Optimal control of linear complementarity systems

Alexandre Vieira

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Automatique-Productique**

Arrêté ministériel : 25 mai 2016

Présentée par

Alexandre Vieira

Thèse dirigée par **Bernard Brogliato**
et codirigée par **Christophe Prieur**

préparée au sein **de l'Inria Grenoble Rhône-Alpes**
dans **l'École Doctorale EEATS**

Commande optimale des systèmes de complémentarité linéaires

Thèse soutenue publiquement le **25 septembre 2018**,
devant le jury composé de :

Éric Blayo
Professeur des Universités, Univ. Grenoble Alpes, Président
Pierre Riedinger
Professeur des Universités, Univ. de Lorraine, Rapporteur
Emmanuel Trélat
Professeur des Universités, Sorbonne Université, CNRS, Rapporteur
Kanat Camlibel
Associate Professor, University of Groningen, Examinateur
Bernard Brogliato
Directeur de Recherche Inria, Univ. Grenoble Alpes, Directeur de thèse
Christophe Prieur
Directeur de Recherche CNRS, Directeur de thèse

Thanks/Remerciements

I have to admit that when I started my PhD, I had already thought about this thanks part (it was a goal for me, as much as the manuscript itself...). It looks like marble to me: the names written here will last as long as there is a copy of this manuscript. So yes: it is going to be a bit long. I will write first in English, and then in Français, so that no one feels excluded. Also, sorry for the repetition of the words *thank* and *merci*: the exercice imposes it.

My first thoughts go of course to Bernard and Christophe, my two advisors. I must say: thank you, from the bottom of my heart, for these three years. Thank you for your presence, for all the meetings we did, and for your persistence, even in the moments when I was making you confused (by the way: sorry for that, I hope I made that up!). Thank you also for granting me freedom in the directions I decided to follow for this research. I admit that some moments were frustrating for me, and I may have been a little stubborn, but I am happy to see that, eventually, we did a job I am feeling proud of.

Talking about being proud, I would like to express my gratitude to the members of my jury, Messrs. Blayo, Camlibel, Riedinger and Trélat. It makes me proud, indeed, to know that my work is being reviewed by such a talented jury. Thank you for the time spent on this thesis, and for adding a taste of credibility to my work.

These three years would have been tasteless without all the people I have met here in Grenoble. I am thankful to all the researchers and permanent engineers of *feu* BIPOP, Florence (I am starting with you to be gallant!), Arnaud, Franck, Guillaume, Jerôme, Pierre-Brice, Vincent; thanks for the discussions we had, the interesting remarks or problems you gave me, which made me come out of my comfort zone to reach greater knowledge. A special thank you for my dear Diane: you have always been the best! I thank all the non-permanent people, all the post-docs, the engineers, the PhD students, the interns that I have met: Camille, Diana, Mounia, Thoï, Alejandro, Dimitar, Éric, Felix, Gilles, Janusz, Hadi, Kirill, Mickaël, Matteo, Nahuel, Narendra, Nestor, Nicolas. I could have written a thousand times *Janusz*, but that would not be fair. Thank you all for the tea times, the cakes (I know, I should bring more of them), the games, the laughs, the restaurants, the movies, the (sometimes serious or technical) discussions... Thank you *mates*, for sharing a slice of joyful life with me. Also, I would blame myself for forgetting all the people I met on the way in the different meetings and conferences I attended. Thank you for all the useful feedback you gave me. A big thank you to Emilio: this week discussing about so many things in Padova was more inspiring than you may think!

J'ai une pensée également pour mes collègues de la prépa de l'INPG, où j'ai eu l'honneur d'assurer deux ans de cours. Merci plus particulièrement à Hélène, Nathalie et Sophie pour m'avoir accompagné dans cette riche aventure, pour leur écoute et la réception toujours attentive des mes remarques et de mes idées, pour m'avoir permis aussi d'expérimenter librement mes tentatives pédagogiques (ces deux semaines de CM restent gravées dans ma mémoire !). Un merci aussi à tous les élèves que j'ai pu avoir, qui sans le savoir ont été un bol d'air frais dans la morosité que

peut connaître un doctorant qui cherche sans trouver. J'espère leur avoir donné l'envie (à certains) de s'intéresser un peu à la recherche, et leur avoir fait comprendre que les chercheurs n'étaient pas tous de vieux messieurs blafards prenant la poussière dans leur bureau.

Je ne serais jamais arrivé à Grenoble (et à la fin de cette thèse) sans toutes ces personnes qui ont jalonné mon parcours, et il me semble opportun de marquer dans ce marbre numérique l'affection que je leur porte. Je ne serai pas là sans les amis de l'INSA de Rouen, et surtout sans Conrad qui aura émulé ma curiosité pendant tous nos projets. Nos longues heures face aux tableaux blancs resteront des moments de bonheur précieux. Vivement nos prochaines randos vieux frère, que je te parle de toutes les beautés mathématiques que j'ai pu voir ! Un grand merci à Florian-*sensei*, qui m'aura finalement fait découvrir le vrai monde de la recherche (et comme tu peux le voir, ça a plutôt bien marché !).

Puis je me dois aussi de remercier les *Kfiens*, ce groupe d'amis qui restera éternellement ma seconde famille. Du fond du cœur, merci. Merci pour votre présence, votre écoute, votre amitié, depuis tant d'années. Promis, un jour j'arrêterai de vous parler de maths (mais pas tout de suite, j'ai une thèse à valider...).

Merci aussi à ma famille, pour son soutien inconditionnel et sa confiance (aveugle quand même !) sur l'avancée de ma carrière.

Enfin, un merci tout particulier à Xavier. Je suis sûr que tu n'as pas conscience de tout ce que je te dois dans cette thèse. Et pourtant, sans toi, sans ton affection, je ne serai pas là à écrire cette dernière ligne de façon si sereine. Sur ces trois années de recherche, tu restes ma plus belle découverte.

Contents

Glossary 2

Introduction 4

List of publications 7

I State of the art 8

1 Linear Complementarity Systems 9
- 1.1 Linear Complementarity Problems . . . 10
- 1.2 Properties of the LCS . . . 13
- 1.3 Numerical simulation of LCS . . . 21

2 Non-smooth Optimal Control Problems 24
- 2.1 Optimal control of non-smooth systems . . . 25
- 2.2 Numerical approximations . . . 30

3 Mathematical Programming with Complementarity Constraints 35
- 3.1 Finite dimension . . . 36
- 3.2 Optimal control . . . 42

II Quadratic optimal control of LCS: the 1D complementarity case 48

4 First order conditions using Clarke's subdifferential 49
- 4.1 Derivation of the maximum principle . . . 50
- 4.2 Analytical solution for a 1D example . . . 54

5 Numerical simulations 58
- 5.1 Numerical schemes . . . 58
- 5.2 Numerical results . . . 60

III Quadratic optimal control of LCS: the general case 63

6 Derivation of the first order conditions 64
- 6.1 First-order necessary conditions for the optimal control problem (6.1)(6.2) . . . 65
- 6.2 Sufficiency of the W-stationarity . . . 74

7 Numerical implementations **77**
7.1 Direct method 77
7.2 Combining direct and indirect methods: the hybrid approach 94

IV Optimality conditions for the minimal time problem **100**

8 Extension of the nonlinear first order conditions **101**
8.1 Necessary conditions 102
8.2 Application to LCS 105

Conclusion **116**

Appendix **119**

A Non-smooth Analysis **119**

B Krein-Milman Theorem **120**

C Viscosity solutions **121**

Bibliography **122**

Glossary

Vectors

$x \geq y$	Usual partial ordering : $x_i \geq y_i \ \forall i$.
$x > y$	Strict ordering : $x_i > y_i \ \forall i$.
$\min(x, y)$	Vector whose ith component is $\min(x_i, y_i)$.
$x^\intercal y = \langle x, y \rangle$	Standard euclidean inner product.
$x \circ y$	Vector whose ith component is $x_i y_i$: the Hadamard product.
$x \perp y$	$x^\intercal y = 0$.

Matrices

$A = (A_{ij})$	Matrix A with components A_{ij}.
$A^\intercal$	Transpose of matrix A.
A_{IJ}	Matrix with entries $(A_{ij})_{i \in I, j \in J}$, a submatrix of A.
$A_{I\bullet}$	The rows of A indexed by I.
$A^{-\intercal}$	$(A^{-1})^\intercal = (A^\intercal)^{-1}$

Sets

$\mathcal{B}_\delta(\overline{x})$	For $\delta > 0$, the open ball of radius δ centered at $\overline{x}$.
conv A	For a set A, the convex hull of A.
cl A	For a set A, the closure of A.
$\text{dist}_A(x)$	The Euclidian distance from x to A.
∂A	For a set A, border or frontier of A: $\partial A =$cl $A \setminus$ int A.
int X	Topological interior of a set X
$\overline{\mathbb{R}}$	Extended real line, $\overline{\mathbb{R}} = (-\infty, +\infty]$
$\mathbb{R}^m_+$	Positive orthant: $\{y \in \mathbb{R}^m \mid y \geq 0\}$
$f : X \rightrightarrows Y$	Multimap (or multifunction, or set valued function): for $x \in X$, $f(x)$ is a subset of Y (possibly empty)

Index sets

$\overline{n}$	For a given $n \in \mathbb{N}$, $\overline{n} = \{1, ..., n\}$
$I_t^{0+}(x, u, v)$	$\{i : v_i(t) = 0 < (Cx(t) + Dv(t) + Eu(t))_i\}$
$I_t^{+0}(x, u, v)$	$\{i : v_i(t) > 0 = (Cx(t) + Dv(t) + Eu(t))_i\}$
$I_t^{0+}(x, u, v)$	$\{i : v_i(t) = 0 = (Cx(t) + Dv(t) + Eu(t))_i\}$
I^c	For $I \subset \overline{n}$, $I^c = \{i \in \overline{n};\ i \notin I\}$

Scalar tools

$\lceil x \rceil$ and $\lfloor x \rfloor$	Ceil and floor value of x: $\lceil x \rceil$, $\lfloor x \rfloor \in \mathbb{Z}$, $\lceil x \rceil \geq x \geq \lfloor x \rfloor$
$\text{sgn}(x)$	For $x \in \mathbb{R}$, $\text{sgn}(x) = 1$ if $x > 0$, $\text{sgn}(x) = -1$ if $x < 0$, $\text{sgn}(x) = [-1, 1]$ if $x = 0$

Functional analysis

$L^p([t_0, t_1], \mathbb{R}^n)$	Set of functions $f : [t_0, t_1] \to \mathbb{R}^n$ such that $\int_{t_0}^{t_1} \|f(t)\|^p dt < \infty$.
$\mathscr{C}^p$	Class of p times continuously differentiable functions.

Dynamical systems

DAE	Differential Algebraic Equation, equation of the form $f(x, \dot{x}) = 0$.
$\text{Acc}_\Omega(x_0, t)$	Accessible set from x_0 at time t with control taking values in Ω.

Mathematical Programming

MPCC	Mathematical Programs with Complementarity Constraints.
MPEC	Mathematical Programs with Equilibrium Constraints.
NLP	Nonlinear Program.

Résumé en français

Cette thèse se concentre sur la commande optimale des systèmes de complémentarité linéaires (notés LCS). Les LCS sont des systèmes dynamiques définis par des équations différentielles algébriques (ÉDA), où une des variables est définie par un problème de complémentarité linéaire, qu'on écrit comme : $0 \leq \lambda \perp D\lambda + q \geq 0$.
Ces systèmes se retrouvent dans la modélisation de nombreux phénomènes, tels que les équilibres dynamiques de Nash, les systèmes dynamiques hybrides, la modélisation des systèmes mécaniques avec contact frottant ou encore des circuits électriques. Les propriétés des solutions à ces ÉDA dépendent essentiellement de propriétés que doit vérifier la matrice D présente dans la complémentarité.
Ces contraintes de complémentarité posent des problèmes à deux niveaux. Premièrement, l'analyse de ces systèmes dynamiques fait souvent appel à des outils pointus, et leur étude laissent encore des questions non résolues. Deuxièmement, la commande optimale pour ces systèmes pose des difficultés à cause d'une part de la présence éventuelle de l'état dans les contraintes, et d'autre part une violation assurée des qualifications des contraintes qui sont une hypothèse récurrente des problèmes d'optimisation.
La recherche de ce manuscrit se concentre sur la commande optimale de ces systèmes. On s'intéresse principalement à la commande quadratique (minimisation d'une fonctionnelle quadratique en l'état et la commande), et à la commande temps minimal. Les résultats se concentrent sur deux pans: d'un côté, on opère une approche analytique du problème afin de trouver des conditions nécessaires d'optimalité (si possible, on démontre qu'elles sont suffisantes) ; dans un deuxième temps, un développement logiciel est effectué, avec le soucis d'obtenir des résultats numériques précis de manière rapide.

English summary

This thesis focuses on the optimal control of Linear Complementarity Systems (LCS). LCS are dynamical systems defined through Differential Algebraic Equations (DAE), where one of the variable is defined by a Linear Complementarity Problem, which can be written as: $0 \leq \lambda \perp D\lambda + q \geq 0$.
These systems can be found in the modeling of various phenomena, as Nash equilibria, hybrid dynamical systems or modeling of electrical circuits. Properties of the solution to these DAE essentially depend on properties that the matrix D in the complementarity must meet.
These complementarity constraints induce two different challenges. First, the analysis of these dynamical systems often uses state of the art tools, and their study still has some unanswered questions. Second, the optimal control of these systems causes troubles due to, on one hand, the presence of the state in the constraints, on the other hand the violation of Constraint Qualifications, that are a recurring hypothesis for optimisation problems.
The research presented in this manuscript focuses on the optimal control of these systems. We mainly focus on the quadratic optimal control problem (minimisation of a quadratic functional involving the state and the control), and the minimal time control. The results present two different aspects: first, we start with an analytical approach in order to find necessary conditions of optimality (if possible, these conditions are proved to be sufficient); secondly, a code is developed, with the aim of getting precise results with a reduced computational time.

Introduction

This dissertation tackles the problem of Optimal Control of Linear Complementarity Systems (LCS). Let us review each of these terms.

Linear Complementarity Systems are systems that are described by dynamical linear equations, which means that there is a variable, often named the *state*, which is changing with time. The evolution of this variable is describe by a differential equation, which is linear. But also, this system has to comply with a constraint, which is expressed with *complementarity*. Complementarity means that two quantities can not be *simultaneously active*. In order to understand this properly, let us give a simple physical example, found in [21].

Let us consider the circuit in Figure 1 with a diode D, resistance R and inductance L. The diode in this case is considered to be *ideal* in the following sense: it possesses a voltage/current characteristic that translates the physical observation.

- When the voltage λ is positive, then the current i is cancelled.
- When the current i is negative, then the voltage λ is cancelled.

This can be written mathematically as $\lambda \geq 0$, $i \geq 0$ and $i\lambda = 0$, which we rewrite more compactly as $0 \leq -i \perp \lambda \geq 0$. Denote $x_1(t) = \int_0^t i_1(s)ds + x_1(0)$ (the charge of the capacitor, in coulomb) and $x_2(t) = i_2(t)$ (the electric current, in ampere). Then the evolution of this system is described as:

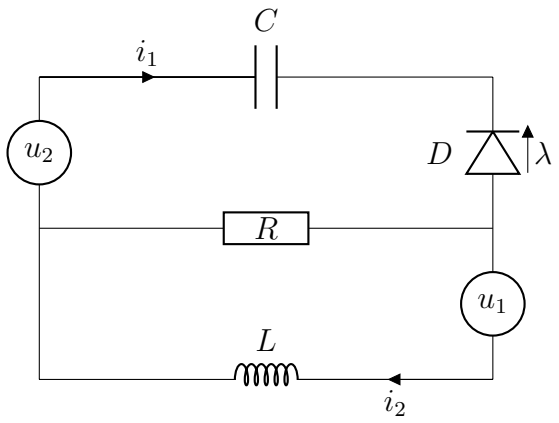


Figure 1: Circuit with an ideal diode and two voltage sources

$$\begin{cases} \dot{x}_1(t) = \dfrac{-1}{RC}x_1(t) + x_2(t) - \dfrac{1}{R}\lambda(t) + \dfrac{1}{R}u_2(t), \\ \dot{x}_2(t) = \dfrac{-1}{LC}x_1(t) - \dfrac{1}{L}\lambda(t) + \dfrac{1}{L}(u_2(t) - u_1(t)), \\ 0 \leq \lambda(t) \perp \dfrac{1}{RC}x_1(t) - x_2(t) + \dfrac{1}{R}\lambda(t) - \dfrac{1}{R}u_2(t) \geq 0, \end{cases} \tag{1}$$

Several questions then arise. For instance, is the mathematical model valid? For this example, it means that the mathematical equation (1) has at least one solution once one sets the source function u, that the solution is preferably unique, and that it has some nice properties (like

continuity on initial data for instance). Also, one could ask if the system is controllable. In this case, the question means that one is wondering if from a given state at time t_0 (the capacity and intensity at time t_0), one can reach an other state later, at a prescribed time t_1.

The second term concerns the Optimal Control problems. These are problems gathering two different facets: the control of dynamical systems, and the optimization. Loosely speaking, these are problems in which one tries to *steer* a dynamical system from one state to another, while *minimizing* (or maximizing) a given criteria. Let us resume the previous example given in Figure 1. Assume one starts from a given state $(x_1(0), x_2(0))$, and the goal is to reach an other state $(x_1(1), x_2(1))$ in one second (since it is possible). But it should not be done recklessly. The aim is to minimize the overall energy at the bounds of the resistance, and the input energy. This can be written as minimizing the functional $\int_0^1 [Rx_2(t)^2 + u_1(t)^2 + u_2(t)^2]dt$. Overall, the Optimal Control problem reads as:

$$\min \int_0^1 [Rx_2(t)^2 + u_1(t)^2 + u_2(t)^2]dt$$
$$\text{such that } \begin{cases} \dot{x}_1(t) = \dfrac{-1}{RC}x_1(t) + x_2(t) - \dfrac{1}{R}\lambda(t) + \dfrac{1}{R}u_2(t), \\ \dot{x}_2(t) = \dfrac{-1}{LC}x_1(t) - \dfrac{1}{L}\lambda(t) + \dfrac{1}{L}(u_2(t) - u_1(t)), \\ 0 \leq \lambda(t) \perp \dfrac{1}{RC}x_1(t) - x_2(t) + \dfrac{1}{R}\lambda(t) - \dfrac{1}{R}u_2(t) \geq 0, \\ (x_1(0), x_2(0)) \text{ fixed} \\ (x_1(1), x_2(1)) \text{ fixed} \end{cases}$$

As expected, the resolution is not trivial. Once again, several questions arise: does there exist at least one solution to this problem, with nice properties (like x_1 and x_2 being absolutely continuous and u square integrable)? How could one characterize the solution(s) with necessary conditions of optimality? Are these conditions sufficient? In most cases, one can not compute explicitly a solution; are there ways to compute a numerical approximation?

This thesis focuses on the optimal control problem for LCS in an abstract framework:

$$\min f(T, x, u)$$
$$\text{such that } \begin{cases} \dot{x}(t) = Ax(t) + Bv(t) + Fu(t) \\ 0 \leq v(t) \perp Cx(t) + Dv(t) + Eu(t) \geq 0 \quad \text{a.e. on } [0, T] \\ (x(0), x(T)) \in \mathcal{T} \end{cases}$$

Two problems will be studied: the quadratic optimal control problem (i.e. f is a quadratic functional in x and u), and the minimal time control problem (i.e. $f(T, x, u) = T$).

Outline

This manuscript is divided in four parts as follows:

- In a first part, a review of the existing results in the literature is made. We will then acknowledge the different tools available in order to tackle the optimal control of LCS, and also why they are eventually too limited.

- In a second part, a first attempt for tackling the quadratic optimal control problem for LCS is made. It relies on assumptions on the underlying complementarity conditions which turns the system into a Lipschitz system, non differentiable. Some first order conditions are then derived, a first code to approximate the solution is written.

- In a third part, the results of Part 2 on the quadratic optimal control problem of LCS are enhanced. This time, properly defined multipliers are added to the necessary conditions, which are in turn transformed in order to be more efficiently handled. Also, these necessary conditions are proved to be also sufficient. With these results, a code is developed in order to compute an approximation of the solution. Two different approaches are tested, which appear to be eventually complementary.

- Finally, the last part focuses on the minimal time problem for LCS. After extending some results of the existing literature, these results are applied specifically on LCS in order to derive sufficient first order conditions. Then, a geometrical analysis of the shape of the complementarity constraints allows us to prove a bang-bang property for LCS.

List of publications

1. A. Vieira, B. Brogliato, and C. Prieur. Preliminary results on the optimal control of linear complementarity systems. IFAC World Congress - Toulouse and IFAC-PapersOnLine, 50(1):2977 – 2982, 2017.

2. A. Vieira, B. Brogliato, and C. Prieur. Quadratic Optimal Control of Linear Complementarity Systems: First order necessary conditions and numerical analysis. Submitted to IEEE Transactions on Automatic Control, January 2018.

A last paper, concerning the minimal time problem (content of Chapter 8) is under preparation. These results were partly presented during the conference **Control of state constrained dynamical systems** (https://events.math.unipd.it/CoSCDS/) at the Department of Mathematics *Tullio-Levi Civita* of Padua University, and during ISMP 2018 (https://ismp2018.sciencesconf.org/) in Bordeaux.

Part I

State of the art

Chapter 1

Linear Complementarity Systems

Abstract. In this chapter, the Linear Complementarity Systems (LCS) are presented. After some results concerning the Linear Complementarity Problem (LCP), which is the specificity of these systems, some properties of the LCS are shown, with an effort for linking them with other types of systems appearing in the literature (like hybrid automata or differential inclusions). A last section focuses on the numerical simulation of LCS.

At the core of the problem tackled by this thesis, one finds Linear Complementarity Systems (LCS). These dynamical systems are usually described as a linear dynamical system where one of the components is defined though a complementarity problem. More precisely, we call LCS($A(\cdot)$, $B(\cdot)$, $C(\cdot)$, $D(\cdot)$) the dynamical system:

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)\lambda(t) \\ 0 \leq \lambda(t) \perp C(t)x(t) + D(t)\lambda(t) &\geq 0 \qquad \text{a.e. on } [t_0, t_1] \\ x(0) &= x_0 \in \mathbb{R}^n \end{aligned} \tag{1.1}$$

where $t_0, t_1 \in \mathbb{R}$, $t_0 < t_1$, $x : [t_0, t_1] \to \mathbb{R}^n$, $\lambda : [t_0, t_1] \to \mathbb{R}^m$, and $A(\cdot), B(\cdot), C(\cdot), D(\cdot)$ are matrices of according dimensions. This provides a modeling paradigm for many problems, as Nash equilibrium games, hybrid engineering systems [19], contact mechanics or electrical circuits [2].
One could use the framework of differential inclusion in order to analyze solutions of problem (1.1), since $\lambda(t)$ could be single or set valued depending on properties of the matrix D. Admit for now that $\lambda(t)$ is uniquely defined for every $t \in [t_0, t_1]$ (for instance if $D(t)$ is a **P**-matrix). As it will be stated in Section 1.1, λ is then a piecewise linear function of x. Therefore, the right-hand side defining the dynamics in (1.1) is a piecewise linear function of x, and therefore a Lipschitz function of x. Then, assuming that every matrix is a continuous function of time t, Cauchy-Lipschitz theorem proves that there exists a unique maximal solution x of (1.1) starting from $x(t_0) = x_0 \in \mathbb{R}^n$ (we could even argue that the solution exists on $[t_0, t_1]$).
Since λ is a piecewise linear function of x, it means that the exists ℓ matrices $\{\Lambda_i(\cdot)\}_{i=1}^{\ell}$ such that:

$$\exists i \in \overline{\ell};\ \dot{x}(t) = \Lambda_i(t)x(t),\ \text{a.e. on } [t_0, t_1]$$

It shows that there exists a connection between LCS and an other class of systems called switching systems. The latter ones encompass both continuous and discrete dynamics, and it has been a popular subject of study in recent years (see for instance [8]). This framework also generalizes Differential Algebraic Equations (DAE), that are often non sufficient in order to describe models naturally occurring in engineering problems that contain inequalities (for instance, unilateral constraints) and disjunctive conditions (for conditional phenomena such as contacts).

