - \frac{T(t)+T_i(t)}{T_{max}}\right] - u(t)kV(t)T(t), \\ T_i'(t) = u(t)kV(t)T(t) - m_2 T_i(t), \\ V'(t) = Nm_2 T_i(t) - m_3 V(t), \\ \text{for all } t \in \langle 0, \tau \rangle, \end{cases}$$

$$
2^{\text{nd}}\text{ mode}\begin{cases} R'(t) = \frac{s}{1+W(t)} - m_1 R(t) + rR(t)\left[1 - \frac{R(t)+R_i(t)}{T_{max}}\right] - kW(t)R(t), \\ R_i'(t) = kW(t)R(t) - m_2 R_i(t), \\ W'(t) = Nm_2 R_i(t) - m_3 W(t), \\ \text{for all } t \in \langle \tau, T_{fin} \rangle, \end{cases}
$$

$$
T(0) = T_0 > 0, \quad T_i(0) = T_{i0} > 0, \quad V(0) = V_0 > 0,
$$

$$
0 \le u(t) \le 1 \quad \text{for all } t \in \langle 0, \tau \rangle,
$$

$$
(T(\tau), T_i(\tau), V(\tau)) = (R(\tau), R_i(\tau), W(\tau)),
$$

where the meaning of symbols is the same as stated at the beginning of this chapter. The linking condition in this optimal control problem is the continuity of state variables in both modes.

Notice that our running costs for both modes remain the same as in previous sections, but we have added the term $-C\tau$ to the cost functional. Note that solution with control turn-off is admissible for the problem solved in the previous sections and hence, control turn-off would not occur in the optimal solution to this hybrid problem, had we not modified the cost functional. We wanted to reward treatment turn-off at time τ and hence, we added the term $-C\tau$.

We use Theorem 3.8 (hybrid principle). From the transversality condition, we get that

$$
p(0) \in \mathbb{R}^3, \quad q(t_{fin}) \in \{\mathbf{0}\},
$$

where $p = (p_1, p_2, p_3)$ are costates for the first mode and $q = (q_1, q_2, q_3)$ are costates for the second mode. For the first mode, we get the same adjoint equations as in previous sections and for the second mode we get the same adjoint equations as in previous sections expressed for $u = 1$. Hamiltonian H_1 for the first mode is the same Hamiltonian as in previous sections and Hamiltonian H_2 for the second mode is H_1 expressed for $u = 1$.

Let τ^* be the optimal switching time. We label the state values at τ^* as $\mathcal{T}^* := (T(\tau^*), T_i(\tau^*), V(\tau^*))$ and $\mathcal{R}^* := (R(\tau^*), R_i(\tau^*), W(\tau^*))$. The switching condition gives us

$$
(h_1 - h_2, -p(\tau^*), q(\tau^*)) \in \partial_L l(\tau^*, \mathcal{T}^*, \mathcal{R}^*) + N_S^L(\tau^*, \mathcal{T}^*, \mathcal{R}^*) = \left\{ (C, -z, z) \,\middle|\, z \in \mathbb{R}^3 \right\},
$$

where $S = \{ (\tau, x, y) \mid \tau \in \mathbb{R}, x, y \in \mathbb{R}^n, x = y \}$ is the set used in linking condition (which, in our case, is the continuity of states). The switching condition implies that $h_1 - h_2 = C$ and that costates are continuous, i.e., $p(\tau^*) = q(\tau^*)$ (observe that in general, the switching condition ensures that costates are continuous whenever

states are continuous). We have

$$C = h_1 - h_2 = H_1(\tau^*) - H_2(\tau^*) \tag{5.3}$$
$$= -(u(\tau^*) - 1)^2 + (u(\tau^*) - 1)kT(\tau^*)V(\tau^*)(p_2(\tau^*) - p_1(\tau^*)),$$

because from the continuity of both states and costates, all the terms that do not contain u in the Hamiltonians are canceled during the subtraction. When the condition (5.3) is satisfied, the system switches from the first mode to the second mode and the treatment is stopped.

The problem was solved using forward-backward sweep method and the result is shown in the Figure 5.7. States for the second mode are plotted as an extension of the states of the first mode and labeled T, T_i and V, as in the first mode. This is logical, because the system itself is not hybrid, we just chose to model the situation as a hybrid problem for convenience.