1.1 Linear Complementarity Problems

Roughly speaking, the Linear Complementarity Problem (LCP) is to find $z \in \mathbb{R}^n$ such that:

$$0 \leq z \perp Mz + q \geq 0, \tag{1.2}$$

where $M \in \mathbb{R}^{n \times n}$ and $q \in \mathbb{R}^n$ is a given vector. This notation means that each component of z and $Mz+q$ must be nonnegative, and both vectors must be perpendicular to each other, which in this case translates to:

$$z_i(Mz+q)_i = 0, \ \forall i \in \{1, ..., n\}.$$

One usually denotes the problem by $\mathrm{LCP}(q, M)$, and the set of its solution by $\mathrm{SOL}(q, M)$. This problem appears naturally when one searches for first order conditions to a finite-dimensional optimization problem - also known as KKT conditions. Indeed, KKT conditions to a linear-quadratic problem of the form:

$$\min_{z \in \mathbb{R}^n} \frac{1}{2} z^\intercal H z + c^\intercal z$$
$$\text{s.t.} \begin{cases} Az \geq b, \\ z \geq 0, \end{cases}$$

where $H \in \mathbb{R}^{n \times n}$ is a symmetric matrix, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$, are expressed as follows:

$$\begin{aligned} Hz + c - A^\intercal \lambda - y &= 0, \\ 0 \leq \lambda \perp Az - b &\geq 0, \\ 0 \leq y \perp z &\geq 0, \end{aligned}$$

where $y \in \mathbb{R}^n$, $\lambda \in \mathbb{R}^m$ are multipliers. By extracting y from the first equality, one obtains:

$$0 \leq \begin{pmatrix} z \\ \lambda \end{pmatrix} \perp \begin{pmatrix} H & -A^\intercal \\ A & 0 \end{pmatrix} \begin{pmatrix} z \\ \lambda \end{pmatrix} + \begin{pmatrix} c \\ -b \end{pmatrix} \geq 0.$$

The existence of solutions and their properties rely heavily on the structure of the matrix M. This chapter will present some results concerning the analysis of this problem that will be useful for the rest of this manuscript. Main information stated here can be found in [45].

1.1.1 Equivalent formulations

In order to analyze properties of the solutions of problem (1.2), we must reformulate the problem in an other framework.

Optimization problem As suggested by the former example, a first way is to express (1.2) as an optimization problem. It is easy to state the following proposition:

Proposition 1.1.1. *Let $q \in \mathbb{R}^n$, $M \in \mathbb{R}^{n \times n}$. $z \in \mathbb{R}^n$ is a solution of $LCP(q, M)$ if and only if it is a global solution to the following quadratic problem:*

$$\begin{aligned} \min_{z \in \mathbb{R}^n} \ & z^\intercal (Mz + q) \\ \textit{such that } & Mz + q \geq 0, \\ & z \geq 0, \end{aligned} \tag{1.3}$$

with an objective value of zero.

C-function A second way is to express this problem as finding the root of a two-argument function, called C-function.

Definition 1.1.1. *We call C-function a function* $f : \mathbb{R}^{2n} \to \mathbb{R}^n$ *such that:*

$$f(a,b) = 0 \iff \langle a,b\rangle = 0,\ a,b \geq 0$$

If f is a C-function, then LCP(q,M) is equivalent to find $z \in \mathbb{R}^n$ such that $f(z, Mz+q) = 0$. Most known C-functions are:

- min function: z is a solution of LCP(q,M) iff $\min(z, Mz+q) = 0$.
- Fischer-Burmeister's C-function: z is a solution of LCP(q,M) iff $\sqrt{z_i^2 + (Mz+q)_i^2} - (z_i + (Mz+q)_i) = 0$, $\forall i \in \{1,n\}$.
- normal map: z is a solution of LCP(q,M) iff $Mz^+ + q - z^- = 0$, where $z^+ = \max(z,0)$, and $z^- = \max(-z,0)$.

In general, C-functions are not Fréchet-differentiable, in particular at the origin. Even if this reformulation seems appealing, this non-differentiability makes the use of this technique tricky in practice. In relation with this fact, we introduce the notion of degeneracy of the solution.

Definition 1.1.2. *A solution z of LCP(q,M) is said non-degenerate if for each $i \in \{1,...,n\}$, $z_i \neq (Mz+q)_i$.*

If a solution is non-degenerate, then points in a neighbourhood around the solution are also non-degenerate. In this case, C-functions are usually Fréchet-differentiable, and locally convergent method for solving root problems of smooth functions (like the Newton method) will work efficiently to solve this problem.

Piecewise affine functions The min function is of a special kind: it is a piecewise affine function. Such functions actually play a major role in the analysis of solutions of LCP. More formally speaking:

Definition 1.1.3. *A function $f : \mathcal{D} \to \mathbb{R}^m$, where $\mathcal{D} \subseteq \mathbb{R}^n$, is said to be piecewise affine function if f is continuous and $\mathcal{D}$ is equal to the union of a finite number of convex polyhedra, called the pieces of f, on each of which f is an affine function.*

The reformulation as the zero of a min-function already underlines the interaction between the two notions, but it goes even a bit further. Indeed, if $f : \mathbb{R}^n \to \mathbb{R}^n$ is an arbitrary piecewise affine function, then under a "nonsingularity" assumption, the system $f(z) = 0$ is equivalent to a certain LCP.
Another link between these two notions lies in the next proposition:

Proposition 1.1.2. *[45] Let $M \in \mathbb{R}^{n\times n}$ be such that the LCP(q,M) has a unique solution for all vectors $q \in \mathbb{R}^n$. Then the unique solution of the LCP(q,M) is a piecewise linear function of q.*

Convex subdifferential Eventually, a final way to express the set of solutions is through the subdifferential of the indicator function $\mathbb{I}_{\mathbb{R}^+_m}$, which is defined as:

$$\mathbb{I}_{\mathbb{R}^m_+}(x) = \begin{cases} 0 & \text{if } x \in \mathbb{R}^m_+ \\ +\infty & \text{if } x \notin \mathbb{R}^m_+. \end{cases}$$

Since $\mathbb{R}^m_+$ is convex, its indicator function is also convex, and we can use tools of convex analysis, and in particular the subdifferential of a convex function. It can be proved easily that the subdifferential of the indicator function $\partial\mathbb{I}_{\mathbb{R}^m_+}(x)$ is equal to the normal cone of convex analysis $\mathcal{N}_{\mathbb{R}^m_+}(x)$. It justifies the following equivalence:

$$\begin{aligned} 0 \leq z \perp \zeta \geq 0 \iff & \quad -z \in \mathcal{N}_{\mathbb{R}^m_+}(\zeta). \\ \iff & \quad -\zeta \in \mathcal{N}_{\mathbb{R}^m_+}(z) \end{aligned}$$

1.1.2 Class of matrices

A sensible question that may be asked is the following: what is the class of matrices M for which LCP(q, M) has a solution for all vectors $q \in \mathbb{R}^n$? This class is denoted $\mathbf{Q}$, and its elements are called $\mathbf{Q}$-matrices. Unfortunately, there is no algebraic description allowing to check in finite time if a matrix is a $\mathbf{Q}$-matrix or not.

Global uniqueness If we narrow the problem by imposing uniqueness of the solution, we have more comprehensive results.

Definition 1.1.4. *A matrix M is said to be a $\mathbf{P}$-matrix if LCP(q, M) admits a unique solution for all $q \in \mathbb{R}^n$. The class of such matrices is denoted $\mathbf{P}$.*

The next theorem gives a full description of these $\mathbf{P}$-matrices.

Theorem 1.1.1. *[45] Let $M \in \mathbb{R}^{n\times n}$. The following statements are equivalent:*

1. *All principal minors of M are positive.*
2. *M reverses the sign of no nonzero vector, i.e.:*
$$[z_i(Mz)_i \leq 0 \text{ for all } i] \implies [z = 0].$$
3. *All real eigenvalues of M and its principal submatrices are positive.*
4. *M is a $\mathbf{P}$-matrix.*

A special subclass of $\mathbf{P}$-matrices are the symmetric positive definite matrices. From a symmetric positive definite matrix M, one can define a norm: for $x \in \mathbb{R}^m$, $\|x\|_M = \sqrt{x^\intercal Mx}$. With this, we define the projection on a convex closed set $K \subset \mathbb{R}^n$ with the metric defined by M, denoted as $\text{proj}_{K,M}$. In this sense, for all $x \in \mathbb{R}^n$, $\text{proj}_{K,M}(x)$ is the closest point to x in K according to $\|\cdot\|_M$. Of course, if M is the identity matrix, then the usual projection operator is retrieved (simply denoted as proj_K). In this framework, we have a formulation for the solution of an LCP.

Proposition 1.1.3. *Let $M \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix and $q \in \mathbb{R}^n$. There exists a unique x, solution of LCP(q, M), and:*

$$x = \text{proj}_{\mathbb{R}^n_+,M}\left(-M^{-1}q\right).$$

Proof. This is a simple application of Propositions 1.5.9 and 4.3.3 of [54]. □

Local uniqueness Asking for a global solution may be too restrictive. By pursuing the analogy with the necessary conditions of the optimization problems, rather than searching for a global solution, we could search for a local one. In other words, if z^* is the solution of LCP(q, M), then there exist no other solution in a neighbourhood around z^*. This problem is entirely described as follows:

Definition 1.1.5. *A matrix $M \in \mathbb{R}^{n\times n}$ is called nondegenerate if all its principal minors are nonzero.*

Theorem 1.1.2. *[45] Let $M \in \mathbb{R}^{n\times n}$. The following statements are equivalent:*

1. *M is nondegenerate.*
2. *For all vectors q, the LCP(q, M) has a finite number (possibly zero) of solutions.*
3. *For all vectors q, any solution of the LCP(q, M), if it exists, must be locally unique.*

1.1.3 Numerical resolution

Of course, most LCP can not be solved analytically, but there exist several techniques to find an approximate solution. There are two major families of techniques:

1. Pivoting techniques, based on the idea of pivoting as found in numerical algebra and linear programming. Examples as the criss-cross algorithm or Lemke's algorithm show that they terminate in a finite number of iterations (for a certain class of problems), but are in the worst case of exponential complexity. Typically, these algorithms produce a sequence of vector pairs $\{(y^k, z^k)\}$ that are extreme points of the feasible region $\{(y, z)|y \geq 0, z \geq 0,\ y = Mz + q\}$ by moving inside the kernel of a basis matrix until a new boundary of the feasible region is encountered. These are good solutions for small to medium size problems, but the efficiency decreases as the problem dimension increases, due to round-off errors and data storage.

2. Iterative methods, which do not solve the problem in finite number of iterations, but converge in the limit. They can exploit the sparsity of the problem, and are less sensitive to the round-off errors. A well known example are the interior point methods, which transform the problem (1.2) into the unconstrained minimisation problem:

$$\min\ \langle x, y\rangle + \|y - Mx - q\| - \mu \sum_i (\log x_i + \log y_i)$$

   where $\mu > 0$ is a parameter continuously driven to 0. As such, the solutions $(x(\mu), y(\mu))$ of this problem trace out a *central path* that leads to a solution of LCP(M, q). Interior methods are well described in [111] and references therein.

1.2 Properties of the LCS

1.2.1 LCS seen as a hybrid automaton

Let us first describe LCS as hybrid automata to see the connection between them and stress the limits. Notations exposed here are taken from [17].

Definition 1.2.1. *A hybrid automaton is given by* (Q, Σ, J, G) *where:*

- Q *is a finite set of modes (sometimes called discrete states or locations).*
- $\Sigma = \{\Sigma_q\}_{q\in Q}$ *is a collection of dynamical systems. For mode* q*, these are given by the ODE* $\dot{x} = f_q(x)$ *or by the DAE* $f_q(\dot{x}, x) = 0$*.*
- $J = \{J_q\}_{q\in Q}$*.* $J_q \subset \mathbb{R}^n$ *is the jump set for mode* q *consisting of the states from which a mode transition and/or state jump occurs.*
- $G = \{G_q\}$ *is the set of jump transition maps where* G_q *is a (possibly multi-valued) map from* J_q *to a subset of* $\mathbb{R}^n \times Q$*.*

Let's now try to describe (1.1) as a hybrid automaton. The following description is inspired by [63] and [64]. For simplicity of exposition, we denote $y = Cx + D\lambda$ and suppose A, B, C, D autonomous.
First, let us notice that the LCP states that $\lambda_i(t) = 0$ or $y_i(t) = 0$ for each $i \in \overline{n}$. This results in a multimodal system with 2^m modes, where each mode is characterized by a subset I of $\overline{m}$. Hence, $Q = \mathcal{P}(\overline{m})$, the power set of $\overline{m}$. The dynamics f_I in mode I are given by the DAE:

$$\begin{aligned} \dot{x} &= Ax + B\lambda, \\ y &= Cx + D\lambda, \\ y_i &= 0, \ i \in I \\ \lambda_i &= 0, \ i \in I^c \end{aligned} \tag{1.4}$$

Also, the LCP imposes the sign condition:

$$\lambda_i(t) \geq 0, i \in I, \ y_i(t) \geq 0, \ i \in I^c \tag{1.5}$$

Therefore, the jump set J_I is given by

$$J_I = \{x_0 \in \mathbb{R}^n \mid \text{there is no smooth solution } (\lambda, x, y) \text{ of (1.4)} \\ \text{for mode } I \text{ satisfying } x(0) = x_0 \text{ and (1.5) on } [0, \varepsilon[\text{ for some } \varepsilon > 0\}$$

A state x_0 is said to be consistent for mode I if $x_0 \notin J_I$. The set of consistent states for mode I is denoted V_I. Define the set T_I as the limit of the sequence:

$$T_0 = \{0\}, \ T_{i+1} = \{x \in \mathbb{R}^n | \exists \lambda \in \mathbb{R}^m, \ \exists \bar{x} \in T_i \text{ such that } \bar{x} = Ax + B\lambda, \ C\bar{x} + D\lambda = y, y_I = 0, \ \lambda_{I^c} = 0\}$$

It is proved in [62] that this sequence converges in at most n steps. The jump transition function G only depends on the state $x(\tau^-)$ just before the event time τ, and not on the previous mode. The jump transition map is given by:

$$G(x) = \{(x^+, I^+) \in \mathbb{R}^n \times \mathcal{P}(\bar{m}) | x^+ = \Pi_{V_{I^+}}^{T_{I^+}}(x)\} \tag{1.6}$$

where $\Pi_{V_I}^{T_I}$ is the projection onto V_I along T_I. There exist then different strategies to select a proper transition in $G(x)$.

Presented that way, we see many computational drawbacks at converting back the LCS (1.1) in the framework of hybrid systems. First, the number of modes grow exponentially as m grows. Secondly, consistent spaces and transition maps are not easy to describe in a useful way. Even in the case when the LCP condition defines λ uniquely, describing precisely the transition map is not an easy task: how do you choose the next mode I^+?
Even though this hybrid representation is an important tool for analysis, it is not the most efficient way to handle this system.

1.2.2 LCS seen as a differential inclusion

As it was shown in Section 1.1.1, the complementarity problem can be equivalently defined as the inclusion of the solution to a normal cone. Therefore, LCS$(A, B, C, 0)$ can be equivalently defined as the Differential Inclusion (DI):

$$\dot{x} \in Ax - B\mathcal{N}_{\mathbb{R}^m_+}(Cx) \tag{1.7}$$

Suppose that there exists a symmetric positive definite matrix R such that $R^2B = C^\intercal$ and define $z = Rx$. Then one can prove (see [19, 20, 56]) that LCS$(A, B, (R^2B)^\intercal, 0)$ can be equivalently expressed as the DI:

$$-\dot{z}(t) + RAR^{-1}z(t) \in \mathcal{N}_S(z(t)) \tag{1.8}$$

where $S = \{Rx | Cx \geq 0\}$. The equivalence is here understood in the sense that the two formalisms are strictly the same way of writing a mathematical object without consideration on the solution. A huge study of differential inclusions can be found in [98], where the author focuses on differential inclusion of the type $\dot{x} \in F(x)$ for some multi-valued map F, but some hypothesis (such as a boundedness property of F) put (1.7) out of its scope. The closest results concerning systems such as (1.8) concern the *sweeping process*, introduced by Moreau [81]. Originally, a sweeping process is a differential inclusion defined as $-\dot{x}(t) \in \mathcal{N}_{C(t)}(x(t))$, for some convex valued multifunction C. These systems, under some hypothesis on C, admit some properties, such as uniqueness of solution for the Cauchy problem defined with this inclusion. Various forms of such systems have been analyzed; see for instance [3, 18, 30, 41, 68].

For a general system LCS(A, B, C, D), we have to put the system under the form of a Differential Algebraic Inclusion (DAI). Create an auxiliary variable λ defined by $Cx + Dv - \lambda = 0$. Then, using the equivalence presented in Section 1.1.1, we know that $-v \in \mathcal{N}_{\mathbb{R}^m_+}(\lambda)$. Reintroducing it into the remaining equations, we obtain the DAI:

$$\begin{cases} \dot{x} - Ax \in -B\mathcal{N}_{\mathbb{R}^m_+}(\lambda) \\ Cx \in \lambda + D\mathcal{N}_{\mathbb{R}^m_+}(\lambda) \end{cases} \tag{1.9}$$

Such systems are to date little studied. One can find some results in [26, 80].

1.2.3 Control of the LCS

Let us now turn to the control of such systems. There exist also some properties known for a certain class of LCS. Consider the input/output system:

$$\dot{x}(t) = Ax(t) + B\lambda(t) + Fu(t), \tag{1.10a}$$

$$y(t) = Cx(t) + D\lambda(t) + Eu(t), \tag{1.10b}$$

$$0 \leq \lambda(t) \perp y(t) \geq 0, \tag{1.10c}$$

where, compared to LCS(A, B, C, D), we just add an input $u : [t_0, t_1] \to \mathbb{R}^k$ and $F \in \mathbb{R}^{n\times k}$, $E \in \mathbb{R}^{m\times k}$. There exist several results concerning properties satisfied by this system. We present here a few of them that will highlight some future results exposed later. In particular, these results show that the set of absolutely continuous functions may be *too small* in order to define a state trajectory $x(\cdot)$ solution of (1.10).

L^2 solutions

The well-posedness of those systems have been analyzed in [32]. We summarise here their results. First of all, in order to state properly well-posedness for LCS, we must define clearly some concepts.

Definition 1.2.2.

- *$LCS(A,B,C,D)$ is said to be passive (or dissipative with respect to the supply rate $\langle \lambda, y\rangle$) if there exists a function $V : \mathbb{R}^n \to \mathbb{R}^+$ (called a storage function) such that:*

$$V(x(t)) + \int_t^{t'} \langle \lambda(t), y(t)\rangle dt \geq V(x(t'))$$

holds for all t, t' with $t \leq t'$, and all $(x,\lambda) \in L^2([t,t'],\mathbb{R}^{n+m})$ satisfying (1.1) and $y(t) = Cx(t) + D\lambda(t)$.

- *A function $f : \mathbb{R}^+ \to \mathbb{R}$ is called Bohl function if it has a rational Laplace transform. The set of such functions is denoted by $\mathcal{B}$. We call $\mathtt{f}$ a Bohl distribution if $\mathtt{f} = \mathtt{f}_{\mathtt{reg}} + \mathtt{f}_{\mathtt{imp}}$ with the impulsive part $\mathtt{f}_{\mathtt{imp}} = \sum_{i=0}^{\ell} u_i \delta_0^{(i)}$, where δ_0 the Dirac function centered at 0, and $\delta_0^{(i)}$ is its ith derivative, and the regular part $\mathtt{f}_{\mathtt{reg}} \in \mathcal{B}$.*

- *A Bohl distribution $\mathtt{f}$ is initially nonnegative if its Laplace transform $\hat{\mathtt{f}}$ satisfies $\hat{\mathtt{f}}(\sigma) \geq 0$ for sufficiently large σ.*

- *f is said to be a piecewise Bohl function if f is right-continuous and there exists a collection of isolated points $\Gamma_w = \{\tau_i\} \subseteq \mathbb{R}^+$ (called the transition points) such that for every i, there exists a function $g \in \mathcal{B}$ such that $f_{|(t_i,t_{i+1})} = g_{|(t_i,t_{i+1})}$. We denote this space $\mathcal{PB}$.*

- *The distribution space $L^{2,\delta}(\mathbb{R}^+)$ is defined as the set of all $\mathtt{u} = \mathtt{u}_{\mathtt{imp}} + \mathtt{u}_{\mathtt{reg}}$ where $\mathtt{u}_{\mathtt{imp}} = \sum_{\theta\in\Gamma} u_\theta \delta_\theta$ for $u_\theta \in \mathbb{R}$ with a set of isolated points $\Gamma \subset \mathbb{R}^+$ and $\mathtt{u}_{\mathtt{reg}} \in L^2_{loc}(\mathbb{R}^+)$.*

Definition 1.2.3 (Initial solution)**.** *The Bohl distribution $(\mathtt{v},\mathtt{x},\mathtt{y})$ is an initial solution to (1.10) with initial state x_0 and input $u \in \mathcal{B}$ if:*

1. *The equations $\dot{\mathtt{x}} = A\mathtt{x} + B\mathtt{v} + Eu + x_0\delta$, $\mathtt{y} = C\mathtt{x} + D\mathtt{v} + Fu$ hold in the distributional sense.*
2. *There exists a $J \subseteq \overline{m}$ such that $\mathtt{v}_i = 0$, $i \in \overline{m}\backslash J$ and $\mathtt{y}_i = 0$, $i \in J$, as equalities of distributions.*
3. *The distributions $\mathtt{v}$ and $\mathtt{y}$ are initially nonnegative.*

In the following Theorem, we denote by Γ_{Eu} the set of times when $Eu(\cdot)$ is discontinuous.

Theorem 1.2.1. *[32, Theorem 7.5] Consider an LCS given by (1.10) such that $LCS(A,B,C,D)$ is passive with the storage function $x \mapsto \frac{1}{2}x^\intercal Kx$ for some matrix K positive definite, and $\begin{pmatrix} B \\ D + D^\intercal \end{pmatrix}$ has full column rank. Then, for any initial state x_0 and any input function $u \in \mathcal{PB}^m$, (1.10) admits a unique global solution $(\mathtt{x},\mathtt{y},\mathtt{v}) \in \left(L^{2,\delta}(\mathbb{R}^+)\right)^{n+m+m}$ satisfying:*

1. *$\mathtt{x}_{imp} = 0$, and impulses in $(\mathtt{v},\mathtt{y})$ only show up at times in $\Gamma = \{0\} \cup \Gamma_{Eu}$.*

2. *For any interval* (a,b) *such that* $(a,b)\cap\Gamma=\emptyset$, $\mathtt{x}_{\mathtt{reg}|_{(a,b)}}$ *is absolutely continuous and satisfy for almost all* $t\in(a,b)$*:*

$$\dot{\mathtt{x}}_{\mathtt{reg}}(t)=A\mathtt{x}_{\mathtt{reg}}(t)+B\mathtt{v}_{\mathtt{reg}}(t)+Fu(t)$$

$$\mathtt{y}_{\mathtt{reg}}(t)=C\mathtt{x}_{\mathtt{reg}}(t)+D\mathtt{v}_{\mathtt{reg}}(t)+Eu(t)$$

$$0\leq\mathtt{v}_{\mathtt{reg}}(t)\perp\mathtt{y}_{\mathtt{reg}}(t)\geq 0$$

3. *For each* $\theta\in\Gamma$*, the corresponding impulse* $(y_\theta\delta_\theta,v_\theta\delta_\theta)$ *is equal to the impulsive part of the unique initial solution to* (1.10) *with initial state* $\mathtt{x}_{\mathtt{reg}}(\theta^-)$ *and input* $t\mapsto u(t-\theta)$*.*

4. *For time* $\theta\in\Gamma$*, it holds that* $\mathtt{x}_{\mathtt{reg}}(\theta^+)=\mathtt{x}_{\mathtt{reg}}(\theta^-)+Bu_\theta$*.*

Some results where the solutions of LCS encompass higher order derivatives of the Dirac function may be found in [3, 63].

BV solutions

Suppose $D=0$ and there exists a matrix R symmetric positive definite such that $R^2B=C^\intercal$. Resuming the presentation made in Section 1.7, one can prove (see [22]) that (1.10) is equivalently defined as the *perturbed sweeping process*:

$$-\dot{z}(t)+RAR^{-1}z(t)+RFu(t)\in\mathcal{N}_{S(t)}(z(t)) \tag{1.11}$$

where $S(t)=\{Rx|Cx+Eu(t)\geq 0\}$. The term *perturbed* comes from the fact that the term $RAR^{-1}z(t)+RFu(t)$ is added. The perturbed sweeping processes offer a framework allowing for a different analysis.
Let us state first some definitions. The variation of $x(\cdot)$ on $[t_0,t_1]$ is the supremum of $\sum\|x(t_i)-x(t_{i-1})\|$ over the set of all finite sets of points $t_2<\ldots<t_k$ of $[t_0,t_1]$. When this supremum is finite, the mapping $x(\cdot)$ is said to be of *bounded variation* on $[t_0,t_1]$. $x(\cdot)$ is of *locally bounded variation* on $[t_0,t_1]$ if it is of bounded variation on each compact subinterval of $[t_0,t_1]$.
Considering a set-valued mapping $S:[t_0,t_1]\rightrightarrows\mathbb{R}^n$ and replacing the above expression $\|z(t_i)-z(t_{i-1})\|$ by the Hausdorff distance $\mathrm{haus}(S(t_i),S(t_{i-1}))$, one obtains the concept of set valued mappings with (locally) bounded variation on $[t_0,t_1]$. The Hausdorff distance between two subsets Q_1 and Q_2 in $\mathbb{R}^n$ is given by

$$\mathrm{haus}(Q_1,Q_2)=\max\left\{\sup_{x\in Q_1}\inf_{y\in Q_2}\|x-y\|,\sup_{x'\in Q_2}\inf_{y'\in Q_1}\|x'-y'\|\right\}$$

Denoting by $\mathrm{var}_S(t)$ the variation of $S(\cdot)$ over $[t_0,t]$, $S(\cdot)$ is said *locally absolutely continuous* on $[t_0,+\infty[$ if $\mathrm{var}_S(\cdot)$ is locally absolutely continuous on $[t_0,+\infty[$.
We make the following assumption on (1.10):

Assumption 1.2.1. *Let* $D=0$*. There exists a symmetric positive definite matrix* R *such that* $R^2B=C^\intercal$*.*

Given $u:[t_0,+\infty[\to\mathbb{R}^m$, define $K(t)=\{x\in\mathbb{R}^n|Cx+Fu(t)\geq 0\}$. We can now state the well-posedness theorem:

Theorem 1.2.2. *[22, Theorem 3.5] Assume that $u(\cdot) \in L^1_{loc}([t_0, +\infty[, \mathbb{R}^m)$, that Assumption 1.2.1 holds and that the set-valued mapping $S(\cdot) = R(K(\cdot))$ is locally absolutely continuous (resp. right continuous of locally bounded variation) with nonempty values. Then (1.10) with initial condition $x(0) \in R(K(0))$ has one and only one locally absolutely continuous (resp. right continuous of locally bounded variation) solution $x(\cdot)$ on $[0, +\infty[$.*

This result can be extended to the case $D \geq 0$ or even for nonlinear systems, under some further hypothesis.
The fact that a solution of the differential equation (1.10) may be only right continuous and of locally bounded variation may seem unnatural. This is due to a reformulation of (1.10) into a *measure differential inclusion* [81], which extends the notion of differential inclusions in order to include state jumps representations. This result will explain some results obtained in numerical simulation presented in Chapter 7.
Also, this may seem in contradiction with the results presented in the previous paragraph about L^2 solutions. Actually, these two results are different: the former solution concept proves global existence and uniqueness of a solution $x(\cdot)$ on $L^2(\mathbb{R}^+)$, and the latter proves the global existence and uniqueness of a solution right continuous with bounded variation. Since a function may belong to $L^2(\mathbb{R}^+)$ and not be right continuous of bounded variation (and vice versa), these two results are actually different.

Complete controllability of a class of LCS

First, let us formulate an Assumption on (1.10):

Assumption 1.2.2. *The following conditions are satisfied for the LCS (1.10)*

- *The matrix D is a **P**-matrix.*
- *The transfer matrix $E + C(sI - A)^{-1}F$ is invertible as a rational matrix.*

These assumptions are actually really restrictive, but they make the analysis easier. The second assumption for instance requires that the number of inputs and the number of complementarity variables be the same. It follows from the first assumption that for each initial state x_0 and bounded locally integrable input u there exist a unique absolutely continuous state trajectory $x^{x_0,u}$ and locally integrable trajectories $(\lambda^{x_0,u}, y^{x_0,u})$ such that $x^{x_0,u}(0) = x_0$ and the triple $(x^{x_0,u}, \lambda^{x_0,u}, y^{x_0,u})$ satisfies the relations (1.10) for almost all $t \geq 0$.
A major result concerning (1.10) is on the complete controllability of the system. We recall the definition here:

Definition 1.2.4. *The LCS (1.10) is said completely controllable if for any pair of states $(x_0, x_f) \in \mathbb{R}^{n+n}$, there exists a locally integrable input u such that the associated trajectory $x^{x_0,u}$ satisfies $x^{x_0,u}(t_1) = x_f$.*

Theorem 1.2.3. *[31] Consider an LCS (1.10) satisfying Assumption 1.2.2. It is completely controllable if and only if the following two conditions hold:*

1. *The pair $(A, [F, B])$ is controllable.*
2. *The system of inequalities*

$$\eta \geq 0 \tag{1.12a}$$

$$\begin{pmatrix}\xi^\intercal & \eta^\intercal\end{pmatrix}\begin{pmatrix}A-\nu I & F\\ C & E\end{pmatrix}=0 \tag{1.12b}$$

$$\begin{pmatrix}\xi^\intercal & \eta^\intercal\end{pmatrix}\begin{pmatrix}B\\ D\end{pmatrix}\leq 0 \tag{1.12c}$$

admits no solution $\nu\in\mathbb{R}$ *and* $0\neq(\xi,\eta)\in\mathbb{R}^{n+m}$.

1.2.4 Zeno behavior

This definition of solution hides a detail. In the community dealing with switching systems, an assumption is often made: the trajectory admits no Zeno-state, meaning there is no time with an infinite accumulation of switch events. We define here more precisely what is the Zeno phenomenon. From LCS(A,B,C,D) (1.1), define three sets of indices, two called the active sets:

$$I_t^{+0}(x,\lambda)=\{i\in\overline{m}\ :\ \lambda_i(t)>0=Cx(t)+D\lambda(t)\} \tag{1.13a}$$

$$I_t^{0+}(x,\lambda)=\{i\in\overline{m}\ :\ \lambda_i(t)=0<Cx(t)+D\lambda(t)\} \tag{1.13b}$$

and the last called the degenerate set:

$$I_t^{00}(x,\lambda)=\{i\in\overline{m}\ :\ \lambda_i(t)=0=Cx(t)+D\lambda(t)\} \tag{1.13c}$$

Definition 1.2.5. *Let* (x,λ) *be a solution trajectory of* (1.1)*,* $t^*\in[t_0,t_1]$*, and let* $x(t_*)=x^*$*. We say* x^* *is:*

- *left non-Zeno relative to* (x,λ) *if a scalar* $\varepsilon_->0$ *and a triple of index sets* $(I_-^{0+},I_-^{+0},I_-^{00})$ *exists such that* $(I_-^{0+},I_-^{+0},I_-^{00})=(I_t^{0+}(x,\lambda),I_t^{+0}(x,\lambda),I_t^{00}(x,\lambda))$ *for every* $t\in[t_*-\varepsilon_-,t_*[$.
- *right non-Zeno relative to* (x,λ) *if a scalar* $\varepsilon_+>0$ *and a triple of index sets* $(I_+^{0+},I_+^{+0},I_+^{00})$ *exists such that* $(I_+^{0+},I_+^{+0},I_+^{00})=(I_t^{0+}(x,\lambda),I_t^{+0}(x,\lambda),I_t^{00}(x,\lambda))$ *for every* $t\in]t_*,t_*+\varepsilon_+]$.

When x^* *is left and right non-Zeno, then we say that* x^* *is non-Zeno.*

The Zeno phenomenon is illustrated in Figure 1.1.

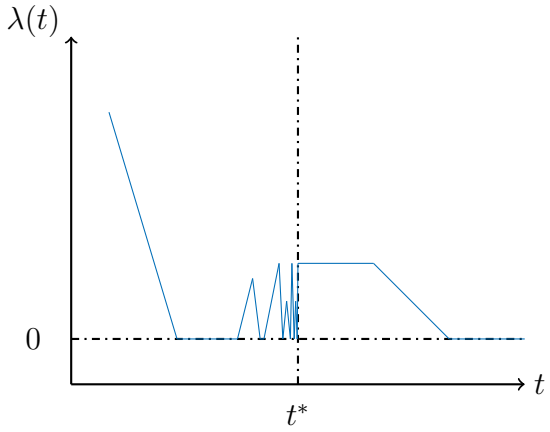


Figure 1.1: Illustration of λ when the solution is left Zeno and right non-Zeno

Theorem 1.2.4. *[97] If* D *is a* $\mathbf{P}$*-matrix, then all states of the LCS*(A,B,C,D) (1.1) *must be non-Zeno.*

Similar results but with assumptions more or less restrictive than $D\in\mathbf{P}$ exist: the interested reader is referred to [33, 97].

1.2.5 Dependence to initial conditions

Consider the Boundary Value Problem (BVP):

$$\begin{aligned} \dot{x} &= Ax + B\lambda \\ 0 \leq \lambda \perp Cx + D\lambda &\geq 0 \\ Mx(0) + Nx(T) &= b \end{aligned} \tag{1.14}$$

where the system is described the same way as in (1.1) but a boundary condition is added (instead of an initial value), defined with matrices $M, N \in \mathbb{R}^{2n\times n}$ and $b \in \mathbb{R}^{2n}$. In the smooth cases, solving BVP uses a technique called *shooting*, which needs a sensitivity in the variation of the trajectory given changes in the initial condition. This sensitivity is usually given by a Jacobian matrix, which is defined though a differential equation involving the derivative of the right-hand side function in the dynamical system. In the case of (1.14), since λ is not a differentiable function of x, this property does not hold. However, Pang and Stewart [85] derived a result giving a sensitivity matrix (through a weaken sense of derivation), which leads to a non-smooth Newton method.

Definition 1.2.6.

- *A set-valued map $F : X \rightrightarrows Y$, where X and Y are normed space, is said upper semi-continuous at $x_0 \in X$ if for all open set M containing $F(x_0)$, there is a neighbourhood Ω of x_0 such that $F(\Omega) \subset M$. F is said upper semi-continuous if it is so at every point $x_0 \in X$.*

- *Let $\Phi : \mathcal{D} \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be a vector function defined on the open set $\mathcal{D}$. The function Φ is said to have a linear Newton approximation at $\bar{x} \in \mathcal{D}$ if for an open neighborhood of $\bar{x}$, a set-valued map $\mathcal{T} : \mathcal{N} \rightrightarrows \mathbb{R}^{n\times n}$ exists such that:*

  (a) $\mathcal{T}(x)$ is a non empty compact subset of $n \times n$ matrices for every $x \in \mathcal{N}$,

  (b) $\mathcal{T}$ is a upper semicontinuous at $\bar{x}$,

  (c) there exists a scalar function $\Delta : \mathbb{R}_+ \to \mathbb{R}_+$ satisfying $\lim\limits_{s\downarrow 0} \Delta(s) = 0$ such that for all $x \in \mathcal{N}$ and all $E \in \mathcal{T}(x)$,

$$\|\Phi(x) - \Phi(\bar{x}) - E(x - \bar{x})\| \leq \|x - \bar{x}\|\Delta(\|x - \bar{x}\|).$$

  If (c) is strengthened to

  (c') a constant $c > 0$ exists such that for all $x \in \mathcal{N}$ and all $E \in \mathcal{T}(x)$,

$$\|\Phi(x) - \Phi(\bar{x}) - E(x - \bar{x})\| \leq c\|x - \bar{x}\|^2,$$

  then we say that the linear Newton approximation $\mathcal{T}$ is strong.

Linear Newton approximations generalize the concept of Jacobian matrix for a broad class of functions (broader than $\mathscr{C}^1$ functions), and allow us to design a nonsmooth Newton method in order to solve the equation $\Phi(x) = 0$ with Φ non differentiable.
Let us use this definition for the solution map of (1.14). Define $\mathcal{L}(z)$ as the family of index sets consisting of all α satisfying:

1. there exists a solution λ to the LCP(Cz, D) such that $\{i : \lambda_i > 0\} \subseteq \alpha \subseteq \mathcal{J} = \{i : (Cz + D\lambda)_i = 0\}$,

2. the columns of the matrix $D_{\mathcal{J}\alpha}$ are linearly independent.

By convention, we let $\mathcal{L}(z)$ consists of only the empty set if no such index set α exists.

Theorem 1.2.5. *[85] Suppose B SOL(Cx, D) is a singleton for all $x \in \mathbb{R}^n$. For $\xi \in \mathbb{R}^n$, denote by $x(\cdot, \xi)$ a solution of the IVP (1.1) with initial condition $x(0, \xi) = \xi$.*

- *Define the family*

$$\mathcal{T}_h(x) = \{-B_{\bullet\alpha}[(D_{\mathcal{J}\alpha})^\intercal D_{\mathcal{J}\alpha}]^{-1}(D_{\mathcal{J}\alpha})^\intercal C_{\mathcal{J}\bullet} : \alpha \in \mathcal{L}(x)\}.$$

  If $\mathcal{L}(x) = \{\emptyset\}$, then $\mathcal{T}_h(x)$ consists of only the zero matrix. Then $\mathcal{T}_h$ provides a linear Newton approximation for the piecewise linear function $x \mapsto B$ SOL(Cx, D).

- *Define $\Phi_b(\xi) = M\xi + Nx(t_1, \xi) - b$. Let us consider the Differential Inclusion (DI) in matrix:*

$$\dot{Y}(t) \in AY(t) + (\text{conv}\mathcal{T}_h(x(t, \xi)))\, Y(t), \;\; Y(0) = \mathbb{I}, \tag{1.15}$$

  where $\mathbb{I}$ is the identity matrix. Then $\mathcal{T}_\Phi(\xi) = \{M + NY(t_1) \;:\; Y(\cdot)$ solves the DI (1.15)$\}$ is a linear Newton approximation of Φ_b.

- *Suppose the BVP (1.14) has a solution $x(\cdot; \xi^*)$. If all matrices $M + NY^*(t_1)$ are nonsingular, where Y^* is a solution of the Differential Inclusion (1.15) with $\xi = \xi^*$, then there exists a neighbourhood $\mathcal{N}_*$ of ξ^* such that for any initial iterate ξ^0 chosen from $\mathcal{N}_*$, the sequence $\{\xi^k\}$ where*

$$\xi^{k+1} = (\tilde{M} + \tilde{N}Y^k(t_1))^{-1}(\tilde{x}_b + \tilde{N}(Y^k(t_1)\xi^k - x(t_1, \xi^k))), \;\; \forall k \geq 0$$

  where Y^k is a solution of the DI (1.15) with $\xi = \xi^k$, is well defined, remains in $\mathcal{N}_*$ and converges Q-superlinearly to ξ^*, meaning:

$$\lim_{k\to\infty} \frac{\|\xi^{k+1} - \xi^*\|}{\|\xi^k - \xi^*\|} = 0.$$

However, the assumption that $x \mapsto B$ SOL(Cx, D) is single-valued for all z is binding. Indeed, as it was noted in [84, Proposition 5.1], the single-valuedness of $x \mapsto B$ SOL(Cx, D) on $\mathbb{R}^n$ is a necessary and sufficient condition for the initial-value LCS (1.1) to have a unique $\mathscr{C}^1$ trajectory $x(\cdot, x_0)$ for all initial conditions $x_0 \in \mathbb{R}^n$.

1.3 Numerical simulation of LCS

1.3.1 Event-driven method

Following the Hybrid representation presented in Section 1.2.1, the event-driven schemes are based on the separation between smooth dynamics and switches when a change of mode occurs. Broadly speaking, the iterations are done as follows:

1. Using a numerical method, the DAE (1.4) is solved between t_k and t_{k+1}.
2. We check the constraints (1.5). If one of the constraints is violated (a switch of mode occurs):
   (a) Search for a time $t^* \in [t_k, t_{k+1}]$ such that: $\exists i \in I$; $\lambda_i(t^*) = 0$ or $\exists i \in I^c$; $y_i(t^*) = 0$, using a Newton method (for smooth cases) or a dichotomy.

(b) Compute the reinitialized state $x(t^*)$ using a jump transition map in (1.6), and all the other variables $y(t^*), \lambda(t^*)$.

3. Set t_k to t_{k+1} if there was no switch, t^* if there was one.

The advantage of such method is that we can achieve precise numerical solutions, by choosing a high order integration method in step 1 (such as Runge-Kutta methods of high order to cite only the most known ones) and a high order interpolation method for step 2(a). However, the event-driven method suffers some problems:

- No accumulation of events is allowed for this scheme: the switching events must be *well-separated* in time, otherwise step 2 is intractable.
- The detection of events may turn out to be tricky. Since with a floating point arithmetic, we can not solve exactly in (1.5) $\lambda_i = 0$ or $y_i = 0$, some threshold will be tuned instead. Step 2(a) then becomes: searching for a time t^* such that $\lambda_i(t^*) < \tau_\lambda$ or $y_i(t^*) < \tau_y$, for some positive scalars τ_λ, τ_y. It is somehow hard, if not impossible, to have threshold values τ_λ and τ_y that will work for any system, especially when m is large and many thresholds have to be tuned.

The convergence of such method depends obviously on the order of the interpolation and the integration methods, as it is proved in [65] for a mechanical system with one degree-of-freedom without accumulations of impacts. Some results concerning the event-driven method (in the more general context of dynamical systems with discontinuous right-hand side) may also be found in [99].

1.3.2 Time-stepping method

A simple approach to compute an approximated trajectory of LCS(A, B, C, D) consists in discretizing the system using a backward Euler method: given $h > 0$, define $N = \left\lceil \frac{t_1 - t_0}{h} \right\rceil$ and solve, for all $k \in \overline{N}$:

$$\frac{x_{k+1} - x_k}{h} = Ax_{k+1} + B\lambda_{k+1} \tag{1.16a}$$

$$0 \leq \lambda_{k+1} \perp Cx_{k+1} + D\lambda_{k+1} \geq 0 \tag{1.16b}$$

In fact, isolating x_{k+1} in (1.16a) (under the hypothesis that $I - hA$ is invertible), the only problem to solve is, for each k:

$$0 \leq \lambda_{k+1} \perp C(I - hA)^{-1}x_k + [D + hC(I - hA)^{-1}B]\lambda_{k+1} \geq 0$$

In order to state a convergence result, we first give some definitions:

Definition 1.3.1.

- *The matrix triple (A, B, C) is said to be minimal if (A, B) is controllable and (C, A) is observable.*
- *For fixed $h > 0$, we say that the functions (x^h, λ^h) are generated by (1.16) if they are piecewise constants, with pieces $\{x_k, \lambda_k\}$ solution of (1.16) and $(x^h(t), \lambda^h(t)) = (x_k, \lambda_k)$, $(k-1)h \leq t < kh$.*

The passivity assumption was already met in a related form in Assumption 1.2.1, where $R^2B = C^\intercal$ is a passive input/output constraint.

Theorem 1.3.1. *[27, Theorem 13] Consider LCS(A, B, C, D) such that (A, B, C) is a minimal representation, B is full column rank and LCS(A, B, C, D) is passive (see Definition 1.2.2). Let $x_0 \in \mathbb{R}^n$ be given. Then, there exists a unique solution (x, λ) on $[t_0, t_1]$ with initial state x_0 (in the L^2 sense presented in Section 1.2.3). Also, let (x^h, λ^h) be the piecewise functions generated by (1.16). Then λ^h converges weakly in L^2 to λ, and x^h converges (strongly) in L^2 to x, as h tends to 0.*

The passivity assumption is rather important (even though it is not proved that it is necessary). A simple example showing no convergence of the time-stepping method is given in [27] by the triple integrator:

$$\dot{x}_1 = x_2,\ \dot{x}_2 = x_3,\ \dot{x}_3 = \lambda$$
$$0 \leq \lambda \perp x_1 \geq 0$$

on $[0, 1]$, with initial state $x(0) = (0\ -1\ 0)^\intercal$. Assume we approximate the solution of this system with a time-stepping method (which generates a sequence $\{x_k, \lambda_k\}_{k=0}^N$), and build with it a function $\tilde{x}$ which is piecewise constant:

$$\tilde{x}(t) = x_k,\ (k-1)h \leq t < kh$$

We can prove that $\{x_{1k}\}$ is defined by

$$x_{1k} = \begin{cases} 0 & \text{if } k = 0 \\ \frac{k(k+1)}{2}h & \text{otherwise} \end{cases}$$

As such, one sees easily that:

$$\|\tilde{x}_1\|_{L^1([0,1],\mathbb{R})} = \sum_{k=1}^{N} \int_{(k-1)h}^{kh} \frac{k(k+1)}{2} h dt = \frac{h^2}{6} N(N+1)(N+2) = O\left(h^{-1}\right)$$

So $\tilde{x}$ is not bounded in L^1 norm (and therefore for any L^p, $p \geq 1$) as h tends to 0. It should be noted that a scheme overcoming this problem is presented in [3].