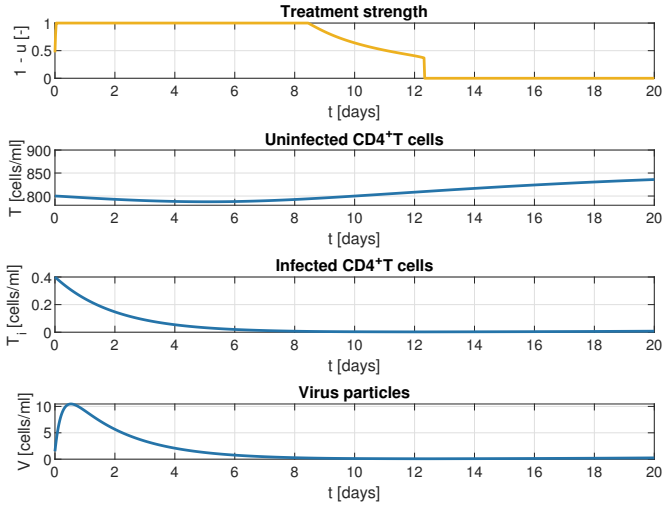


Figure 5.7: Optimal HIV treatment turn-off for $A = 0.05$, $C = 0.6$

The optimal switching time was found as $\tau^* = 12.327$ and the final state values are $T(t_{fin}) = 835.878$, $T_i(t_{fin}) = 0.008$, $V(t_{fin}) = 0.249$, which is slightly worse that what we achieved without switching the treatment off completely in section 5.2.2, as expected. Nevertheless, the treatment turn-off might be beneficial for reasons discussed above.

6 Conclusions & Outlook

6.1 Achievements

In this thesis, nonsmooth analysis was introduced in order to formulate the extended maximum principle. Various specific instances of this extended maximum principle were presented and then used in applications, which include steering boat in a vector field to a target set and optimal HIV treatment. Therefore, I consider the assignment of the thesis fulfilled in its full extent. As some of the greatest contributions of this thesis I consider:

- optimal trajectories for steering boat to the moving ellipse;
- found optimal HIV treatment in pre-defined intervals;
- found optimal treatment turn-off and suggested approach to STI HIV treatment optimization.

6.2 Future work

Some of the straightforward improvements in terms of the mathematical theory are the exploration of sufficient conditions of optimality and analysis of the existence of the solution in the optimal HIV treatment (although the latter was done in many articles for our system). Further development in the part of applied mathematics could include more types of vector fields in the problem of boat steering, deeper exploration and understanding of theorems for pure state constraints or more complicated target set, which could involve even randomness in the position of the target set. Improvements in the optimal HIV treatment could cover optimal STI treatment with more intervals of treatment turn-off. A more complicated model of HIV could be used, which could incorporate more types of treatment. Future work could also be done in the field of numerical mathematics, which was discussed in this thesis only briefly. Advanced numerical methods could be used to solve the STI HIV treatment for more treatment turn-offs covering a longer time interval.

Bibliography

[Ada+04] B. M. Adams, H. T. Banks, Hee-Dae Kwon, and Hien T. Tran. **Dynamic Multidrug Therapies for HIV. Optimal and STI Control Approaches**. *Mathematical Biosciences and Engineering* 1:2 (2004), 223–241. ISSN: 1551-0018. DOI: 10.3934/mbe.2004.1.223. URL: http://www.aimspress.com/article/10.3934/mbe.2004.1.223 (see pages 71, 80).

[AEP20] N. Andreasson, A. Evgrafov, and M. Patriksson. **An Introduction to Continuous Optimization. Foundations and Fundamental Algorithms**. Third, Revised Third edition. Dover Publications, 2020. ISBN: 978-0486802879 (see page 22).

[And+18] Joel A E Andersson, Joris Gillis, Greg Horn, James B Rawlings, and Moritz Diehl. **CasADi – A software framework for nonlinear optimization and optimal control**. *Mathematical Programming Computation* (2018) (see page 93).

[ARK21] Shohel Ahmed, Sumaiya Rahman, and Md Kamrujjaman. **Optimal treatment strategies to control acute HIV infection**. *Infectious Disease Modelling* 6 (2021), 1202–1219. ISSN: 24680427. DOI: 10.1016/j.idm.2021.09.004. URL: https://linkinghub.elsevier.com/retrieve/pii/S2468042721000658 (see page 71).