Conclusion

The study of LCS, including the control of such systems, is now well developed, even though there are still some unanswered questions. Once the control of a system is studied, it seems logical to turn our interest on how to choose the good control in order to respect some criteria, like minimizing a functional depending on the state and the control. This is the subject of the next chapter.

Chapter 2

Non-smooth Optimal Control Problems

Abstract. This chapter focuses on the optimal control of various types of dynamical systems on two different aspects. First of all, the necessary conditions of optimality concerning different systems are considered, including hybrid automata, differential inclusions, Lipschitz systems and mixed constraints. Along these lines, the presentation will show the limitations of these results when they are applied to LCS. Secondly, two methods of numerical approximations of the optimal solutions are presented, called respectively the Direct and Indirect methods.

Let us now focus on the analysis of Optimal Control Problems. These problems, which find their origin in the Calculus of Variation, have many applications in a wide variety of topics. Mathematically speaking, these problems are formulated as the optimization of a functional over functions satisfying a differential equation (or a differential inclusion), alongside with constraints over the whole domain of integration and/or at the boundaries. We focus here only on Optimal Control Problems of Ordinary Differential Equations of the form:

$$\begin{aligned} \min \quad & \int_0^T g^{int}(t, x(t), u(t))dt + g^b(x(0), x(T)) \\ \text{s.t.} \quad & \begin{cases} \dot{x}(t) = f(t, x(t), u(t)) \\ g^c(t, x(t), u(t)) \in \mathcal{E}(t) \ \text{ a.e. on } [0, T], \\ (x(0), x(T)) \in \mathcal{B} \end{cases} \end{aligned} \tag{2.1}$$

where $x : [0, T] \to \mathbb{R}^n$ is called the state, $u : [0, T] \to \mathbb{R}^m$ is called the control, $g^{int} : [0, T] \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ called the running cost, $g^b : \mathbb{R}^{2n} \to \mathbb{R}$ called the final cost, $f : [0, T] \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, $g^c : [0, T] \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^k$, $\mathcal{E} : [0, T] \rightrightarrows \mathbb{R}^k$, $\mathcal{B} \subseteq \mathbb{R}^{2n}$. Further hypothesis can be made concerning smoothness of functions appearing in (2.1), making the analysis of this problem more or less easy. Several names are given to this problem whether the running cost g^{int} and the final cost g^b are present or not. We call it a Mayer problem if $g^{int} \equiv 0$, a Lagrange problem if $g^b \equiv 0$, and a Mayer problem if none of them is identically 0.
By analogy with what is made for finite-dimensional optimization problems, the analysis of (2.1) often focuses on two different topics: existence of an optimal solution, and searching for necessary conditions that the optimal solution should comply with. For Optimal Control Problems, these conditions take the form of the Pontryagin equations, defining an adjoint state through a Differential Algebraic Equation. These first order necessary conditions were first proved by Pontryagin et al in their seminal work (see [88]). Their work focused on optimal control with smooth data, and only constraints in the control were allowed. We state here their version for illustration:

Theorem 2.0.1 (Pontryagin's Maximum Principle)**.** *[104]*
Consider the Optimal Control problem:

$$\min \int_0^T g^{int}(t,x(t),u(t))dt + g^b(x(T)) \\ s.t. \begin{cases} \dot{x}(t) = f(t,x(t),u(t)) & \\ u(t) \in \mathcal{E} & a.e.\ on\ [0,T], \\ x(0) \in \mathcal{B}_0, x(T) \in \mathcal{B}_1 & \end{cases} \tag{2.2}$$

If (x^*,u^*) is optimal for this problem, then it is the projection of an extremal $(x(\cdot),p(\cdot),p^0,u(\cdot))$, where $p^0 \leq 0$ and $p(\cdot):[0,T]\to\mathbb{R}^n$ is an absolutely continuous function called the adjoint state, with $(p(\cdot),p^0)\neq(0,0)$, such that (x^*,p) solve the Euler equations:

$$\dot{x}(t) = \frac{\partial H}{\partial p}(t,x^*(t),p(t),p^0,u^*(t)),\ \dot{p}(t) = -\frac{\partial H}{\partial x}(t,x^*(t),p(t),p^0,u^*(t))$$

almost everywhere in $[0,T]$, where $H(t,x,p,p^0,u) = \langle p, f(t,x,u)\rangle + p^0 g^{int}(t,x,u)$ is the Hamiltonian, and there holds:

$$H(t,x^*(t),p(t),p^0,u^*(t)) = \max_{v\in\mathcal{E}} H(t,x^*(t),p(t),p^0,v)$$

Additionally, if $\mathcal{B}_0$ and $\mathcal{B}_1$ are submanifolds of $\mathbb{R}^n$, then the adjoint vector can be built in order to satisfy the transversality conditions at both extremities (or just one of them):

$$p(0) \perp T_{x^*(0)}\mathcal{E}_0,\ p(T) - p^0\frac{\partial g}{\partial x}g(x^*(T)) \perp T_{x^*(T)}\mathcal{E}_1$$

where T_xM denotes the tangent space to M at point x.

Subsequent research focused on searching first-order conditions which resembles much this Theorem, and also kept the same terminology (Hamiltonian, adjoint state, Pontryagin Maximum Principle,...).
These equations are defined even for systems with weak smoothness (for instance if g^{int} is only Lipschitz), but one has to use tools of non-smooth analysis in order to tackle such problems. Even though the problem of Optimal Control of LCS may seem smooth, the complementarity conditions makes the use of non-smooth analysis necessary. After going over some definitions, some results concerning Optimal Control of various systems linked with LCS will be presented, along with their limits for the application for LCS. Then, the last Section will focus on numerical approximations of the optimal solutions.

2.1 Optimal control of non-smooth systems

2.1.1 Optimal control of hybrid automata

Before defining the optimal control problem for hybrid automata, we must define a hybrid control system.

Definition 2.1.1. *A hybrid control system is given by a 7-tuple $(Q,\Sigma,J,G,M,U,\mathcal{U})$, where:*

- (Q, Σ, J, G) *is a hybrid automaton as defined in Definition 1.2.1 (where this time,* Σ *is defined through ODEs of the form* $\dot{x} = f_q(t, x, u)$*).*
- $M = \{M_q\}_{q \in Q}$ *is a collection of smooth manifolds where the state can have values for each mode.*
- $U = \{U_q\}_{q \in Q}$ *is a collection of sets defining acceptable values for the control for each mode.*
- $\mathcal{U} = \{\mathcal{U}_q\}_{q \in Q}$ *is a collection of functional sets admissible for controls:* $u : Dom(u) \to U_q$*.*

Optimal control of such systems have been widely analyzed: see for instance [11, 29, 49, 55, 86, 87, 101, 102, 105, 106]. The functional to minimize may take into account a running cost partitioned over the different modes followed by the trajectory, along with a cost on the state at switching times; the switching times may also be part of the optimal solution searched. As it was underlined in Section 1.2.1, this approach has a lot of drawbacks for LCS:

- First of all, describing the LCS as a switching system is not an easy task, and the resulting system may become too big to apply properly these results (for instance, because of the 2^m different modes possible);
- Secondly, most results concerning the Optimal Control of Switching Systems only assume that the dynamical systems are described with ODEs for each mode. However, in most cases of LCS, the dynamical system will be described with DAEs. Furthermore, adding further constraints involving at the same time the state and the control does not seem to be taken into account in the above articles;
- The resulting necessary conditions make strong assumptions on the system, on the admissible transitions at switching times, or on the number of switches on the interval of integration. Therefore, the results become intractable for a high number of switches or modes.

For all these reasons, the hybrid approach will not be used in order to tackle the Optimal Control problem for LCS.

2.1.2 Differential inclusions and the sweeping process

The Optimal Control Problem for Differential Inclusions also attracted a lot of interest. We summarize here some results concerning this.

Lipschitz bounded differential inclusions

In [98], the problem of optimality of solution is tackled for system of the form $\dot{x} \in F(x)$ for some multifunction $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ complying with some assumptions. Let $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a upper semi-continuous multifunction (see Definition 1.2.6) with compact convex values. The problem is then to solve:

$$\begin{aligned} &\min \ \phi(x(T)) \\ &\text{s.t.} \ \begin{cases} \dot{x} \in F(x) \\ x(0) \in C_0, \ x(T) \in C_1 \end{cases} \end{aligned} \tag{2.3}$$

over absolutely continuous functions $x : [0, T] \to \mathbb{R}^n$, where C_0, C_1 are two closed subsets of $\mathbb{R}^n$, and $\phi : \mathbb{R}^n \to \mathbb{R}$ is a Lipschitz function. One could see this problem as a generalization of optimal

control of systems defined by the ODE $\dot{x} = f(x, u)$, $u \in U$ (by defining $F(x) = f(x, U)$). For this problem, one can prove two results: one concerning existence of an optimal solution, one giving necessary condition this optimal solution should comply with [98, 109]. These conditions are close to the PMP.

However, since the multifunction F is supposed to have compact values, we can not apply it directly to the controlled LCS (1.11). Indeed, except for some peculiar cases, the normal cone, seen as a multifunction, may admit unbounded values; for instance, $\mathcal{N}_{\mathbb{R}^+}(0) = -\mathbb{R}^+$.

Sweeping process

The optimal control of sweeping processes $-\dot{x} \in \mathcal{N}_{C(t)}(x(t))$ (see Section 1.2.3) has been a topic of research gathering a lot of attention in the recent past years. Of course, as it was mentioned in Section 1.2.3, a Cauchy problem described with a sweeping process may admit a unique solution; there is, in these cases, no hope to optimize any trajectory, unlike in many cases of systems defined with a differential inclusion. In the literature, three different approaches attempt to add control in the sweeping process:

1. Adding an additive perturbation in the differential inclusion, taking the form

$$-\dot{x} \in \mathcal{N}_{C(t)}(x(t)) + f(t, x(t), a(t))$$

for some function $a : [0, T] \to \mathbb{R}^m$ and $f : [0, T] \times \mathbb{R}^{n+m} \to \mathbb{R}^n$, and where $C(\cdot)$ is unchanged. The research in this direction mainly focuses on existence and relaxation; see [5, 6, 44, 77] and references therein.

2. The second approach couples a sweeping process and a controlled ordinary differential equation, where $C(t)$ is a given multimap but uncontrolled. The results provide necessary optimality conditions [4, 23].

3. The third approach adds control in the definition of the moving set. For instance, [43] focuses on the case of moving controlled polyhedra:

$$C(t) = \{x \in \mathbb{R}^n \mid \langle u_i(t), x\rangle \leq b_i(t),\ i = 1, ..., m\}$$

where the control functions are u_i and b_i, $i = 1, ..., m$. The results focus on necessary optimality conditions: see [42, 43]. As noted in [28], the discrete approximation approach to optimization of the differential inclusions $\dot{x} \in F(x)$ (see [79]) heavily relies on a Lipschitz property of the multifunction F. However, this assumption is not verified with such controlled polyhedra, and therefore the approach can not be applied.

The closest result to the optimal control problem of LCS may be found in [28], where the optimal control problem reads as:

$$\min\ \phi(x(T)) + \int_0^T \ell(t, x(t), u(t), a(t), \dot{x}(t), \dot{u}(t), \dot{a}(t))dt$$
$$\text{s.t.} \begin{cases} -\dot{x}(t) \in \mathcal{N}_{C(t)}(x(t)) + f(x(t), a(t)) \\ C(t) = C + u(t),\ C = \{x \in \mathbb{R}^n \mid Gx \geq 0\} \\ \|u(t)\| = r \end{cases}$$

for $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$, $\ell : [0,T] \times \mathbb{R}^{4n+2d} \to \overline{\mathbb{R}}$, a scalar $r > 0$ and a matrix $G \in \mathbb{R}^{m \times n}$, and where the controls are $a : [0,T] \to \mathbb{R}^d$ and $u : [0,T] \to \mathbb{R}^n$. If one supposes $f(x,a) = -Ax - Fa$, the system can be easily put in the form of a LCS. Notice that $C(t) = \{x \in \mathbb{R}^n | Gx + Gu(t) \geq 0\}$. Moreover, using [22, Proposition A.2], we prove that:

$$\partial \mathbb{I}_{C(t)}(x) = G^\intercal \partial \mathbb{I}_{\mathbb{R}^m_+}(Gx + Eu(t))$$

where $\mathbb{I}_K$ is the indicator function of the set K and ∂ denotes the convex subdifferential (see Section 1.2.2). Therefore, since $\partial \mathbb{I}_K(x) = \mathcal{N}_K(x)$, it yields: $-\dot{x} \in G^\intercal \mathcal{N}_{\mathbb{R}^m_+}(Gx + Gu) + Ax + Fa$. Finally, we use the following equivalence (presented Section 1.1.1):

$$-v \in \mathcal{N}_{\mathbb{R}^n_+}(Gx + Gu) \iff 0 \leq v \perp Gx + Gu \geq 0$$

such that we prove the equivalence between the sweeping process and the following LCS:

$$\begin{aligned} &\dot{x}(t) = Ax(t) + G^\intercal v(t) + Fa(t) \\ &0 \leq v(t) \perp Gx(t) + Gu(t) \geq 0 \\ &\|u(t)\| = r \end{aligned}$$

The main results of [28] show well-posedness of the system, convergence of the discrete approximations, and necessary optimality conditions. However, the results developed in [28] do not apply to our problem because of the presence of decoupled controls (u, a), acting either on the polyhedron or on the dynamical inclusion.

2.1.3 Lipschitz systems

As presented in Proposition 1.1.2, a certain class of LCS can be described with a piecewise linear Lipschitz differential equation, but not differentiable. Most results concerning optimality analysis of non constrained systems require some differentiability of the data, but this assumption has to be weakened. Clarke in [40] tackled the optimal control problem of a non-constrained system under minimal hypotheses; namely, the data have to be measurable and Lipschitz in the state variable. We state here more clearly his results, as they will be used in further development.

Definition 2.1.2. *Let $g : \mathbb{R}^n \to \mathbb{R}^k$ be a locally Lipschitz function. We define the Clarke generalized Jacobian $\partial^C g(s)$ of g at s as:*

$$\partial^C g(s) = \text{conv}(\lim(Dg(s_i))^\intercal)$$

where we consider all sequences $\{s_i\}$ converging to s where the usual Jacobian $Dg(s_i)$ exists, as well as the limit of the sequence $\{Dg(s_i)\}$. For $f : \mathbb{R}^n \times \mathbb{R}^\ell \to \mathbb{R}^k$, we denote by $\partial^C_s f(t,x)$ the Clarke generalized partial Jacobian, which is the generalized Jacobian of the function $s \mapsto f(t,s)$ at point x.

Consider the following optimal control problem:

$$\begin{aligned} \min \quad & \int_0^T f^0(t, x(t), u(t)) dt \\ \text{s.t.} \quad & \begin{cases} \dot{x}(t) = f(t, x(t), u(t) \\ u(t) \in U(t), \ x(t) \in X \\ x(0) \in C_0, \ x(T) \in C_1 \end{cases} \end{aligned} \tag{2.4}$$

where $T > 0$ is given, $f^0 : [0,T] \times \mathbb{R}^n \times \mathbb{R}$, $f : [0,T] \times \mathbb{R}^n \times \mathbb{R}^m$, $U : [0,T] \rightrightarrows \mathbb{R}^m$ is a multifunction, X, C_0 and C_1 are closed subsets of $\mathbb{R}^n$. $u : [0,T] \to \mathbb{R}^m$ satisfying $u(t) \in U(t)$ a.e. $t \in [0,T]$ is a Lesbesgue measurable function (the control). An admissible trajectory associated with u is an absolutely continuous function $x : [0,T] \to \mathbb{R}^n$ satisfying the constraints in (2.4) and $x(t) \in X$ a.e. on $[0,T]$. x is said interior if moreover, $x(t) \in \text{int } X$ a.e.

Define $\tilde{f} = (f, f_0)$. We do the following hypotheses:

H1 For each s in a neighborhood of X, the function $(t,u) \mapsto \tilde{f}(t,s,u)$ is $L \times B^m$-measurable[1].

H2 There is a function $k \in L^1([0,1],\mathbb{R})$ such that for $t \in [0,T]$, $u \in U(t)$ and s_1, s_2 in a neighbourhood of X:

$$\|\tilde{f}(t,s_1,u) - \tilde{f}(t,s_2,u)\| \leq k(t)\|s_1 - s_2\|$$

H3 The graph of U is $L \times B^m$-measurable.

Theorem 2.1.1. *[40] Let (x^*,u^*) be the optimal solution to (2.4), where x^* is an interior admissible trajectory associated with u^*. If* **H1** - **H3** *are satisfied, there exist an absolutely continuous function $p : [0,T] \to \mathbb{R}^n$ and a scalar $\lambda \leq 0$ such that $|p(t)| + |\lambda| \neq 0$, and:*

$$-\dot{p}(t) \in \partial_x^C f(t,x^*(t),u^*(t))p(t) + \lambda \partial_x^C f^0(t,x^*(t),u^*(t))$$

$$\langle p(t), f(t,x^*(t),u^*(t))\rangle + \lambda f^0(t,x^*(t),u^*(t)) = \sup\{\langle p(t), f(t,x(t),u(t))\rangle + \lambda f^0(t,x(t),u(t)),\ u \in U(t)\} \text{ a.e.,}$$

$p(0)$ is normal to C_0 at $x^*(0)$, and $-p(T)$ is normal to C_1 at $x^*(T)$.

As explained in [40, Remark 4], the state constraint set X may also depend on time, allowing to define X as an ε-neighbourhood of an optimal state trajectory.

2.1.4 Mixed constraints

In the classical Pontryagin Maximum Principle (PMP) (see [88]), the optimal control problem tackled only admitted pathwise constraint on the control function ($u(t) \in U(t)$ for all $t \in [0,T]$ and some multifunction U). Subsequently, these results were adapted in order to tackle a broader problem by adding some pathwise mixed control/state constraints:

$$\phi(t,x(t),u(t)) \in \Phi(t),\ \text{a.e. } t \in [0,T]$$

for a given function ϕ and a multifunction Φ. In order to state necessary conditions with these constraints, some *constraint qualification* has to be made on the data. We find in the literature two different approaches:

- The first one assumes a full-rank condition on the derivative of the data with respect to u (when the data are smooth enough). Some other qualification can be found, implying this full-rank condition (such as Mangasarian Fromovitz condition). The idea is then to analyze an augmented cost, where the integral $\int_0^T \phi(t,x(t),u(t))^\intercal \zeta(t)dt$ is added, along with

[1] $L \times B^m$ denotes the σ-algebra of subsets of $[0,T] \times \mathbb{R}^m$ generated by all products of a Lebesgue measurable subset of $[0,T]$ and a Borel subset of $\mathbb{R}^m$.

complementarity slackness conditions, where ζ is a measurable and bounded function, viewed as a Lagrange multiplier. Such results may be found in [47, 50, 76]. On a side note, Arutyunov and al. in [9] extended these previous results to impulsive systems, expressed in form of measure differential equations. The assumption they make on the mixed constraints enter in this same framework.

- The second approach, whose assumptions actually imply the first results, derives *stratified* necessary conditions. In this approach, a measurable *radius function* is defined in order to fix a neighbourhood around the optimal control, giving sense to what is a local optimum. More precisely, denote by $J(x,u) = \int_0^T g^{int}(t,x(t),u(t))dt + g^b(x(0),x(T))$ (the cost in (2.1)). One says that (x^*,u^*) is a local minimum of radius $R:[0,T]\to[0,+\infty]$ if there exists $\varepsilon>0$ such that, for all admissible functions (x,u) such that:

$$\|x^*(t)-x(t)\|\leq\varepsilon,\ \|u^*(t)-u(t)\|\leq R(t) \text{ a.e. on } [0,T],\ \int_0^T\|\dot{x}^*(t)-\dot{x}(t)\|dt\leq\varepsilon$$

  one has $J(x^*,u^*)\leq J(x,u)$. This radius function is now part of the definition of a local minimum, and the necessary conditions are expressed in its term. Of course, if $R(t)=+\infty$ for almost all t on $[0,T]$, then one retrieves the classical $W^{1,1}$ minimizers (see [109]). This approach is typified by [37, 38].

In both cases, the necessary conditions take the same form as the PMP, meaning the existence of an absolutely continuous function (the adjoint state) defined by a differential equation (or inclusion), maximization of a Hamiltonian and transversality conditions. One could argue that "pure" state constraints may be added also in this framework. The previous results rest upon constraint qualification that prevents most constraints of the form $\phi(t,x(t))\in\Phi(t)$. An other argument to support this observation is the study of optimal control with unilateral state constraints ($\phi(t,x(t))\leq 0$). In this context, the necessary conditions involve a discontinuous adjoint state defined by means of a Radon measure supported on intervals of active constraints. More details can be found in [13] and references therein.

2.2 Numerical approximations

Obviously, most optimal control problems can not be solved analytically. Therefore, most problems are approximated by numerical schemes. In order to solve these problems, two different methods clearly stand out: the direct methods and the indirect methods.

2.2.1 Direct approach

In the direct approach, one discretizes directly the optimal control problem in order to solve a finite dimensional optimization problem. Once again, many discretization choices can be made. For ease of presentation, we focus on the Mayer problem (meaning $g^{int}\equiv 0$ in (2.1)). We identify three approaches:

- the single shooting, which consists in discretizing only the control. The choice of discretization consists in choosing a finite-dimensional basis in which the control is represented

(piecewise constant, piecewise affine, splines, etc). The problem then reduces to the optimal choices of pieces $u_0, ..., u_M$ for some fixed positive integer M:

$$\min_{u_0,...,u_M} g^b(x(0), x(T))$$
$$\text{s.t.} \begin{cases} \dot{x}(t) = f(t, x(t), u(t)) \\ g^c(t, x(t), u(t)) \in \mathcal{E}(t) & \text{a.e. on } [0, T] \\ (x(0), x(T)) \in \mathcal{B} \\ u(t) = \mathcal{U}(t, u_i, u_{i-1}) & \forall t \in [t_{i-1}, t_i],\ \forall i \in \overline{M}, \end{cases}$$

where $t_0 = 0$ and $t_M = T$. The stage times t_i can also be part of the decision variables used for optimization (especially if the final time T is free). The path constraint $g^c(t, x(t), u(t)) \in \mathcal{E}(t)$ has to be also re-expressed: either it is satisfied only at stage points t_i, or it can also be reformulated as an integral constraint. The differential equation is numerically integrated with any ODE solver, but it should also provide sensitivities, since they are needed by the optimization solver.
The strong points of the single shooting method is that it creates a relatively small optimization problem, with little degrees of freedom, which can use fully adaptive ODE or DAE solvers. Furthermore, the ODE will be solved at each iteration of the solving algorithm (it is a feasible path method). However, any knowledge on the trajectory can not be used for initialization, unstable systems seem to be difficult to treat, the method becomes computationally demanding, and the sensitivities for the state trajectory x may depend non trivially on the choices of the pieces u_i.