[Aru12] A. V. Arutyunov. **Properties of the Lagrange multipliers in the Pontryagin maximum principle for optimal control problems with state constraints**. *Differential Equations* 48:12 (2012), 1586–1595. ISSN: 0012-2661. DOI: 10.1134/S0012266112120051. URL: http://link.springer.com/10.1134/S0012266112120051 (see page 65).

[AT17] Adam Attarian and Hien Tran. **An Optimal Control Approach to Structured Treatment Interruptions for HIV Patients. A Personalized Medicine Perspective**. *Applied Mathematics* 08:07 (2017), 934–955. ISSN: 2152-7385. DOI: 10.4236/am.2017.87074. URL: http://www.scirp.org/journal/doi.aspx?DOI=10.4236/am.2017.87074 (see page 80).

[Axl20] Sheldon Axler. **Measure, Integration & Real Analysis**. Cham: Springer International Publishing, 2020. ISBN: 978-3-030-33142-9 (see page 28).

[BKL98] S Butler, D Kirschner, and S Lenhart. **Optimal control of chemotherapy affecting the infectivity of HIV**. In: *Advances in Mathematical Population Dynamics — Molecules, Cells and Man*. WORLD SCIENTIFIC, 1998-02-28, 557–69. ISBN: 978-981-02-3176-7. DOI: 10.1142/9789814529594. URL: https://www.worldscientific.com/doi/abs/10.1142/9789814529594 (see page 74).

[BP07] Alberto Bressan and Benedetto Piccoli. **Introduction to the mathematical theory of control**. Springfield: American Institute of Mathematical Sciences, 2007. ISBN: 978-1-60133-002-4 (see page 52).

[BPV16] A. Boccia, M. D. R. de Pinho, and R. B. Vinter. **Optimal Control Problems with Mixed and Pure State Constraints**. *SIAM Journal on Control and Optimization* 54:6 (2016), 3061–3083. ISSN: 0363-0129. DOI: 10.1137/15M1041845. URL: http://epubs.siam.org/doi/10.1137/15M1041845 (see page 63).

[Bry96] A.E. Bryson. **Optimal control-1950 to 1985**. *IEEE Control Systems Magazine* 16:3 (1996), 26–33. DOI: 10.1109/37.506395 (see page 1).

[Che+18] Roman Chertovskih, Dmitry Karamzin, Nathalie T. Khalil, and Fernando Lobo Pereira. **An indirect numerical method for a time-optimal state-constrained control problem in a steady two-dimensional fluid flow**. In: *2018 IEEE/OES Autonomous Underwater Vehicle Workshop (AUV)*. IEEE, 2018, 1–6. ISBN: 978-1-7281-0253-5. DOI: 10.1109/AUV.2018.8729750. URL: https://ieeexplore.ieee.org/document/8729750/ (see page 65).

[CKP20] Roman Chertovskih, Nathalie T. Khalil, and Fernando Lobo Pereira. "Time-Optimal Control Problem with State Constraints in a Time-Periodic Flow Field." In: *Optimization and Applications*. Cham: Springer International Publishing, 2020, 340–354. ISBN: 978-3-030-38602-3. DOI: 10.1007/978-3-030-38603-0_25. URL: http://link.springer.com/10.1007/978-3-030-38603-0_25 (see page 65).

[Cla01] Francis Clarke. **Nonsmooth Analysis in Control Theory. A Survey**. *European Journal of Control* 7:2-3 (2001), 145–159. ISSN: 09473580. DOI: 10.3166/ejc.7.145-159. URL: https://linkinghub.elsevier.com/retrieve/pii/S0947358001711454 (see page 2).

[Cla09] Francis Clarke. "Nonsmooth Analysis in Systems and Control Theory." In: *Encyclopedia of Complexity and Systems Science*. New York, NY: Springer New York, 2009, 6271–6285. ISBN: 978-0-387-75888-6. DOI: 10.1007/978-0-387-30440-3_370. URL: https://link.springer.com/10.1007/978-0-387-30440-3_370 (see page 12).

[Cla13] Frank H. Clarke. **Functional analysis, calculus of variations and optimal control**. London: Springer-Verlag, 2013. ISBN: 978-1-4471-4819-7 (see pages 1, 5, 11, 16, 17, 25, 28–30, 32, 35–37, 41, 47, 63).

[Cla90] Frank H. Clarke. **Optimization and Nonsmooth Analysis**. III. Society for Industrial and Applied Mathematics, 1990. ISBN: 978-0-89871-256-8 (see pages 25, 30).

[Cla+98] F.H. Clarke, Yu.S. Ledyaev, R.J. Stern, and P.R. Wolenski. **Nonsmooth Analysis and Control Theory**. New York: Springer-Verlag, 1998. ISBN: 0-387-98336-8 (see pages 5, 12, 30).

[CMN19] Ştefan Cobzaş, Radu Miculescu, and Adriana Nicolae. **Lipschitz Functions**. Cham: Springer International Publishing, 2019. ISBN: 978-3-030-16488-1. DOI: 10.1007/978-3-030-16489-8 (see page 11).

[CP10] Francis Clarke and M. R. de Pinho. **Optimal Control Problems with Mixed Constraints**. *SIAM Journal on Control and Optimization* 48:7 (2010), 4500–4524. ISSN: 0363-0129. DOI: 10.1137/090757642. URL: http://epubs.siam.org/doi/10.1137/090757642 (see pages 41, 43).

[CR18] Jessica M. Conway and Ruy M. Ribeiro. **Modeling the immune response to HIV infection**. *Current Opinion in Systems Biology* 12 (2018), 61–69. ISSN: 24523100. DOI: 10.1016/j.coisb.2018.10.006. URL: https://linkinghub.elsevier.com/retrieve/pii/S2452310018300428 (see page 71).

[Gam19] R. V. Gamkrelidze. **History of the Discovery of the Pontryagin Maximum Principle**. English. *Proceedings of the Steklov Institute of Mathematics* 304:1 (2019), 1–7 (see page 1).

[GG23] Sofia A. Battistini Garcia and Nilmarie Guzman. **Acquired Immune Deficiency Syndrome CD4+ Count**. *StatPearls* (2023). URL: https://www.ncbi.nlm.nih.gov/books/NBK513289/ (see page 71).

[HPH20] Wei-Ting Hsu, Sung-Ching Pan, and Szu-Min Hsieh. **10-year outcome of temporary structured treatment interruption (STI) among HIV-1-infected patients. An observational study in a single medical center**. *Journal of the Formosan Medical Association* 119:1 (2020), 455–461. ISSN: 09296646. DOI: 10.1016/j.jfma.2019.07.029. URL: https://linkinghub.elsevier.com/retrieve/pii/S0929664619304383 (see page 79).

[Hug+14] Alison J. Hughes, Christine L. Mattson, Susan Scheer, Linda Beer, and Jacek Skarbinski. **Discontinuation of Antiretroviral Therapy Among Adults Receiving HIV Care in the United States**. *JAIDS Journal of Acquired Immune Deficiency Syndromes* 66:1 (2014), 80–89. ISSN: 1525-4135. DOI: 10.1097/QAI.0000000000000084. URL: https://journals.lww.com/00126334-201405010-00011 (see pages 72, 80).

[HW15] Ernst Hairer and Gerhard Wanner. "Runge–Kutta Methods, Explicit, Implicit." In: *Encyclopedia of Applied and Computational Mathematics*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, 1282–1285. ISBN: 978-3-540-70528-4. DOI: 10.1007/978-3-540-70529-1_144. URL: https://link.springer.com/10.1007/978-3-540-70529-1_144 (see pages 91, 92).

[KP19] Dmitry Karamzin and Fernando Lobo Pereira. **On a Few Questions Regarding the Study of State-Constrained Problems in Optimal Control**. *Journal of Optimization Theory and Applications* 180:1 (2019), 235–255. ISSN: 0022-3239. DOI: 10.1007/s10957-018-1394-2. URL: http://link.springer.com/10.1007/s10957-018-1394-2 (see pages 65, 66).

[Lib12] Daniel Liberzon. **Calculus of Variations and Optimal Control Theory: A Concise Introduction**. Princeton University Press, 2012. ISBN: 978-0-691-15187-8 (see page 38).

[LR14] Hartmut Logemann and Eugene P. Ryan. **Ordinary Differential Equations**. London: Springer London, 2014. ISBN: 978-1-4471-6397-8 (see page 26).