- the simultaneous method, which consists in a full discretization of the control and of the state. In this case, the ODE $\dot{x}(t) = f(t, x(t), u(t))$ is replaced by equalities $f^h(t_i, x_i, x_{i-1}, u_i, u_{i-1}) = 0$, for $i \in \overline{N}$ for some positive integer N. The problem then becomes:

$$\min_{u_0,...,u_N,x_0,...,x_N} g^b(x_0, x_N)$$
$$\text{s.t.} \begin{cases} f^h(t_i, x_i, x_{i-1}, u_i, u_{i-1}) = 0 \\ g^c(t_i, x_i, u_i) \in \mathcal{E}(t_i) & \forall i \in \overline{N} \\ (x_0, x_N) \in \mathcal{B}. \end{cases}$$

  The advantages of the simultaneous method are that we can use an a priori initialization on the state trajectory x, the local convergence is fast, it can treat unstable systems, and it can easily handle state or terminal constraints. The disadvantages are a very large NLP, and the numbers of time stages that must be chosen a priori. Aside these observations, note also that the dynamical equation will only be satisfied at the converged solution of the optimization process. This infeasible path method is a drawback if the algorithm is stopped too soon, but it has the advantage of saving computational efforts and allows to deal with unstable systems.

- the multiple shooting, that combines the spirit of the two previous methods. We present the method for controls discretized with constants pieces. Denote by $x(\cdot; s_i, u_i)$ the solution of:

$$\dot{x}(t) = f(t, x(t), u_i) \text{ a.e. on } [t_i, t_{i+1}],\ x(t_i) = s_i$$

These s_i, each approximating the state x at time t_i, become a new decision variable (as in the simultaneous method). Of course, some continuity property of the state must be preserved, which is enforced by imposing $s_{i+1} = x(t_{i+1}; s_i, u_i)$. The problem eventually reads as:

$$\min_{u_0,\dots,u_N,s_0,\dots,s_N} g^b(s_0, s_N)$$
$$\text{s.t.} \begin{cases} s_{i+1} - x(t_{i+1}; s_i, u_i) = 0 \\ g^c(t_i, s_i, u_i) \in \mathcal{E}(t_i) \qquad \forall i \in \overline{N} \\ (s_0, s_N) \in \mathcal{B}. \end{cases}$$

This method mostly has the same advantages as the simultaneous method (a priori initialization of the state, deals robustly with unstable systems, fast local convergence). Furthermore, the integration can easily be parallelized. However, the trajectory will be valid on the whole interval $[0, T]$ only once the optimization algorithm fully converged (in order to assume continuity of the trajectory). And again, computing an optimal descent direction will use sensitivities, normally given by the ODE solver.

Note that there exist also some probabilistic approaches, in which the optimal control is expressed as searching an occupation measure. This measure is approximated by a finite number of its moment in order to reduce it to a finite dimensional optimization problem; see [70].

These finite dimensional optimization problems can then be solved by different algorithms, such as Sequential Quadratic Programming (SQP) or interior-point methods. The direct methods are often used because of their simple construction and their modularity to change of model (for adding constraints, for instance). Also, no a priori analysis of the problem is needed, and they are not so sensitive to the choice of the initial condition. However, reaching high precision becomes really difficult with this problem. High precision is met for instance by shrinking the time-step, which causes a rise of variables in the optimization problem. Various variations entering the framework of direct methods can be found in [25, 48, 53, 57, 89, 93, 104]. We make a special focus on [12], where the author emphasises the importance of sparsity for the resolution of the optimization problems.

2.2.2 Indirect approach

The indirect approach relies on the Pontryagin's equations in its Maximum Principle in order to compute the solution. Let us present the method in the simpler context given in Theorem 2.0.1 (namely, smooth system with only constraint on the control). Suppose we can make explicit the maximization condition (leading to an expression of u^* as a function of x^* and p). Then the Euler equations can be put in the form $\dot{z}(t) = F(t, z(t))$, where $z(t) = (x^*(t), p(t))$. Moreover, initial, final and transversality conditions can be put in the form $R(z(0), z(T)) = 0$. This forms a Boundary Value Problem (BVP).
Denote by $z(t; z_0)$ the solution of the Cauchy problem $\dot{z}(t) = F(t, z(t))$, $z(0) = z_0$. The BVP problem is reduced to a root-search of the function $G(z_0) = R(z_0, z(T; z_0))$.

This method, called the *simple shooting indirect method*, is known to suffer from instability. In order to improve stability, the interval $[0, T]$ is divided in subinterval $[t_i, t_{i+1}]$, where $0 = t_0 < t_1 < .... < t_N = T$ are called the shooting nodes, and considering as unknown not only $z(0)$ but all $z(t_i)$ for i between 0 and N. Then on each subinterval $[t_i, t_{i+1}]$, the Cauchy problem $\dot{z}(t) = F(t, z(t))$, $z(t_i) = z_i$ is solved. Assuring also continuity of the solution leads to add further constraints $z_{i+1} - z(t_{i+1}; z_i) = 0$ into the function R. All is reduced eventually to a root-search problem, with with more variables (namely, z_i, $i \in \{0, ..., N\}$). This is called the *multiple shooting*

indirect method. This approach can be generalized to other problems, for instance with pure state constraints (which add jump in the adjoint state).

There exist many different root-search algorithms, such as the Newton methods or the BFGS method. These methods use the Jacobian (or an approximation) of the function G in order to find a descent direction. Augmenting the number of subintervals improves the stability, as the determinant of this Jacobian, seen as a function of $t_{i+1} - t_i$, is an exponentially growing function (see [100]).

The indirect methods are advantageous for obtaining a precise solution, thanks to the root-finding algorithms that may present fast convergence. Also, shrinking the time-step used for integrating the Euler equations on each subinterval will just increase the number of iterations but not the complexity of the integration. Finally, the multiple shooting can be easily parallelized (each Cauchy problem can be solved in parallel on each subinterval $[t_i, t_{i+1}]$). The drawbacks of such methods are their complexity, both analytically (deriving the necessary conditions may turn out to be difficult in presence of pure state constraints), and numerically. The root-finding algorithms are hard to make convergent unless the initial guess is already in a neighbourhood of the solution. Also, the optimal control is computed as an open-loop controller; the PMP is only a necessary condition, thus the obtained solution needs to be checked a posteriori; and the structure of the switches (in case of bang-bang control for instance) should be known a priori. More details on these methods can be found in [14, 15, 16, 24, 104] and references therein.

2.2.3 Hybrid numerical approach

There is some duality between the direct and indirect approaches. We could sketch the idea as follows: the direct approach consists in 1) discretize and 2) dualize, while the indirect approach consists in 1) dualize and 2) discretize. The hybrid numerical approach consists in applying both the direct and indirect methods in order to achieve a high precision solution. Broadly speaking, it consists in applying first a direct method in order to get a rough solution, and then solving the Pontryagin equations using as first guess the solution given by the direct method. Of course, the adjoint state and multipliers, appearing in the first order conditions, must be computed from the rough solution. Since the solution of the direct method should already be in a neighbourhood of the optimal solution, there is high hope that the indirect method would converge. There are two reasons for applying such scheme. First, the necessary conditions assure that the solution found actually is a stationary solution. Secondly, the first rough solution can be recomputed using a thiner grid while solving the Pontryagin equations. Since it is not an optimization problem anymore, decreasing the time-step will just increase the number of iterations made to integrate the equation, but not the complexity of each of them.

We could wonder if passing from the direct method to the indirect method is valid, in the sense that there exists a mapping such that, when applied to the solution given by the direct method, it can be used as a good approximation of the solution of the Pontryagin equation, and vice versa. Stated in an other way, the question is to know if the dualization and discretization steps for the direct and indirect approaches commute or not. Some examples show that it is not the case. Even under classical assumptions of consistency and stability of the discretization scheme used, the indirect method converges while the direct method diverges (see simple examples in [60]). Nonetheless, some results concerning this commutation theory have been derived, under the name of *Covector Mapping Principle*. Denote by B the original optimal control problem (2.1), by B^N its direct approach discretization, and by $B^{N,\lambda}$ the stationary conditions associated with the optimization problem in B^N. On the other hand, denote by B^λ the stationary conditions associated

with B, and by $B^{\lambda,N}$ the discretization of B^{λ}. This is schematized in Figure 2.1.

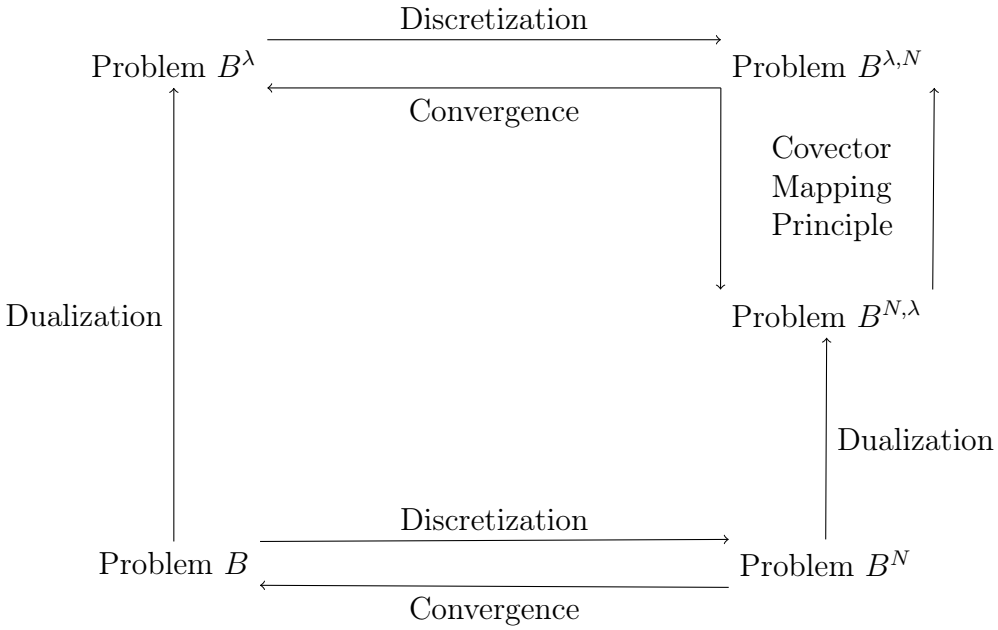


Figure 2.1: Sketch of the different numerical methods for Optimal Control problems (inspired by [57]).

Obtaining results ensuring commutation is not obvious. The class of Legendre pseudospectral methods is up to now the class of methods with the richest results. These are typified by [53, 92]. Ross and al in [93] show that the pseudospectral discretization leads to a Covector Mapping Theorem, under so-called *closure condition*. Gong and al in [57] show more precisely the connection between the covector mapping theorem and convergence of the different methods using pseudospectral methods. Aside pseudospectral methods, Hager in [60] assert convergence of the direct method provided that the method be based on Runge-Kutta schemes with positive coefficients. Dontchev and Hager in [51] show convergence of an Euler approximation scheme for an optimal control problem with pure state constraints. Other results for RK schemes can be found in [35, 52, 94] and references therein. In [69], the authors prove convergence of a scheme for an LQ problem (the state is approximated by a piecewise linear function, and the control by a piecewise constant function).
However, all these results rely on a certain degree of "smoothness" of the data and/or of the optimal solution. Otherwise, it seems that there is no systematic method for designing approximation leading for sure to convergence of the numerical methods.

Conclusion

The study of optimal control problems is already broad and undertakes a lot of different systems. However, it can not tackle yet the formulations including complementarity, as it violates most hypothesis formulated. As shown in the following chapter, some definitions, especially tailored for optimization problems with complementarity, must first be explained. With these new tools, one can then properly study the optimal control problem of LCS.

Chapter 3

Mathematical Programming with Complementarity Constraints

Abstract. The presence of complementarity in the optimization problem leads us to the study of Mathematical Programming with Equilibrium Constraints (MPEC). These are problems with tailored theoretical developments, as they do not suit most hypothesis made in the prior study of most Nonlinear Programming (NLP) problems. After some theoretical results concerning finite dimensional MPEC, and their numerical resolution, we will see how it can be adpated to the study of optimal control problem involving complementarity constraints.

Mathematical Programming with Equilibrium Constraints (MPEC) are optimization problems in which there are two variables: one which is a vector of parameters, the other one which is a primary variable submitted to equilibrium constraints. More specifically, for two positive integers n and m, denote by $f : \mathbb{R}^{n+m} \to \mathbb{R}$ an objective function to minimize, $F : \mathbb{R}^{n+m} \to \mathbb{R}^m$ a function called the equilibrium function, $Z \subseteq \mathbb{R}^{n+m}$ a non-empty closed set, and $\mathcal{C} : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ a multimap with (possibly empty) closed convex values. Denote by $\mathrm{SOL}(\mathcal{C}(x), F(x,\cdot))$ the set $\{y \in \mathcal{C}(x) \mid (v-y)^\intercal F(x,y) \geq 0, \forall v \in \mathcal{C}(x)\}$ (this is the set of solutions of the variational inequality defined by $F(x,\cdot)$ and $\mathcal{C}(x)$). Then, we call MPEC the optimization problem:

$$\begin{array}{l} \min\ f(x,y) \\ \text{s.t.} \begin{cases} (x,y) \in Z \\ y \in \mathrm{SOL}(\mathcal{C}(x), F(x,\cdot)) \end{cases} \end{array}$$

Suppose that: $\forall x \in \mathbb{R}^n$, $\mathcal{C}(x) = \mathbb{R}^m_+$. Then the variational inequality defined by $F(x,\cdot)$ and $\mathcal{C}(x)$ is a complementarity problem [54] and the MPEC becomes:

$$\begin{array}{l} \min\ f(x,y) \\ \text{s.t.} \begin{cases} (x,y) \in Z \\ 0 \leq y \perp F(x,y) \geq 0 \end{cases} \end{array}$$

which is more usually calles Mathematical Programming with Complementarity Constraints (MPCC).

MPCC may look like any other Nonlinear Program (NLP), but most tools used in order to analyze NLP do not apply, because of the violation of all standard constraint qualification (CQ), such as the Magasarian-Fromovitz CQ (MFCQ) or the Linear Independence CQ (LICQ) (see [115, Proposition 1.1]). Thus, most results concerning MPCC had to first redefine their own tailored

CQ, in order to derive cases for which first order conditions exist. These definitions are most important in order to tackle efficiently optimal control of LCS.

3.1 Finite dimension

This section is devoted to the study of the MPCC:

$$\begin{array}{l} \min f(x) \\ \text{s.t.} \begin{cases} g(x) \leq 0 \\ h(x) = 0 \\ 0 \leq G(x) \perp H(x) \geq 0 \end{cases} \end{array} \tag{3.1}$$

with continuously differentiable functions $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^q$, $h : \mathbb{R}^n \to \mathbb{R}^p$, $G, H : \mathbb{R}^n \to \mathbb{R}^m$. For "standard" nonlinear programs, there exist different ways to obtain first-order optimality conditions:

- a geometric approach (equality of the tangent cone and a linearized version);
- via exact penalty function (see Section 1.1.1 where the LCP is re-expressed with non-smooth functions), using then non-smooth analysis to derive first order conditions;
- via the Fritz-John conditions, which give first order conditions without using any CQ, but at the cost of having a multiplier in front of the objective function.

Under suitable CQ, all these methods give eventually the same results: the Karush-Kuhn-Tucker (KKT) conditions. For MPEC, things are slightly different, in the sense that each method will give its own results, that under some CQ will happen to be the same.

3.1.1 Analysis

When the Fritz-John conditions are directly applied to MPEC, they do not give efficient results. Thus, the results had to be refined (as shown in [66, Theorem 3.1]). Moreover, as it was mentioned in the introduction of this Section, the classical CQs are not verified for MPEC. Therefore, some special MPEC-tailored CQs have to be defined. A lot of different CQ can be found in the literature (see a zoology in [114]). We define here three of them. For that, define the index sets:

$$I^g(x) = \{i \in \overline{q} \mid g_i(x) = 0\},$$

$$I^{+0}(x) = \{i \in \overline{m} \mid G_i(x) > 0 = H_i(x)\},$$
$$I^{0+}(x) = \{i \in \overline{m} \mid G_i(x) = 0 < H_i(x)\},$$

and the biactive (or degenerate) set:

$$I^{00}(x) = \{i \in \overline{m} \mid G_i(x) = 0 = H_i(x)\}.$$

Also, denote by $I^{0\bullet}(x) = I^{0+}(x) \cup I^{00}(x) = \{i \in \overline{m} \mid G_i(x) = 0\}$ and $I^{\bullet 0}(x) = I^{+0}(x) \cup I^{00}(x) = \{i \in \overline{m} \mid H_i(x) = 0\}$

Definition 3.1.1. *Let $x^* \in \mathbb{R}^n$.*

- *We say that the MPEC LICQ holds at x^* if the family of gradients*

  $$\{\nabla g_i(x^*) \ : \ i \in I_g(x^*)\} \cup \{\nabla h_i(x^*) \ : \ i \in \overline{p}\} \cup \{\nabla G_i(x^*) : \ i \in I^{0\bullet}(x^*)\} \cup \{\nabla H_i(x^*) : \ i \in I^{\bullet 0}(x^*)\}$$

  is linearly independent.

- *The MPEC linear condition holds if all the functions g, h, G, H are affine.*

- *The MPEC MFCQ holds at x^* if the family of gradients*

  $$\{\nabla h_i(x^*) \ : \ i \in \overline{p}\} \cup \{\nabla G_i(x^*) \ : \ i \in I^{0\bullet}(x^*)\} \cup \{\nabla H_i(x^*) \ : \ i \in I^{\bullet 0}(x^*)\}$$

  is linearly independent, and there exists a vector $d \in \mathbb{R}^n$ such that:

  $$\nabla h_i(x^*)^\intercal d = 0, \ \forall i \in \overline{p}$$
  $$\nabla G_i(x^*)^\intercal d = 0, \ \forall i \in I^{0\bullet}(x^*)$$
  $$\nabla H_i(x^*)^\intercal d = 0, \ \forall i \in I^{\bullet 0}(x^*)$$
  $$\nabla g_i(x^*)^\intercal d < 0, \ \forall i \in I^g(x^*).$$

These constraint qualifications are often sufficient conditions for the local error bound condition to hold, which we define here. For this, define by $\mathcal{C}^m = \{(v, w) \in \mathbb{R}^m \mid 0 \leq v \perp w \geq 0\}$

Definition 3.1.2. *Let $S = \{x \in \mathbb{R}^n \ : \ g(x) \leq 0, \ h(x) = 0, \ (G(x), H(x)) \in \mathcal{C}^m\}$. We say that the local error bound condition holds (for the constrained system representing S) at $x^* \in S$ if there exist positive constants τ and δ such that:*

$$\operatorname{dist}_S(x) \leq \tau \left(\| \max\{0, g(x)\} \| + \|h(x)\| + \operatorname{dist}_{\mathcal{C}^m}(G(x), H(x)) \right), \ \forall x \in \mathcal{B}_\delta(x^*)$$

The analysis of MPEC involves different types of stationarity conditions, more or less strong, which will be verified by a local minimum depending on the CQ it complies with. Among all stationarity concepts designed for MPEC, we give here three definitions used for the rest of the manuscript.

Definition 3.1.3. *Let x^* be an admissible point for (3.1).*

1. *x^* is a W(eak) stationary point of (3.1) if there exist multipliers $(\lambda, \mu, \theta, \nu) \in \mathbb{R}^{p+q+2m}$ such that:*

   $$\nabla f(x^*) + \nabla g(x^*)^\intercal \lambda + \nabla h(x^*)\mu - \nabla G(x^*)\theta - \nabla H(x^*)\nu = 0,$$
   $$0 \leq \lambda \perp -g(x^*) \geq 0,$$
   $$\theta_i = 0, \ \forall i \in I^{+0}(x^*), \nu_i = 0, \ \forall i \in I^{0+}(x^*). \tag{3.2}$$

2. *x^* is a C(larke) stationary point of (3.1) if there exist multipliers $(\lambda, \mu, \theta, \nu) \in \mathbb{R}^{p+q+2m}$ such that it is W-stationary and $\forall i \in I^{00}(x^*)$:*

   $$\theta_i \nu_i \geq 0. \tag{3.3}$$

3. *x^* is a M(ordukovich) stationary point of (3.1) if there exist multipliers $(\lambda, \mu, \theta, \nu) \in \mathbb{R}^{p+q+2m}$ such that it is W-stationary and $\forall i \in I^{00}(x^*)$:*

   $$\text{either } [\theta_i \nu_i = 0] \text{ or } [\theta_i > 0 \text{ and } \nu_i > 0]. \tag{3.4}$$

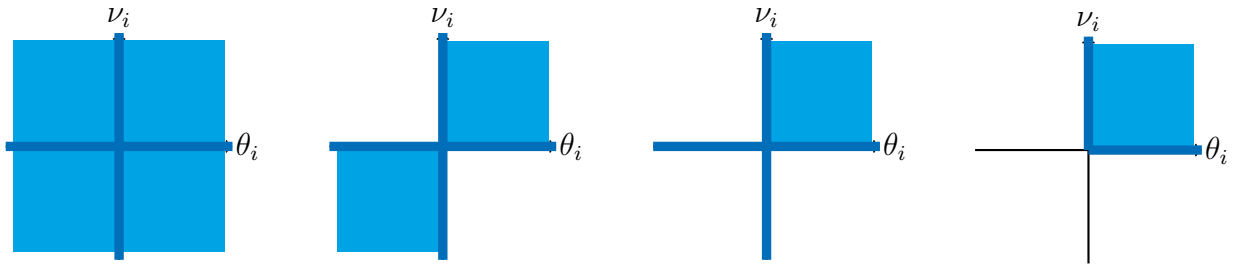
4. *x^* is a S(trong) stationary point of (3.1) if there exist multipliers $(\lambda, \mu, \theta, \nu) \in \mathbb{R}^{p+q+2m}$ such that it is W-stationary and $\forall i \in I^{00}(x^*)$:*

$$\theta_i \geq 0 \text{ and } \nu_i \geq 0. \tag{3.5}$$

From the definition, it is clear that one has the implication:

$$\text{S-stationarity} \implies \text{M-stationarity} \implies \text{C-stationarity} \implies \text{W-stationarity}$$

The difference between the different stationarity conditions lies in the biactive (or degenerate) set $I^{00}(x^*)$. This is illustrated in Figure 3.1.



(a) W-stationarity (b) C-stationarity (c) M-stationarity (d) S-stationarity

Figure 3.1: Geometric illustration of W-, M- and S-stationarity for i in the biactive set $I^{00}(x^*)$ (taken from [67]).

As explained in [114], these stationarity conditions are actually related to associated programs, for which they are the usual KKT conditions. For instance, the W-stationarity is the KKT condition of the *tightened* MPEC:

$$\min f(z)$$
$$\text{s.t.} \begin{cases} g(z) \leq 0,\ h(z) = 0, & \\ G_i(z) = 0,\ H_i(z) \geq 0 & \forall i \in I^{0+}(x^*), \\ G_i(z) \geq 0,\ H_i(z) = 0 & \forall i \in I^{+0}(x^*), \\ G_i(z) = 0,\ H_i(z) = 0 & \forall i \in I^{00}(x^*), \end{cases}$$

C-stationarity condition is the *nonsmooth KKT* condition (see [36, Chapter 6]) using Clarke subdifferential (see Definition 2.1.2) of:

$$\min f(z)$$
$$\text{s.t.} \begin{cases} g(z) \leq 0,\ h(z) = 0, & \\ G_i(z) = 0,\ H_i(z) \geq 0 & \forall i \in I^{0+}(x^*), \\ G_i(z) \geq 0,\ H_i(z) = 0 & \forall i \in I^{+0}(x^*), \\ \min\{G_i(z), H_i(z)\} = 0 & \forall i \in I^{00}(x^*), \end{cases}$$

while S-stationarity is the KKT condition for the *relaxed* MPEC:

$$\min f(z)$$
$$\text{s.t.} \begin{cases} g(z) \leq 0,\ h(z) = 0, & \\ G_i(z) = 0,\ H_i(z) \geq 0 & \forall i \in I^{0+}(x^*), \\ G_i(z) \geq 0,\ H_i(z) = 0 & \forall i \in I^{+0}(x^*), \\ G_i(z) \geq 0,\ H_i(z) \geq 0 & \forall i \in I^{00}(x^*). \end{cases}$$

With the previous definitions, we can now state some first-order conditions for (3.1).

Theorem 3.1.1. *1. [95, Theorem 2 (1)] Let x^* be a local optimal solution for (3.1), and suppose MPEC MFCQ holds at x^*. Then x^* is C-stationary.*

2. [114, Theorem 2.2] Suppose the MPEC linear CQ is met. Let x^* be a local optimal solution for (3.1). Then x^* is M-stationary.

3. [83, Theorem 3] Let x^* be a local optimal solution for (3.1), and suppose MPEC LICQ hold at x^*. Then x^* is S-stationary.

The definition of stationarities in Definition 3.1.3 uses some index sets that depend on the optimal solution (namely, $I^{+0}(x^*)$ and all the other (bi)active sets). This dependence may sometimes be cumbersome, and prevent the resolution of these stationary conditions directly. The next Theorem reformulates them, in order to get rid of these index sets.

Theorem 3.1.2. *[58, Theorem 3.1] For any x admissible for (3.1), we have the following statements:*

1. *Conditions (3.2) and (3.3) are equivalent to the equations*
$$\theta_i G_i(x) = \nu_i H_i(x) = 0,\ \theta_i \nu_i \geq 0,\ i \in \overline{m}$$
2. *Conditions (3.2) and (3.4) are equivalent to the equations*
$$\theta_i G_i(x) = \nu_i H_i(x) = 0,\ \theta_i \nu_i \geq 0,\ \max\{u_i, v_i\} \geq 0,\ i \in \overline{m}$$
3. *Conditions (3.2) and (3.5) are equivalent to the equations*
$$\begin{aligned} \theta_i = \alpha_i - \zeta H_i(x),\ \nu_i = \beta_i - \zeta G_i(x) \\ \alpha_i G_i(x) = \beta_i H_i(x) = 0,\ \alpha_i \beta_i \geq 0,\ i \in \overline{m} \end{aligned} \tag{3.6}$$
for $\alpha, \beta \in \mathbb{R}^m$ and $\zeta \in \mathbb{R}$.

Remark 3.1.1. *Concerning the reformulation of S-stationarity, one may see α and β as the multipliers associated with the constraints $G(x) \geq 0$, $H(x) \geq 0$, and ζ as the multiplier associated with the constraint $G(x)^\intercal H(x) = 0$, where (3.1) is considered as a standard NLP. Since neither the (standard) LICQ nor MFCQ apply directly to these constraints, [110, Theorem 1] states that these multipliers (α, β, ζ) can not be unique, nor can they remain in a bounded set.*

Other results related to this problem can be found in the aforementioned references and in [75, 82, 113].

3.1.2 Numerical resolution

Of course, most MPEC can not be solved analytically. In general, the numerical treatment of MPEC is not an easy task. Due to the violation of standard CQs, most optimization solvers are not guaranteed to converge. The algorithms designed for their resolution try to tackle this problem while focusing on two points:

- Of course, the algorithm has to converge to a local optimum, or at least to a stationary point.
- Preferably, it should fit in the already known algorithms used for classical optimization.

It should be noted also that there exists a benchmark with numerous MPEC problems and their solution, used for testing the different algorithms; see [72]. We present here two schemes that may converge to M- or S-stationary points.

Complementarity relaxation

In [67], the idea is to relax the complementarity condition using the following C-function (see Definition 1.1.1): define $\varphi : \mathbb{R}^2 \to \mathbb{R}$ by:

$$\varphi(a,b) = \begin{cases} ab & \text{if } a+b \geq 0 \\ -\frac{1}{2}(a^2+b^2) & \text{if } a+b < 0 \end{cases}$$

and then relax the problem as: let $\Phi : \mathbb{R}^n \times \mathbb{R}^+ \to \mathbb{R}^m$ be defined component-wise by:

$$\Phi_i(z,\tau) = \varphi(G_i(z) - \tau, H_i(z) - \tau)$$

With this function, we define the relaxed problem NLP(τ) for $\tau \geq 0$ as:

$$\begin{aligned} \min\ & f(z) \\ \text{s.t. } & g(z) \leq 0,\ h(z) = 0 \\ & G(z) \geq 0,\ H(z) \geq 0 \\ & \Phi(z,\tau) \leq 0 \end{aligned} \tag{NLP(τ)}$$

Algorithm 1: Relaxation algorithm $(z^0, \tau_0, \sigma, \tau_{min}, \varepsilon)$.

Input: A starting vector z^0, an initial relaxation parameter τ_0, and parameters $\sigma \in (0,1)$, $\tau_{min} > 0$, and $\varepsilon > 0$

1 Set k:=0

2 **while** *($\tau_k \geq \tau_{min}$ and compVio(z^k) $> \varepsilon$) or k=0* **do**

3 Find an approximate solution z^{k+1} of NLP(τ_k). To solve NLP(τ_k), use z^k as starting vector. If NLP(τ_k) is not feasible, terminate the algorithm

4 Let $\tau_{k+1} \leftarrow \sigma \min\{\tau_k, \text{compVio}(z^{k+1})\}$ and $k \leftarrow k+1$

5 **end**

Output: The final iterate $z_{opt} = z^k$, the corresponding function value $f(z_{opt})$, and the maximum constraint violation $maxVio(z_{opt})$.

In this algorithm, we denote:

$$\text{compVio}(z) = \max\{\min\{G_i(z), H_i(z)\}, i \in \overline{m}\}$$

$$\text{maxVio}(z) = \max\{\max\{0, g_j(z)\}, |h_k(z)|, |\min\{G_i(z), H_i(z)\}|, j \in \overline{q},\ k \in \overline{p},\ i \in \overline{m}\}$$

Theorem 3.1.3. *[67, Theorem 4.1, 4.2] Let $\{\tau_k\} \downarrow 0$ and $\{(z^k, \lambda^k, \mu^k, \gamma^k, \nu^k, \delta^k\}$ be a sequence of KKT points of NLP(τ_k) with $z^k \to z^*$. If MPEC LICQ holds in z^*, then z^* is an M-stationary point of the MPEC (3.1).*
Furthermore, if there is a subsequence $K \subseteq \mathbb{N}$ such that:

$$G_i(z^k) \leq \tau_k,\ H_i(z^k) \leq \tau_k,\ \forall k \in K,\ \forall i \in I^{00}(z^*)$$

then z^* is a S-stationary point of (3.1).

However, it should be noted that this convergence is actually sensitive to instabilities.

Definition 3.1.4. *Let $\varepsilon > 0$. We say that z^* is an ε-stationary point of the problem*

$$\min f(z) \ s.t. \ g(z) \leq 0, \ h(z) = 0$$

if there are multipliers λ and μ such that:

$$\|\nabla f(z^*) + (\nabla g(z^*))^\intercal \lambda + (\nabla h(z^*))^\intercal \mu\|_\infty \leq \varepsilon$$

$$g(z^*) \leq 0, \ \lambda \geq 0, \ \lambda_i g_i(z^*) \geq -\varepsilon, \ \forall i$$

$$|h_i(z^*)| \leq \varepsilon, \ \forall i$$

Theorem 3.1.4. *[67, Theorem 4.13] Let $\{\tau_k\} \downarrow 0$, $\varepsilon_k = o(\tau_k)$, and z^k be a sequence of ε_k-stationary points of NLP(τ_k) with multipliers $(\lambda^k, \mu^k, \gamma^k, \nu^k, \delta^k)$. Assume that $z^k \to z^*$. If MPEC LICQ holds at z^*, then z^* is a W-stationary point of MPEC (3.1). Furthermore, if there is a subsequence $K \subseteq \mathbb{N}$ such that:*

$$G_i(z^k) \leq \tau_k, \ H_i(z^k) \leq \tau_k, \ \forall k \in K, \ \forall i \in I^{00}(z^*)$$

then z^* is a C-stationary point of (3.1).

Cost penalization

The technique used in [73] is the penalization of the objective function. The complementarity is moved to the objective function in the form of an ℓ_1-penalty term, so that the objective becomes:

$$f(z) + \pi G(z)^\intercal H(z)$$

The associated barrier problem is defined as:

$$\begin{aligned} \min \ & f(z) + \pi G(z)^\intercal H(z) - \mu \left(\sum_i \log s_i + \sum_i \log G_i(z) + \sum_i \log H_i(z) \right) \\ \text{s.t.} \ & h(z) = 0 \\ & g(z) - s = 0 \end{aligned} \tag{3.7}$$

The Lagrangian of this barrier problem is given by:

$$\begin{aligned} \mathcal{L}_{\mu,\pi}(z, s, \lambda, \theta) = f(z) + \pi G(z)^\intercal H(z) - \mu \left(\sum_i \log s_i + \sum_i \log G_i(z) + \sum_i \log H_i(z) \right) \\ - \sum_i \theta_i h_i(z) - \sum_i \lambda_i (g_i(z) - s_i) \end{aligned}$$

Algorithm 2: Classic: A Practical Interior-Penalty Method for MPECs.

Input: Let $z^0, s^0, \lambda^0, \theta^0$ be the initial value of the primal and dual variables.

Set $k = 1$.

repeat

Choose a barrier parameter μ^k, a stopping tolerance ε^k_{pen} and ε^k_{comp}

Find π^k and an approximate solution $(x^k, s^k, \lambda^k, \theta^k)$ of problem (3.7) with parameter μ^k and π^k that satisfy $G(z^k) > 0$, $H(z^k) > 0$, $s^k > 0$, $\lambda^k > 0$ and the following conditions:

$$\|\nabla_z \mathcal{L}_{\mu^k,\pi^k}(z^k, s^k, \lambda^k, \theta^k)\| \leq \varepsilon^k_{pen}$$

$$\|s^k_i \lambda^k_i - \mu^k\| \leq \varepsilon^k_{pen}, \ \forall i$$

$$\left\| \begin{matrix} h(z^k) \\ g(z^k) - s^k \end{matrix} \right\| \leq \varepsilon^k_{pen}$$

$$\|\min\{G(z^k), H(z^k)\}\| \leq \varepsilon^k_{comp}$$

Let $k \leftarrow k + 1$

until *a stopping test for the MPEC is satisfied*

Theorem 3.1.5. *[73, Theorem 3.4 and Corollary 3.5] Suppose that Algorithm 2 generates an infinite sequence of iterates $\{z^k, s^k, \lambda^k, \theta^k\}$ and parameters $\{\pi^k, \mu^k\}$, for sequences $\{\varepsilon^k_{pen}\}$, $\{\varepsilon^k_{comp}\}$, $\{\mu^k\}$ converging to zero. If z^* is a limit point of the sequence $\{z^k\}$, and f, g and h are continuously differentiable in an open neighborhood $\mathcal{N}(z^*)$ of z^*, then z^* is feasible for (3.1). If in addition, MPEC LICQ holds at z^*, then z^* is C-stationary. Moreover, if $\{\pi^k\}$ is bounded, then z^* is an S-stationary point of (3.1).*

3.2 Optimal control

Recently in [59], different results concerning MPEC have been adapted in order to analyze an Optimal Control problem with complementarity constraints:

$$\begin{aligned} \min \ & J(x, \tilde{u}) = \int_0^T F(t, x(t), \tilde{u}(t))dt + f(x(0), x(T)) \\ & \left.\begin{aligned} & \dot{x}(t) = \phi(t, x(t), u(t)) \\ & g(t, x(t), u(t)) \leq 0 \\ \text{s.t.} \quad & h(t, x(t), u(t)) = 0 \\ & 0 \leq G(t, x(t), u(t)) \perp H(t, x(t), u(t)) \geq 0 \\ & u(t) \in U(t) \end{aligned} \right\} \text{ a.e. } t \in [0, T] \\ & (x(0), x(T)) \in \mathscr{E}. \end{aligned} \tag{3.8}$$

with $F : [0,T] \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, $f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, $\phi : [0,T] \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, $g : [0,T] \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^q$, $h : [0,T] \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^p$, $G, H : [0,T] \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^l$, $U : [0,T] \rightrightarrows \mathbb{R}^m$ a multifunction, and $\mathscr{E}$ a closed subset of $\mathbb{R}^{2n}$. This section will use different notions of normal cones and subdifferentials, which are defined in the Appendix A.

3.2.1 Necessary conditions

We suppose that F and ϕ are $\mathcal{L}\times\mathcal{B}$-measurable, where $\mathcal{L}\times\mathcal{B}$ denotes the σ-algebra of subsets of appropriate spaces generated by product sets $M\times N$, where M is a Lebesgue ($\mathcal{L}$) measurable subset in $\mathbb{R}$, and N is a Borel ($\mathcal{B}$) measurable subset in $\mathbb{R}^n\times\mathbb{R}^m$.

Definition 3.2.1.

- *We refer to any absolutely continuous function as an arc.*
- *An admissible pair for (3.8) is a pair of functions (x,u) on $[0,T]$ for which u is a control and x is an arc, that satisfy all the constraints in (3.8).*
- *The complementarity cone is defined by*

$$\mathcal{C}^m=\{(v,w)\in\mathbb{R}^m \mid 0\le v\perp w\ge 0\}$$

- *We define the set constraint at time $t\in[0,T]$, $S(t)$, by:*

$$S(t)=\{(x,u)\in\mathbb{R}^n\times U(t): g(t,x,u)\le 0,\ h(t,x,u)=0,\ (G(t,x,u),H(t,x,u))\in\mathcal{C}^m\}\,.$$

- *We say that the local error bound condition holds (for the constrained system representing S) at $(x^*,u^*)\in S$ if there exist positive constants τ and δ such that: $\forall(x,u)\in\mathcal{B}_\delta(x^*,u^*)$*

$$\mathrm{dist}_S(x,u)\le\tau\left(\|\max\{0,g(x,u)\}\|+\|h(x,u)\|+\mathrm{dist}_{\mathcal{C}^l}(G(x,u),H(x,u))\right)$$

- *For every given $t\in[0,T]$ and two positive constants R and ε, we define a neighbourhood of the point $(x^*(t),u^*(t))$ as:*

$$S_*^{\varepsilon,R}(t)=\{(x,u)\in S(t):\|x-x^*(t)\|\le\varepsilon,\ \|u-u^*(t)\|\le R\}. \tag{3.9}$$

(x^*,u^*) is a local minimizer of radius R if there exists ε such that for every pair (x,u) admissible for (3.8) such that:

$$(x(t),u(t))\in S_*^{\varepsilon,R}(t),\ \text{a.e. } t\in[0,T],\ \int_0^T\|\dot{x}^*(t)-\dot{x}(t)\|dt\le\varepsilon$$

we have $J(x^*,u^*)\le J(x,u)$.

Assumption 3.2.1. *(a) There exist measurable functions k_x^ϕ, k_x^F, k_w^ϕ, k_w^F such that for almost every $t\in[0,T]$ and for every (x^1,w^1), $(x^2,w^2)\in S_*^{\varepsilon,R}(t)$, we have:*

$$\begin{aligned}&(a)\ \|\phi(t,x^1,w^1)-\phi(t,x^2,w^2)\|\le k_x^\phi(t)\|x^1-x^2\|+k_w^\phi(t)\|w^1-w^2\|\\&(b)\ |F(t,x^1,w^1)-F(t,x^2,w^2)|\le k_x^F(t)\|x^1-x^2\|+k_w^F(t)\|w^1-w^2\|.\end{aligned} \tag{3.10}$$

(b) There exists a positive measurable function k_S such that for almost every $t\in[0,T]$, the bounded slope condition holds:

$$(x,w)\in S_*^{\varepsilon,R}(t),(\alpha,\beta)\in\mathcal{N}_{S(t)}^P(x,w)\implies\|\alpha\|\le k_S(t)\|\beta\|. \tag{3.11}$$

(c) The functions k_x^ϕ, k_x^F and $k_S[k_w^\phi+k_w^F]$ are integrable, and there exists a positive number η such that $R(t)\ge\eta k_S(t)$ a.e. $t\in[0,T]$.

(d) F and ϕ are $\mathcal{L}\times\mathcal{B}$-measurable, f is locally Lipschitz continuous, g, h, G and H are $\mathcal{L}$-measurable in variable t and strictly differentiable in variable (x,u), f is locally Lipschitz continuous, and $\mathscr{E}$ is a closed subset in $\mathbb{R}^{2n}$. U is a $\mathcal{L}$ measurable multifunction with convex values.

Assumption 3.2.1(b) is the central in [46], but it is rarely trivial to check if a given system complies with this assumption. As it is stated at the end of Section 2 in [46], this bounded slope condition excludes unilateral state constraints; that is, constraints of the type $x(t)\in X(t)$ (since in this case, there exists normals of the form $(\alpha,0)$ in (3.11)).

Define the sets

$$\begin{aligned}
I_t^-(x,u) &= \{i\in\overline{q}\ :\ g_i(t,x(t),u(t))<0\},\\
I_t^{+0}(x,u) &= \{i : G_i(t,x(t),u(t))>0=H_i(t,x(t),u(t))\},\\
I_t^{0+}(x,u) &= \{i : G_i(t,x(t),u(t))=0<H_i(t,x(t),u(t))\},\\
I_t^{00}(x,u) &= \{i : G_i(t,x(t),u(t))=0=H_i(t,x(t),u(t))\},
\end{aligned}$$

and for any $(\lambda^g,\lambda^h,\lambda^G,\lambda^H)\in\mathbb{R}^{q+p+m+m}$, denote:

$$\Psi(t,x,u;\lambda^g,\lambda^h,\lambda^G,\lambda^H)=g(t,x,u)^\intercal\lambda^g+h(t,x,u)^\intercal\lambda^h-G(t,x,u)^\intercal\lambda^G-H(t,x,u)^\intercal\lambda^H. \tag{3.12}$$

Theorem 3.2.1. *[59] Let (x^*,u^*) be a local minimizer of radius R for (3.8) and let Assumption 3.2.1 hold. If for almost every $t\in[0,T]$ the local error bound condition for the system representing $S(t)$ holds at $(x^*(t),u^*(t))$, then there exist a number $\lambda_0\in\{0,1\}$, an arc p and measurable functions $\lambda^g:\mathbb{R}\to\mathbb{R}^q$, $\lambda^h:\mathbb{R}\to\mathbb{R}^p$, $\lambda^G,\lambda^H:\mathbb{R}\to\mathbb{R}^m$ such that the following conditions hold:*

1. *the non-triviality condition: $(\lambda_0,p(t))\neq 0$, $\forall t\in[0,T]$*

2. *the transversality condition:*

$$(p(0),-p(T))\in\lambda_0\partial^L f(x^*(0),x^*(T))+\mathcal{N}_{\mathscr{E}}(x^*(0),x(T)) \tag{3.13}$$

3. *the Euler adjoint inclusion: for almost every $t\in[0,T]$,*

$$\begin{aligned}
(\dot p(t),0)\in\ &\partial^C\left\{\langle -p(t),\phi(t,\cdot,\cdot)\rangle+\lambda_0F(t,\cdot,\cdot)\right\}(x^*(t),u^*(t))\\
&+\nabla_{x,u}\Psi(t,x^*(t),u^*(t);\lambda^g(t),\lambda^h(t),\lambda^G(t),\lambda^H(t))\\
&+\{0\}\times\mathcal{N}_{U(t)}(u^*(t))
\end{aligned} \tag{3.14}$$

$$\lambda^g(t)\geq 0,\ \lambda_i^g(t)=0,\ \forall i\in I_t^-(x^*,u^*)$$

$$\lambda_i^G(t)=0,\ \forall i\in I_t^{+0}(x^*(t),u^*(t)),\ \lambda_i^H(t)=0,\ \forall i\in I_t^{0+}(x^*(t),u^*(t))$$

4. *the Weierstrass condition for radius R: for almost every $t\in[0,T]$,*

$$(x^*(t),u)\in S(t),\ \|u-u^*(t)\|<R(t)$$

$$\implies \langle p(t),\phi(t,x^*(t),u)\rangle-\lambda_0F(t,x^*(t),u)\leq\langle p(t),\phi(t,x^*(t),u^*(t))\rangle-\lambda_0F(t,x^*(t),u^*(t))$$

The Weierstrass condition can be re-expressed as searching a local minimizer of the following MPEC:

$$\begin{aligned} &\max \ \langle p(t), \phi(t, x^*(t), u)\rangle - \lambda_0 F(t, x^*(t), u) \\ &\text{s.t.} \ \begin{cases} (G(t, x^*(t), u), H(t, x^*(t), u)) \in \mathcal{C}^m \\ g(t, x^*(t), u) \leq 0, \ h(t, x^*(t), u) = 0 \end{cases} \end{aligned} \tag{3.15}$$

where $\mathcal{C}^m$ is defined in Definition 3.2.1. For each $t \in [0, T]$, this is an MPEC which admits stationarity conditions as exposed in Section 3.1.

Definition 3.2.2. *Let (x^*, u^*) be an admissible pair for (3.8).*

- *The Fritz-John (FJ) type W-stationarity holds at (x^*, w^*) if there exist a number $\lambda_0 \in \{0, 1\}$, an arc p and measurable functions λ^G, λ^H such that Theorem 3.2.1 (1)-(4) hold.*
- *The FJ-type M-stationarity holds at (x^*, u^*) if (x^*, u^*) is W-stationarity with arc p and there exist measurable functions η^g, η^h, η^G, η^H such that, for almost every $t \in [0, T]$,*

$$\begin{aligned} 0 \in \partial^L \{\langle -p(t), \phi(t, x^*(t), \cdot) + \lambda_0 F(t, x^*(t), \cdot)\rangle\}(u^*(t)) \\ + \nabla_u \Psi(t, x^*(t), u^*(t); \eta^g(t), \eta^h(t), \eta^G(t), \eta^H(t)) + \mathcal{N}_{U(t)}(u^*(t)) \end{aligned}$$

$$\eta^g(t) \geq 0, \ \eta^g_i(t) = 0, \ \forall i \in I_t^-(x^*, u^*)$$

$$\eta^G_i(t) = 0, \ \forall i \in I_t^{+0}(x^*(t), u^*(t)), \ \eta^H_i(t) = 0, \ \forall i \in I_t^{0+}(x^*(t), u^*(t))$$

$$\text{either } [\eta^G_i(t)\eta^H_i(t) = 0] \text{ or } [\eta^G_i(t) > 0 \text{ and } \eta^H_i(t) > 0], \ \forall i \in I_t^{00}(x^*(t), u^*(t)).$$

- *The FJ-type S-stationarity holds at (x^*, u^*) if (x^*, u^*) is W-stationarity with arc p and there exist measurable functions η^G, η^H such that, for almost every $t \in [0, T]$,*

$$\begin{aligned} 0 \in \partial^L \{\langle -p(t), \phi(t, x^*(t), \cdot) + \lambda_0 F(t, x^*(t), \cdot)\rangle\}(u^*(t)) \\ + \nabla_u \Psi(t, x^*(t), u^*(t); \eta^g(t), \eta^h(t), \eta^G(t), \eta^H(t)) + \mathcal{N}_{U(t)}(u^*(t)) \end{aligned}$$

$$\eta^g(t) \geq 0, \ \eta^g_i(t) = 0, \ \forall i \in I_t^-(x^*, u^*)$$

$$\eta^G_i(t) = 0, \ \forall i \in I_t^{+0}(x^*(t), u^*(t)), \ \eta^H_i(t) = 0, \ \forall i \in I_t^{0+}(x^*(t), u^*(t))$$

$$\eta^G_i(t) \geq 0, \ \eta^H_i(t) \geq 0, \ \forall i \in I_t^{00}(x^*(t), u^*(t)).$$

We refer to the FJ-type W-, M- and S-stationarities as the W-, M- and S-stationarities, respectively, if $\lambda_0 = 1$.