[Luc+12] Rishi Vishal Luckheeram, Rui Zhou, Asha Devi Verma, and Bing Xia. **CD4 + T Cells. Differentiation and Functions**. *Clinical and Developmental Immunology* 2012 (2012), 1–12. ISSN: 1740-2522. DOI: 10.1155/2012/925135. URL: http://www.hindawi.com/journals/jir/2012/925135/ (see page 71).

[LW07] Suzanne Lenhart and John T. Workman. **Optimal Control Applied to Biological Models**. Chapman and Hall/CRC, 2007. ISBN: 1-58488-640-4 (see pages 71–73, 90, 91).

[MG09] Mike Mesterton-Gibbons. **A Primer on the Calculus of Variations and Optimal Control Theory**. I. Title. American Mathematical Society, 2009. ISBN: 978-0-8218-4772-5 (see pages 52, 65).

[MHH05] J. Montaner, M. Harris, and R. Hogg. **Structured Treatment Interruptions. A Risky Business**. *Clinical Infectious Diseases* 40:4 (2005-02-15), 601–603. ISSN: 1058-4838. DOI: 10.1086/427707. URL: https://academic.oup.com/cid/article-lookup/doi/10.1086/427707 (see page 79).

[Pon+87] L.S. Pontryagin, V.G. Boltyanskii, R.V. Gamkrelidze, and E.V. Mishchenko. **Mathematical Theory of Optimal Processes**. Classics of Soviet Mathematics. New York: Taylor & Francis, 1987. ISBN: 978-2881240775 (see page 1).

[PP09] Hans Pesch and Michael Plail. **The maximum Principle of optimal control: A history of ingenious ideas and missed opportunities**. eng. *Control and Cybernetics* 38:4A (2009), 973–995. URL: http://eudml.org/doc/209682 (see page 1).

[Rao10] Anil Rao. **A Survey of Numerical Methods for Optimal Control**. *Advances in the Astronautical Sciences* 135 (Jan. 2010) (see pages 92, 93).

[SBS21] Jesse A. Sharp, Kevin Burrage, and Matthew J. Simpson. **Implementation and acceleration of optimal control for systems biology**. *Journal of The Royal Society Interface* 18:181 (2021). ISSN: 1742-5662. DOI: 10.1098/rsif.2021.0241. URL: https://royalsocietypublishing.org/doi/10.1098/rsif.2021.0241 (see page 90).

[SW97] H.J. Sussmann and J.C. Willems. **300 years of optimal control. from the brachystochrone to the maximum principle**. *IEEE Control Systems* 17:3 (1997), 32–44. ISSN: 1066-033X. DOI: 10.1109/37.588098. URL: https://ieeexplore.ieee.org/document/588098/ (see page 1).

A Forward-Backward Sweep Method

Forward-backward sweep method, implemented according to in [LW07] and [SBS21], is a numerical method used to solve fixed-time (we assume time interval $\langle a, b \rangle$) and free endpoint problems. In these problems, the maximum principle provides us with a final condition for costates $p(b) = p_{fin}$ and we assume the initial condition $x(a) = x_0$. We discretize the system at N points with a sampling period h. It is necessary we provide the algorithm with an initial guess of the optimal control u_{guess}, again discretized at N points with sampling period h. It is crucial that $u_{guess}[k] \in U(a + k \cdot h)$, that is, u_{guess} is an admissible control for the optimal control problem. Our main variables x, p and u are matrices of the states, costates and controls (respectively) in discrete times with dimensions $n \times N$, $n \times N$ and $m \times N$ respectively, where n is the order of the system and m is the number of controls. Forward-backward sweep method is explained in Algorithm 1.

Algorithm 1: Forward-Backward Sweep Method

Data: $u_{guess}, x_0, p_{fin}, N, h, \omega, a$
Result: x, p, u

$u \leftarrow u_{guess}$;
$x[0] \leftarrow x_0$;
$p[N-1] \leftarrow p_{fin}$;
while *not converged* **do**
    $u_{old} \leftarrow u$;
    $x_{old} \leftarrow x$;
    $p_{old} \leftarrow p$;
    $x \leftarrow$ F_RK4$(x_{old}[0], u_{old}, h)$;
    $p \leftarrow$ B_RK4$(p_{old}[N-1], x, u_{old}, h)$;
    $u[kh] \leftarrow g(a + kh, x[kh], p[kh])$ for all $k = 0, ..., N-1$;
    $u \leftarrow (1 - \omega)u + \omega u_{old}$;
    test convergence of u, p, x;