Notice that we have now two different sets of multipliers, and as shown in [59, Example 3.4], these new multipliers η^G, η^H can be different in measure from the corresponding λ^G, λ^H. Nonetheless, the results concerning stationarity for MPEC still can be applied to (3.15):

Theorem 3.2.2. *[59, Theorem 3.5] Let (x^*, u^*) be a local minimizer of radius R for (3.8), and let Assumption 3.2.1 hold. Suppose also that the MPEC linear condition holds for $S(t)$ for almost every $t \in [0, T]$, i.e. functions $g(t, \cdot, \cdot)$, $h(t, \cdot, \cdot)$, $G(t, \cdot, \cdot)$ and $H(t, \cdot, \cdot)$ are affine and $U(t)$ is a union of finitely many polyhedral sets. Then the FJ-type M-stationarity holds at (x^*, u^*).*

The difference between these two sets however vanishes if we further suppose that MPEC-LICQ holds:

Theorem 3.2.3. *[59] Let (x^*, u^*) be a local minimizer of radius R for (3.8), and let Assumption 3.2.1 hold. If for almost every $t \in [0,T]$, the functions $F(t,\cdot,\cdot)$ and $\phi(t,\cdot,\cdot)$ are strictly differentiable at $(x^*(t), u^*(t))$, and the MPEC LICQ holds at $u^*(t)$ for problem (3.15), i.e., the family of gradients*

$$\{\nabla_u g_i(t, x^*(t), u^*(t)) : \ i \in I_t^0(x^*(t), u^*(t))\} \cup \{\nabla_u h(t, x^*(t), u^*(t)) \ : \ i \in \overline{p}\} \cup$$
$$\{\nabla_u G_i(t, x^*(t), u^*(t)) : i \in I_t^{0\bullet}(x^*(t), u^*(t))\} \cup \{\nabla_u H_i(t, x^*(t), u^*(t)) : i \in I_t^{\bullet 0}(x^*(t), u^*(t))\}$$

is linearly independent, where

$$I_t^0(x,u) = \{i \in \overline{q} \ : \ g_i(t, x(t), u(t)) = 0\}$$

$$I_t^{0\bullet}(x^*(t), u^*(t)) = I_t^{0+}(x^*(t), u^*(t)) \cup I_t^{00}(x^*(t), u^*(t)),$$
$$I_t^{\bullet 0}(x^*(t), u^*(t)) = I_t^{+0}(x^*(t), u^*(t)) \cup I_t^{00}(x^*(t), u^*(t)),$$

then the FJ-type S-stationarity holds at (x^*, u^*). Moreover, the multipliers η^g, η^h, η^G, η^H can be taken as equal to λ^g, λ^h, λ^G, λ^H, respectively, almost everywhere.

3.2.2 Sufficient condition for the Bounded Slope Condition (3.11)

The Bounded Slope Condition (3.11) is not an easy assumption to check. There are, however, some results that give sufficient conditions for it to hold. We give here two of them.

Proposition 3.2.1. *[59] Assume that the local error bound condition holds at every $(x,u) \in S_*^{\varepsilon,R}(t)$ and:*

$$\left.\begin{array}{r} (x,u) \in S_*^{\varepsilon,R}(t), \zeta \in \mathcal{N}_{U(t)}^L(u), \\ \lambda^g \geq 0, \ \lambda_i^g = 0 \ \forall i \in I_t^-(x,u), \\ \lambda_i^G = 0, \ \forall i \in I_t^{+0}(x,u), \ \lambda_i^H = 0, \ \forall i \in I_t^{0+}(x,u), \\ \lambda_i^G > 0, \ \lambda_i^H > 0, \ \text{or } \lambda_i^G \lambda_i^H = 0, \ \forall i \in I_t^{00}(x,u) \end{array}\right\}$$
$$\implies \|\nabla_x \Psi(t,x,u;\lambda^g,\lambda^h,\lambda^G,\lambda^H)\| \leq k_S(t)\|\nabla_u \Psi(t,x,u;\lambda^g,\lambda^h,\lambda^G,\lambda^H) + \zeta\|,$$

where Ψ is defined in (3.12). Then the bounded slope condition (3.11) holds.

Let us define the following set:

$$C_*^{\varepsilon,R} = cl\{(t,x,u) \in [0,T] \times \mathbb{R}^n \times \mathbb{R}^m : (x,u) \in S_*^{\varepsilon,R}(t)\}$$

Proposition 3.2.2. *[59] Let the mappings g, h, G, H, U be autonomous. Assume that $C_*^{\varepsilon,R}$ is compact for some $\varepsilon > 0$, the local error bound holds, and that, for every (x,u) such that $(t,x,u) \in C^{\varepsilon,R}$, the system complies with the following implication:*

$$\left.\begin{array}{r} 0 \in \nabla_u \Psi(t,x,u;\lambda^g,\lambda^h,\lambda^G,\lambda^H) + \mathcal{N}_{U(t)}^L(u) \\ \lambda^g \geq 0, \ \lambda_i^g = 0 \ \forall i \in I_t^-(x,u), \\ \lambda_i^G = 0, \ \forall i \in I_t^{+0}(x,u), \ \lambda_i^H = 0, \ \forall i \in I_t^{0+}(x,u), \\ \lambda_i^G > 0, \ \lambda_i^H > 0, \ \text{or } \lambda_i^G \lambda_i^H = 0, \ \forall i \in I_t^{00}(x,u) \end{array}\right\} \implies \nabla_x \Psi(t,x,u;\lambda^g,\lambda^h,\lambda^G,\lambda^H) = 0,$$

where Ψ is defined in (3.12). Then there exists a positive constant k_S such that for every $t \in [0,T]$, the bounded slope condition (3.11) holds with $k_S(t) = k_S$.

Conclusion

This presentation of MPEC showed how specific is the study of such systems, and which tools are needed to tackle them properly. These will enable us to tackle efficiently the optimal control of LCS, both analytically and numerically, as shown in the following chapters.

Part II

Quadratic optimal control of LCS: the 1D complementarity case

Chapter 4

First order conditions using Clarke's subdifferential

Abstract. The optimal control of LCS is first tackled when the complementarity is one dimensional. In this case, the system can be re-expressed as a Lipschitz dynamical system, and Clarke's results concerning the optimal control of such system is used. After that, the optimality conditions are equivalently expressed in a different form. All these results are used to derive the analytical solution of a simple 1D class of problems. This Chapter has been published in [107].

We start the analysis of this problem with a simplified problem, with a complementarity which is one dimensional:

$$\begin{cases} \dot{x}(t) = Ax(t) + B\lambda^x(t) + Fu(t), \\ 0 \leq \lambda^x(t) \perp Cx(t) + d\lambda^x(t) + eu(t) \geq 0, \quad \text{a.e. on } [0,T] \\ x(0) = x_0, \ x(T) \text{ free} \end{cases} \tag{4.1}$$

where $A \in \mathbb{R}^{n\times n}$, $B, F \in \mathbb{R}^n$, $C \in \mathbb{R}^{1\times n}$, $d, e \in \mathbb{R}$, $d > 0$, $T > 0$, $x : [0,T] \to \mathbb{R}^n$ is absolutely continuous, $u : [0,T] \to \mathbb{R}$ is square integrable, and $\lambda^x : [0,T] \to \mathbb{R}$ is measurable. In order to avoid trivial cases, we assume that $(C, e) \neq (0, 0)$.

We define now the optimal control problem of finding the trajectories of (4.1) minimizing the functional:

$$C(x,u) = \int_0^T f^0(x(t), u(t))dt = \int_0^T (x(t)^\intercal x(t) + u(t)^2)dt. \tag{4.2}$$

In this case, an optimal trajectory (x^*, u^*) is a $W^{1,1}$ local minimizer, meaning: there exists $\varepsilon > 0$ such that, for any admissible trajectory for (4.1) (x, u) which verifies:

$$\|x - x^*\|_{W^{1,1}} = \int_0^T \|\dot{x}(t) - \dot{x}^*(t)\|dt \leq \varepsilon$$

one has $C(x^*, u^*) \leq C(x, u)$.

The reason to focus on this simplified problem is that basic convex analysis proves that $\lambda^x(t) = \frac{1}{d}\Pi_{\mathbb{R}^+}(-Cx(t) - eu(t))$, where $\Pi_K(x)$ is the projection of x on the set K. Therefore, (4.1) becomes (where the argument t is omitted for simplification):

$$\begin{aligned} \dot{x} &= f^x(x,u), \\ &= Ax + Fu + \frac{B}{d}\Pi_{\mathbb{R}^+}(-Cx - eu), \\ &= \begin{cases} Ax + Fu & \text{if } Cx + eu \geq 0, \\ \left(A - \frac{BC}{d}\right)x + \left(F - \frac{e}{d}B\right)u & \text{otherwise,} \end{cases} \end{aligned} \tag{4.3}$$

with obvious definition for f^x. With this simple system, some first order conditions of optimality will be derived, that will show some advantages that LCS possess over other modeling paradigms.

4.1 Derivation of the maximum principle

4.1.1 Preliminary results

Remark that (4.3) is actually a Lipschitz differential equation, and the optimal control problem (4.1)(4.2) admits no further constraints. Therefore, we can use Theorem 2.1.1. It yields the following result:

Proposition 4.1.1. *Let (x^*, u^*) be a local minimizer. There exists an absolutely continuous function $p : [0, T] \to \mathbb{R}^n$ such that:*

$$-\dot{p}(t) \in f^p(x^*(t), p(t), u^*(t)) = \partial_x^C f^x(x^*(t), u^*(t))p(t) - \frac{1}{2}\partial_x^C f^0(x^*(t), u^*(t)), \tag{4.4}$$

$p(0)$ is free, and $p(T) = 0$. It also complies with the Weierstrass condition:

$$\langle f^x(x^*(t), u^*(t)), p(t)\rangle - \frac{1}{2}f^0(x^*(t), u^*(t)) = \sup_{v\in\mathbb{R}^m}\left\{\langle f^x(x^*(t), v), p(t)\rangle - \frac{1}{2}f^0(x^*(t), v)\right\} \tag{4.5}$$

Recall that ∂_x^C denotes the Clarke subdifferential with respect to the x variable, defined in the Appendix A.

Proof. First, we need to define properly the set X in (2.4). As it has been remarked in [40, Remark 4], X can depend on time as long as an ε-neighbourhood of $x^*(t)$ lies in the interior of $X(t)$. Let us define $X(t)$ as an $\tilde{\varepsilon}$-neighbourhood of $x^*(t)$ with $\tilde{\varepsilon}$ big enough.
Since $x^*(T)$ is free, it implies $p(T) = 0$. We easily prove that the non-positive number λ appearing in Theorem 2.1.1 can not be 0 (since otherwise, $|p(T)| + \lambda = 0$). Therefore it can be chosen as $-\frac{1}{2}$. The differential inclusion and the Weierstrass condition are just simple transcripts of the result.
We just need to verify the hypotheses **H1**-**H3**. **H1** and **H3** are easily verified. Concerning **H2** (which asserts that the dynamical system is Lipschitz with respect to x), since f^x and f^0 are both locally Lipschitz, and since $X(t)$ is bounded for all t in $[0, T]$, the inequality trivially holds. □

Let us compute these differential inclusions. Since f^0 is smooth, its subdifferential only contains the gradient:

$$\partial_x^C f^0(x, u) = \{2x\}$$

Concerning the right-hand side of the dynamical system, its subdifferential is expressed as:

$$\partial_x^C f^x(x, u) = \begin{cases} \{A\} & \text{if} \quad Cx + eu > 0, \\ \{A - \frac{BC}{d}\} & \text{if} \quad Cx + eu < 0, \\ [A - \frac{BC}{d}, A] & \text{if} \quad Cx + eu = 0. \end{cases}$$

using the notation:

$$[M_1, M_2] = \text{conv}\{M_1, M_2\}$$

for any pair (M_1, M_2) of according dimensions matrices and where $\text{conv}\{M_1, M_2\}$ stands for the convex hull of M_1 and M_2.

Therefore, (4.4) becomes:

$$-\dot{p}(t) \in \begin{cases} \{-x(t) + A^\intercal p(t)\} & \text{if } Cx + eu > 0, \\ \left\{-x(t) + \left(A - \frac{BC}{d}\right)^\intercal p(t)\right\} & \text{if } Cx + eu < 0, \\ \{-x(t)\} + \left[\left(A - \frac{BC}{d}\right)^\intercal, A^\intercal\right] p(t) & \text{if } Cx + eu = 0. \end{cases} \tag{4.6}$$

In order to compute an optimal control, the system should obviously be controllable between the initial and final points. We will only focus on completely controllable systems, relying on Theorem 1.2.3 stating the complete controllability of some LCS.

4.1.2 Equations (4.1) and (4.6) as a mixed LCS (MLCS)

We denote $z = \begin{pmatrix} x \\ p \end{pmatrix}$. Let us rewrite (4.1) and (4.6) in the compact form:

$$\dot{z} \in \begin{pmatrix} f^x(z,u) \\ f^p(z,u) \end{pmatrix} \tag{4.7}$$

where the right-hand side is a set-valued function defined by:

- if $Cx + eu > 0$:
$$\begin{pmatrix} f^x(z,u) \\ f^p(z,u) \end{pmatrix} = \left\{ \begin{pmatrix} Ax + Fu \\ x - A^\intercal p \end{pmatrix} \right\},$$
- if $Cx + eu < 0$:
$$\begin{pmatrix} f^x(z,u) \\ f^p(z,u) \end{pmatrix} = \left\{ \begin{pmatrix} \left(A - \frac{CB}{d}\right) x + \left(F - \frac{e}{d}B\right) u \\ x + \left(\frac{C^\intercal B^\intercal}{d} - A^\intercal\right) p \end{pmatrix} \right\},$$
- if $Cx + eu = 0$:
$$\begin{pmatrix} f^x(z,u) \\ f^p(z,u) \end{pmatrix} = \begin{pmatrix} \left(A - \frac{FC}{e}\right) x \\ x + \left[\frac{C^\intercal B^\intercal}{d} - A^\intercal, -A^\intercal\right] p \end{pmatrix}.$$

We can recast the differential inclusion (4.7) in the framework of complementarity systems with linear dynamics:

$$\dot{z} = \begin{pmatrix} g^x(z,u) \\ g^p(z,u) \end{pmatrix} = \begin{pmatrix} A & 0 \\ I & -A^\intercal \end{pmatrix} z + \begin{pmatrix} B & 0 \\ 0 & \frac{C^\intercal B^\intercal}{d} \end{pmatrix} \begin{pmatrix} \lambda^x \\ \lambda_1^p \end{pmatrix} + \begin{pmatrix} F \\ 0 \end{pmatrix} u, \tag{4.8}$$

$$0 \le \lambda^x \perp Cx + d\lambda^x + eu \ge 0, \tag{4.9a}$$
$$0 \le |\lambda_1^{p_j}| \perp Cx + eu + |Cx + eu| \ge 0, \tag{4.9b}$$
$$0 \le |\lambda_2^{p_j}| \perp |Cx + eu| - (Cx + eu) \ge 0 \qquad j = 1...n, \tag{4.9c}$$
$$|\lambda_1^{p_j}| + |\lambda_2^{p_j}| = |p_j| \qquad j = 1...n, \tag{4.9d}$$
$$\lambda_1^p + \lambda_2^p = p, \tag{4.9e}$$

where the subscript j denotes the j-th component of a vector, and $\lambda_i^p = (\lambda_i^{p_1}, ..., \lambda_n^{p_n})^\intercal$.

Proposition 4.1.2. *The right-hand side of (4.7) is the same as the right-hand side of (4.8) defined with the complementarity conditions in (4.9).*

Proof. The first line (4.9a) gives obviously the same right-hand side as in the second line of (4.1), which is $f^x(z,p)$. Then, $f^x(z,u) = g^x(z,u)$. Therefore, we have to check that the other lines in (4.9) are the same as in (4.6). To do so, we have to distinguish 3 cases:

- if $Cx + eu > 0$, then from (4.9b), we deduce that $\lambda_1^p = 0$. It follows that:

$$g^p(z,u) = x - A^\intercal p = f^p(z,u).$$

- if $Cx + eu < 0$, then from (4.9c), $\lambda_p^2 = 0$, and therefore from (4.9e), $\lambda_1^p = p$. We then have the following equality:

$$g^p(z,u) = x + \left(\frac{C^\intercal B^\intercal}{d} - A^\intercal\right) p = f^p(z,u).$$

- if $Cx + eu = 0$, then from (4.9d), we have that $|\lambda_1^{p_j}| + |\lambda_2^{p_j}| = |p_j|$, so that $\lambda_i^{p_j} \in [-p_j, p_j]$, $i = 1,2$, $j = 1,...n$. But, in order to comply with the last equality (4.9e), we must have $\lambda_i^{p_j} \in [0,1]p_j$, $i = 1,2$, $j = 1,...,n$, and hence, $\lambda_i^p \in [0,1]p$, $i = 1,2$. This gives us

$$\begin{aligned} g^p(z,u) &= \{x - A^\intercal p\} + [0,1]\frac{C^\intercal B^\intercal}{d} p \\ &= \{x\} + \left[\frac{C^\intercal B^\intercal}{d} - A^\intercal, -A^\intercal\right] p \\ &= f^p(z,u). \end{aligned}$$

□

We can still have an even more interesting form of (4.9) by noticing that the function $|\cdot|$ is piecewise linear and so, admits a representation in the form of an LCP. This is the topic of the next lemma:

Lemma 4.1.1. *Define $\mathbb{1}_n = (1,...,1)^\intercal \in \mathbb{R}^n$. The multipliers λ^x and λ_1^p given by system (4.9) are equally defined by the following system:*

$$\left\{\begin{array}{lcr} 0 \leq \lambda^x & \perp & Cx + d\lambda^x + eu \geq 0 \\ 0 \leq \mu_{x,u} & \perp & \mu_{x,u} - 2(Cx + eu) \geq 0 \\ 0 \leq \mu_p & \perp & \mu_p - 2p \geq 0 \\ 0 \leq \lambda_1^{abs} & \perp & \mu_{x,u}\mathbb{1}_n \geq 0 \\ 0 \leq \lambda_2^{abs} & \perp & (\mu_{x,u} - 2(Cx + eu))\mathbb{1}_n \geq 0 \\ 0 \leq \mu_1 & \perp & \mu_1 - 2\lambda_1^p \geq 0 \\ 0 \leq \mu_2 & \perp & \mu_2 - 2\lambda_2^p \geq 0 \\ & \lambda_1^{abs} + \lambda_2^{abs} = \mu_p - p & \\ & \lambda_1^p + \lambda_2^p = p & \\ & \mu_1 - \lambda_1^p = \lambda_1^{abs} & \\ & \mu_2 - \lambda_2^p = \lambda_2^{abs} & \end{array}\right. \tag{4.10}$$

where (4.10) is a mixed LCP (MLCP).

Proof. First, we need to establish the following simple result: for any scalar x, $|x| = \mu - x$ where μ is given by:

$$0 \leq \mu \perp \mu - 2x \geq 0.$$

Indeed, if $x \leq 0$, then we must take $\mu = 0$, so that $\mu - x = -x = |x|$. If $x > 0$, then we must take $\mu = 2x$, and $\mu - x = x = |x|$. Let us use (4.9) to rewrite equivalently the absolute values as:

$$\begin{aligned}
0 \leq \mu_{x,u} &\perp \mu_{x,u} - 2(Cx + eu) \geq 0,\\
0 \leq \mu_1 &\perp \mu_1 - 2\lambda_1^p \geq 0,\\
0 \leq \mu_2 &\perp \mu_2 - 2\lambda_2^p \geq 0,\\
0 \leq \mu_p &\perp \mu_p - 2p \geq 0,
\end{aligned}$$

$$\begin{aligned}
\lambda_{x,u}^{abs} &= \mu_{x,u} - (Cx + eu) = |Cx + eu|,\\
\lambda_1^{abs} &= \mu_1 - \lambda_1^p = |\lambda_1^p|,\\
\lambda_2^{abs} &= \mu_2 - \lambda_2^p = |\lambda_2^p|,\\
\lambda_p^{abs} &= \mu_p - p^\intercal = |p|,
\end{aligned}$$

where notation $|x|$ on a vector x is here understood component-wise, e.g. $|x| = (|x_1|, ..., |x_n|)^\intercal$. Therefore, (4.9) becomes:

$$\left\{\begin{array}{rcr}
0 \leq \lambda^x & \perp & Cx + d\lambda^x + eu \geq 0\\
0 \leq \mu_{x,u} & \perp & \mu_{x,u} - 2(Cx + eu) \geq 0\\
0 \leq \mu_1 & \perp & \mu_1 - 2\lambda_1^p \geq 0\\
0 \leq \mu_2 & \perp & \mu_2 - 2\lambda_2^p \geq 0\\
0 \leq \mu_p & \perp & \mu_p - 2p \geq 0\\
0 \leq \lambda_1^{abs} & \perp & (Cx + eu + \lambda_{x,u}^{abs})\mathbb{1}_n \geq 0\\
0 \leq \lambda_2^{abs} & \perp & (\lambda_{x,u}^{abs} - (Cx + eu))\mathbb{1}_n \geq 0\\
\multicolumn{3}{c}{\lambda_1^{abs} + \lambda_2^{abs} = \lambda_p^{abs}}\\
\multicolumn{3}{c}{\lambda_1^p + \lambda_2^p = p}\\
\multicolumn{3}{c}{\lambda_{x,u}^{abs} = \mu_{x,u} - (Cx + eu)}\\
\multicolumn{3}{c}{\lambda_1^{abs} = \mu_1 - \lambda_1^p}\\
\multicolumn{3}{c}{\lambda_2^{abs} = \mu_2 - \lambda_2^p}\\
\multicolumn{3}{c}{\lambda_p^{abs} = \mu_p - p}
\end{array}\right.$$

Noticing that we can use the two equalities on $\lambda_{x,u}^{abs}$ and λ_p^{abs} and insert them above, we have proven that λ^x and λ_1^p are equally defined by (4.9) or (4.10). □

Therefore, we infer that the right-hand side of the differential inclusion in (4.7) is equal to the right-hand side of system (4.8):

$$\dot{z} = \tilde{A}z + \tilde{B}\Lambda + \tilde{F}u, \tag{4.11}$$

where $\tilde{A}$, $\tilde{B}$, $\tilde{F}$ and Λ are easily identifiable from (4.8), and subject to the MLCP (4.10).