The algorithm itself consists of three steps, which are repeated until convergence. These steps are: forward integration (denoted F_RK4 in the algorithm), backward integration (denoted B_RK4 in the algorithm) and control update. Firstly, we

approximate the solution of the state equation by forward run of Runge-Kutta method (in our implementation, we use the classical (explicit) Runge-Kutta method of the fourth order, descibed in [HW15]). It is necessary to choose sufficiently small sampling period h. Then, we approximate the solution of the costate equation by backwards run of Runge-Kutta method (we use state values from the forward run). Afterwards, we have approximate solutions for both states and costates, which we use to find the optimal control u. Reacll that we have the expression for optimal control u from the maximum principle in the form $u(t) = g(t, x(t), p(t))$ for some function $g : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m$. Then, we update u as $u \leftarrow (1-\omega)u + \omega u_{old}$, where $\omega \in (0, 1)$ is a parameter (we used $\omega = 0.4$), u is the newly found optimal control and u_{old} is the control from previous step. Lastly, we check for convergence of the algorithm by a convergence test presented in [LW07, p. 51] imposed on all vectors (or matrices, where we impose the convergence test on each row of the matrix) x, p, u. We use relative error to determine, if the current solution, namely, vectors (or matrices) x, p, u, is sufficiently close to the previous one. When the criterion

$$\frac{\|z - z_{old}\|_1}{\|z\|_1} < \delta,$$

where z is a vector for which we test the convergence, is satisfied for some small $\delta > 0$, we stop the algorithm. The criterion can be rewritten to a form

$$\delta\|z\|_1 - \|z - z_{old}\|_1 > 0,$$

which is used in the algorithm. We use the 1-norm of a vector and $\delta = 10^{-4}$.

B Collocation Method

Collocation method, explained in [Rao10], is a standard numerical method for solving optimal control problems. We discretize the system at N points with a sampling period h, which divides the time interval $\langle a, b\rangle$ into $N-1$ intervals $\langle t_i, t_{i+1}\rangle$. The essence of this method lies in approximating the state equation by polynomials of order l on each interval. We require these polynomials match the value of the function at the beginning of each interval, i.e., $\tilde{x}(t_i) = x(t_i)$, where $\tilde{x}$ is the polynomial and x is the state. This also enforces the initial condition. Then, to enforce the dynamics of the approximated solution is an approximation of the system's dynamics, we divide each subinterval $\langle t_i, t_{i+1}\rangle$ into l points $\tilde{t}_j$ ($j = 1, ..., l$) and ensure the derivative of the polynomial at each point $\tilde{t}_j$ corresponds to the state equation (for every interval $\langle t_i, t_{i+1}\rangle$), i.e.,

$$\tilde{x}'(\tilde{t}_j) = f(\tilde{t}_j, x(\tilde{t}_j), u(\tilde{t}_j)) \quad \text{for } j = 1, ..., l. \tag{B.1}$$

Runge-Kutta methods belong to the category of collocation methods. Therefore, we can use them to force the validity of equation (B.1). We used a Runge-Kutta method of the fourth order ($l = 4$), which is explained in [Rao10] and [HW15]. We assume an autonomous system. Then, we can formulate a nonlinear program as follows

$$\begin{gathered}\min J(x, u) \\ \text{s.t.: } x[i+1] - x[i] - h(k_1/6 + k_2/3 + k_3/3 + k_4/6) = \mathbf{0} \quad \text{for } i = 1, ..., N-1\end{gathered}$$

where

$$\begin{aligned} k_1 &= f(x[i], u[i]), \\ k_2 &= f(x[i] + hk_1/2, u[i]), \\ k_3 &= f(x[i] + hk_2/2, u[i]), \\ k_4 &= f(x[i] + hk_3, u[i+1]) \end{aligned}$$

are different constants for each time interval.

Then, we add additional constraints specific to our problem, e.g., initial condition,

final condition, state constraints, control constraints. If we encounter a variable-time problem, we are forced to discretize the system at N points with an unknown sampling period h and hence, we do not have the sampling period under control and we have to pay extra attention to adjusting N. We used CasADi [And+18] tool for solving the nonlinear program with an *IPOPT* solver, which, according to [Rao10], uses second derivatives and therefore, provides us with faster convergence near the optimal solution.

Index