4.2 Analytical solution for a 1D example

This result can be used in order to find the unique stationary trajectory of a special class of LCS. Suppose in (4.1) that $n = 1$ and $C = 0$; it means we try to tackle the following problem:

$$\begin{aligned} &\text{minimize} \int_0^T \left(x(t)^2 + u(t)^2\right) dt, \\ &\text{such that:} \begin{cases} \dot{x}(t) = ax(t) + b\lambda(t) + fu(t), \\ 0 \leq \lambda(t) \perp d\lambda(t) + eu(t) \geq 0, \text{ a.e. on } [0,T] \\ x(0) = x_0,\ x(T) \text{ free}, \end{cases} \end{aligned} \tag{4.12}$$

where all variables are scalar, $d > 0$, $b, e \neq 0$.

4.2.1 Complete controllability conditions for this 1D example

In order to analyze this problem, we will first specify the necessary and sufficient complete controllability conditions in the 1D case. Applying Theorem 1.2.3, we must check that the following system:

$$(a - \lambda)\zeta = 0, \tag{4.13}$$
$$f\zeta + e\eta = 0, \tag{4.14}$$
$$\eta \geq 0, \tag{4.15}$$
$$b\zeta + d\eta \leq 0, \tag{4.16}$$

has no solution $\lambda \in \mathbb{R}$ and $(\zeta, \eta) \neq 0$

If $e > 0$: we deduce through (4.14): $\eta = -\frac{f\zeta}{e}$.

1. If $f = 0$, then $\eta = 0$. In (4.13), we can take $\lambda = a$. However, with (4.16), we have that $\zeta b \leq 0$. Let us take $\zeta = -\text{sign}(b)$. Then we found a solution with $\zeta \neq 0$: the system is not completely controllable.
2. If $f < 0$, then with (4.15), we have that $\zeta \geq 0$. Through (4.13), we take $\lambda = a$.
   - If $b \geq 0$, then (4.16) is a sum of positive terms which must be nonpositive, so $\eta = 0$ and $\zeta = 0$: the system is completely controllable.
   - If $b < 0$, then (4.16) becomes $\zeta(b - \frac{fd}{e}) \leq 0$ with $\zeta \geq 0$.
     - If $b - \frac{fd}{e} \leq 0$ then we can take any $\zeta \geq 0$: the system is not completely controllable.
     - Otherwise, only $\zeta = 0$ suits, so $\eta = 0$, and then the system is completely controllable.
3. If $f > 0$, then in (4.13), we take $\lambda = a$. Through (4.15), we have that $\zeta \leq 0$.
   - If $b \leq 0$, then (4.16) is a positive terms sum which must be nonpositive, so $\eta = 0$ and $\zeta = 0$: the system is completely controllable.
   - If $b > 0$, then (4.16) becomes $\zeta(b - \frac{fd}{e}) \leq 0$ with $\zeta \leq 0$.
     - If $b - \frac{fd}{e} \geq 0$ then then we can take any $\zeta \leq 0$: the system is not completely controllable.
     - Otherwise, only $\zeta = 0$ suits, so $\eta = 0$, and then the system is completely controllable.

If $e < 0$: we have the same cases as with $e > 0$ by inverting the sign of f.

4.2.2 Search for the explicit optimal solution

The dynamic system in (4.12) can be rewritten as:

$$\dot{x} = ax + fu + \frac{b}{d}\Pi_{\mathbb{R}_+}(-eu). \tag{4.17}$$

Therefore, the Hamiltonian function is written as:

$$H(x,p,u) = p\left(ax + fu + \frac{b}{d}\Pi_{\mathbb{R}_+}(-eu)\right) - \frac{1}{2}\left(x^2 + u^2\right). \tag{4.18}$$

We notice that this equation is smooth in x. Therefore, using (4.6), the adjoint equation is smooth, and is written as:

$$\dot{p}(t) = -ap(t) + x(t).$$

We can even differentiate twice p, and obtain the following second-order differential equation:

$$\begin{aligned}
\ddot{p} &= -a\dot{p} + \dot{x} \\
&= a^2 p + fu + \frac{b}{d}\Pi_{\mathbb{R}_+}(-eu) \\
&= \begin{cases} a^2 p + fu & \text{if } eu \geq 0 \\ a^2 p + \left(f - \frac{be}{d}\right) u & \text{if } eu \leq 0. \end{cases}
\end{aligned}$$

We now search for an expression of the optimal control u^*, function of x and p, maximizing the Hamiltonian function $H(x,p,u^*)$. To that aim, we use the subdifferential of H with respect to u, written $\partial_u^C H(x,p,u)$, and the fact that if u^* maximizes H, then

$$0 \in \partial_u^C H(x,p,u^*).$$

In our problem, the subdifferential is written as

$$\partial_u^C H(x,p,u) = \begin{cases} \{fp - u\} & \text{if } eu > 0, \\ \left\{\left(f - \frac{eb}{d}\right) p - u\right\} & \text{if } eu < 0, \\ -\left[f - \frac{eb}{d}, f\right] p & \text{if } eu = 0. \end{cases}$$

We now only focus on the complete controllable cases in order to find a control u maximizing this function:

If $e > 0$: In that case, $\text{sgn}(eu) = \text{sgn}(u)$.

1. We consider first $f < 0$.
   - If $b > 0$, then if $p \leq 0$, then $fp \geq 0$, $\left(f - \frac{eb}{d}\right) p \geq 0$, and if $p \geq 0$, then $fp \leq 0$, $\left(f - \frac{eb}{d}\right) p \leq 0$. We also notice that $0 \notin \left[f, f - \frac{eb}{d}\right]$. So we have:

$$u^* = \begin{cases} fp & \text{if } p \leq 0, \\ \left(f - \frac{eb}{d}\right) p & \text{if } p \geq 0. \end{cases}$$

- If $b < 0$, then we must make sure that $f < \frac{eb}{d}$. We notice that in this case, $0 \notin \left[f, f - \frac{eb}{d}\right]$. We are then in the exact same case as the previous one, and therefore, the control is expressed the same way:

$$u^* = \begin{cases} fp & \text{if } p \leq 0, \\ \left(f - \frac{eb}{d}\right) p & \text{if } p \geq 0. \end{cases}$$

2. We consider now $f > 0$.

- If $b < 0$, then if $p \leq 0$, then $fp \leq 0$, $\left(f - \frac{eb}{d}\right) p \leq 0$, and if $p \geq 0$, then $fp \geq 0$, $\left(f - \frac{eb}{d}\right) p \geq 0$. We also notice that $0 \notin \left[f, f - \frac{eb}{d}\right]$. So we have:

$$u^* = \begin{cases} fp & \text{if } p \geq 0, \\ \left(f - \frac{eb}{d}\right) p & \text{if } p \leq 0. \end{cases}$$

- If $b > 0$, then we must make sure that $f > \frac{eb}{d}$. We notice that in this case, $0 \notin \left[f, f - \frac{eb}{d}\right]$. We are then in the exact same case as the previous one, and therefore, the control is expressed the same way:

$$u^* = \begin{cases} fp & \text{if } p \geq 0, \\ \left(f - \frac{eb}{d}\right) p & \text{if } p \leq 0 \end{cases}.$$

If $e < 0$: we have the same cases as with $e > 0$ by inverting the sign of f.

Therefore, we can summarize this result as follows:

$$u^* = \begin{cases} fp & \text{if } efp \geq 0, \\ \left(f - \frac{eb}{d}\right) p & \text{if } efp \leq 0. \end{cases} \tag{4.19}$$

Finally, we use the optimal control found in (4.19) in the equation found on $\ddot{p}$. Surprisingly, we end up with a rather simple equation:

$$\ddot{p} = \begin{cases} \left(a^2 + f^2\right) p & \text{if } efp \geq 0, \\ \left(a^2 + \left(f - \frac{be}{d}\right)^2\right) p & \text{if } efp \leq 0, \end{cases}$$

which we rewrite in the more simple form:

$$\ddot{p} = \gamma(p)p \tag{4.20}$$

with $\gamma(p) > 0$ and piecewise constant.

We need now to find $p(0)$ such that $p(T) = 0$ (since $x(T)$ is free, according to the maximum principle). Moreover, we know that the initial value for the derivative $\dot{p}$ is given by:

$$\dot{p}(0) = x(0) - ap(0)$$

The phase portrait is depicted in Figure 4.1.

It is clear that, in order to have $p(T) = 0$, the sign of $p(0)$ is determined by the sign of the constants in the model:

- If $a > 0$, $x(0) > 0$, then $p(T) = 0 \implies p(0) < 0$,
- If $a > 0$, $x(0) < 0$, then $p(T) = 0 \implies p(0) > 0$,
- If $a < 0$, $x(0) > 0$, then $p(T) = 0 \implies p(0) < 0$,
- If $a < 0$, $x(0) < 0$, then $p(T) = 0 \implies p(0) > 0$.

We can summarize this by:

$$\text{sgn}(p(0)) = -\text{sgn}(x(0))$$

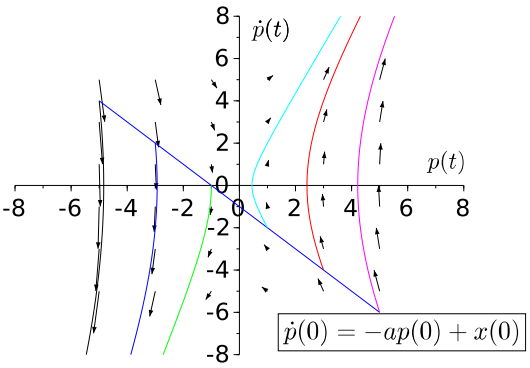


Figure 4.1: Phase portrait of (4.20) - $a = 1$, $b = -0.5$, $d = 1$, $e = -2$, $f = 3$, $x(0) = -1$,

Furthermore, p will always have the same sign on $[0, T]$, so the optimal control u^* given in equation (4.19) always has the same sign on $[0, T]$, and is smooth (since p is smooth). Furthermore, γ will be constant on $[0, T]$. Consequently, we know explicitly the solution $p(t)$ on $[0, T]$, namely:

$$p(t) = \frac{1}{\sqrt{\gamma}} \left[(\sqrt{\gamma} \cosh(\sqrt{\gamma} t) - a \sinh(\sqrt{\gamma} t)) p(0) + \sinh\left(\sqrt{\gamma} t\right) x(0) \right]$$

where cosh and sinh are the hyperbolic cosine and sine functions. In order to have $p(T) = 0$, we must take:

$$p(0) = -\frac{\sinh\left(\sqrt{\gamma} T\right) x(0)}{\left(\sqrt{\gamma} \cosh(\sqrt{\gamma} T) - a \sinh(\sqrt{\gamma} T)\right)}.$$

From that, it is easy to have the expression of the optimal trajectory x, using the fact that

$$x(t) = \dot{p}(t) + ap(t).$$

Conclusion

This approach enabled us to obtain some stationarity results that have a convenient formulation. Furthermore, an analytical result have been also obtained. This will let us design, in the next Chapter, numerical schemes for approximating the optimal solution. Using the analytical solution, we will be able to appreciate the exactness of the solution.

Chapter 5

Numerical simulations

Abstract. Three numerical schemes are presented here: one which is a naive Direct method, and two other Indirect methods, based on the optimality condition derived in Chapter 4. Thanks to the analytical solution on which these schemes are tested, the correctness of the solutions is illustrated. This Chapter has been published in [107].

Obviously, there is little hope to solve analytically all possible problems of the form (4.1)(4.2). Therefore, some numerical schemes are designed in order to obtain a numerical approximations. Following the presentation of Section 2.2, we present here two numerical methods in order to approximate the solution of (4.1)(4.2): a direct method, and an indirect method.

5.1 Numerical schemes

5.1.1 Direct method

A first way to solve the optimal problem is to discretize directly the dynamics (4.1) and the cost (4.2) in order to obtain a constrained optimization problem. In fact, a simple discretization that we can use here is the following one:

$$\min_{u\in\mathbb{R}^N} \sum_{i=0}^{N} f^0(x_i,u_i),$$

$$s.t. \left\{ \begin{array}{ccc} 0 \leq \lambda_k^x & \perp & Cx_{k+1}+d\lambda_k^x+eu_k \geq 0 \\ \frac{x_{k+1}-x_k}{h} & = & Ax_{k+1}+B\lambda_k^x+Fu_k \end{array} \right. k=0...N-1,$$

where the subscript k in z_k denotes the k-th step in the discretization of the variable $z(t)$ at time t_k, and h denotes the (here uniform) time-step. Moreover, we can choose to integrate the dynamics with an implicit (as presented here) or an explicit method. This is a Mathematical Program (MP) constrained by a Mixed Linear Complementarity Problem with parameters $(G(\cdot,u,\cdot),H(\cdot,u,\cdot))$ where $G : \mathbb{R}^{n(N+1)} \times \mathbb{R}^{N+1} \times \mathbb{R}^N_+ \to \mathbb{R}^{nN}$ and $H : \mathbb{R}^{n(N+1)} \times \mathbb{R}^{N+1} \times \mathbb{R}^N_+ \to \mathbb{R}^N$ are defined component-wise by:

$$\begin{aligned} G_k(x,u,\lambda) &= \frac{x_{k+1}-x_k}{h} - Ax_{k+1} - B\lambda_k^x - Fu_k, \\ H_k(x,u,\lambda) &= Cx_{k+1} + d\lambda_k^x + eu_k, \end{aligned}$$

$k = 0, ..., N-1$. We notice that we can isolate x_{k+1} in the dynamics, and reintroduce it in the complementarity conditions:

$$0 \leq \lambda_k^x \perp (d + hC(I - hA)^{-1}B)\lambda_k^x + (e + hC(I - hA)^{-1}F)u_k + C(I - hA)^{-1}x_k \geq 0$$

5.1.2 Indirect method

A second way to compute an approximate solution of the optimal control problem is to discretize (4.11) under the constraints (4.10) and with the condition (4.5). This method is known as the indirect method (since it is using an a priori study of the system, and the result obtained with the Pontryagin equations). Obviously, the choice of the discretization will eventually have an impact on the accuracy and the stability of the numerical solution. So as to have a first idea of the extent of these issues, these equations are discretized with an Euler scheme which will use implicit or explicit terms in its formulation, and we will investigate how these choices affect the solution. We present two formulations, that we name of *explicit* and *implicit* type. As we will see, these two formulations lead to different types of optimization problems.

Explicit type

This first formulation leads to a problem where we can identify two (almost) independent problems at each step. Let us assume we already know variables values of x and p at time t_k, e.g. z_k, and we want to compute the solution at time t_{k+1}. We first solve the following discretization of (4.5) with (4.10):

$$\max_{v \in \mathbb{R}} \{\langle p_k, f^x(x_k, v)\rangle + p^0 f^0(x_k, v)\},$$

$$s.t. \begin{cases} 0 \leq \lambda_k^x \perp Cx_k + d\lambda_k^x + ev \geq 0 \\ 0 \leq \mu_{x,u} \perp \mu_{x,u} - 2(Cx_k + ev) \geq 0 \\ 0 \leq \mu_p \perp \mu_p - 2p_k \geq 0 \\ 0 \leq \lambda_1^{abs} \perp \mu_{x,u}\mathbb{1}_n \geq 0 \\ 0 \leq \lambda_2^{abs} \perp (\mu_{x,u} - 2(Cx_k + ev))\mathbb{1}_n \geq 0 \\ 0 \leq \mu_1 \perp \mu_1 - 2\lambda_{1k}^p \geq 0 \\ 0 \leq \mu_2 \perp \mu_2 - 2\lambda_{2k}^p \geq 0 \\ \lambda_1^{abs} + \lambda_2^{abs} = \mu_p - p_k \\ \lambda_{1k}^p + \lambda_{2k}^p = p_k \\ \mu_1 - \lambda_{1k}^p = \lambda_1^{abs} \\ \mu_2 - \lambda_{2k}^p = \lambda_2^{abs} \end{cases}$$

Solving this problem will give us u_k and the associated Λ_k, which is unique for a given u_k as we have seen from the derivation of system (4.10), except in the case where $Cx_k + eu_k = 0$. We can rewrite the complementarity conditions of this MPEC in the following compact form :

$$0 \leq \Omega_k \perp \Delta\Omega_k + \Psi \geq 0$$

where $\Omega_k = (\lambda_k^x, \mu_{x,u}, \mu_p, \lambda_1^{abs}, \lambda_2^{abs}, \mu_1, \mu_2)^\intercal$,

$$\Delta = \begin{pmatrix} d & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbb{1}_n & 0 & 0 & 0 & 0 \\ 0 & \mathbb{1}_n & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathbb{1}_n & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathbb{1}_n & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbb{1}_n \end{pmatrix} \tag{5.1}$$

and Ψ is easily identifiable. Finally, we just need to integrate the dynamics (4.11), which is integrated with the following discretization:

$$\frac{z_{k+1} - z_k}{h} = \tilde{A} z_{k+1} + \tilde{B}\Lambda_k + \tilde{F} u_k.$$

Here again, we can use an implicit integration, or an explicit one by introducing z_k instead of z_{k+1} in the right-hand side.

Implicit type

This second formulation is expressed in the form of a single MPEC solved at each timestep. Here, every variable will be used *implicitly*, as the dynamics is introduced inside the constraints of the MP. Namely, we need to solve at each step the following MPEC:

$$\max_{v \in \mathbb{R}} \{\langle p_{k+1}, f(x_{k+1}, v)\rangle + p^0 f^0(x_{k+1}, v)\}$$

$$s.t. \begin{cases} 0 \le \lambda_{k+1}^x \perp Cx_{k+1} + d\lambda_{k+1}^x + ev \ge 0 \\ 0 \le \mu_{x,u} \perp \mu_{x,u} - 2(Cx_{k+1} + ev) \ge 0 \\ 0 \le \mu_p \perp \mu_p - 2p_{k+1} \ge 0 \\ 0 \le \lambda_1^{abs} \perp \mu_{x,u}\mathbb{1}_n \ge 0 \\ 0 \le \lambda_2^{abs} \perp (\mu_{x,u} - 2(Cx_{k+1} + ev))\mathbb{1}_n \ge 0 \\ 0 \le \mu_1 \perp \mu_1 - 2\lambda_{1k+1}^p \ge 0 \\ 0 \le \mu_2 \perp \mu_2 - 2\lambda_{2k+1}^p \ge 0 \\ \lambda_1^{abs} + \lambda_2^{abs} = \mu_p - p_{k+1} \\ \lambda_{1k+1}^p + \lambda_{2k+1}^p = p_{k+1} \\ \mu_1 - \lambda_{1k+1}^p = \lambda_1^{abs} \\ \mu_2 - \lambda_{2k+1}^p = \lambda_2^{abs} \\ \frac{z_{k+1}-z_k}{h} = \tilde{A} z_{k+1} + \tilde{B}\Lambda_{k+1} + \tilde{F} v \end{cases}$$

5.2 Numerical results

5.2.1 Direct method

Consider the system given in (4.12). The direct method gives good, whatever the parameters used. We used the solver GAMS (available at `http://www.gams.com/`) which includes a powerful MPEC solver. The only trouble noticed was that some fluctuations around the analytical solution were found (see Figure 5.1 for $t \in [0.5, 1]$), thus the results were difficult to achieve thin precision. Nonetheless, these fluctuations are still admissible in all the calculations we made. Some results are shown in Figure 5.1. The numerical performance of the direct method (including curves showing its order) are further extended in Section 7.1.

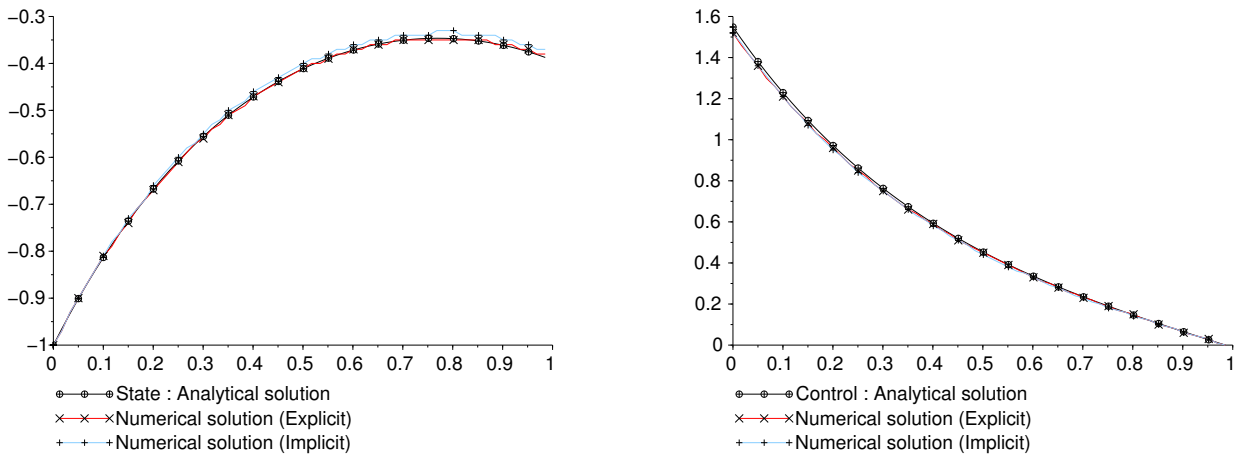


Figure 5.1: Numerical solution: direct approach - $a = 1$, $b = -0.5$, $d = 1$, $e = -2$, $f = 3$, $x(0) = -1$, $N = 60$

5.2.2 Indirect method

The indirect method gives more disappointing results. As shown in Figure 5.2, even with the good initial value $p(0)$, this approach fails to give a good solution, close to the analytical one. Hope for a better solution with this method seems small since in general application, $p(0)$ should be found numerically. Here again, GAMS was used to solve the MPEC at each step. The reason why this is not working still is unknown to us: changing the parameters does not seem to enhance the precision of the numerical solution, nor the reduction of the timestep h. At each time step, the resolution of the MPEC seems to fail, the constraints being often largely violated. A first way to explain this may be found in matrix Δ introduced in (5.1): even in this scalar example, it is of rank 5, when Δ is 7×7. In order to tackle this instability, the Forward-Backward Sweep Method (FBSM) could be applied. This method consists in solving the differential equation on x with p fixed between 0 and T (forward), and then solving the differential equation on p with x fixed between T and 0 (backward), and iterate until the boundary conditions are met. This method has been analysed in [78]. So far, the convergence results relies on restrictive assumptions (such as T being small enough). Also, since the good solution for $p(0)$ is already given, one should not need to converge to the boundary conditions. For all these reasons, this method has not been tested. However, an enhanced version of the indirect method, working properly, is shown in Section 7.2.

Conclusion

The two main families of numerical methods for Optimal Control problems were tested here. It shows that the Direct methods gives satisfactory results, while the Indirect method seem to fail. The formulation of the necessary conditions may be the problem; a different approach, using multipliers, is presented in the subsequent Chapters 6 and 7.

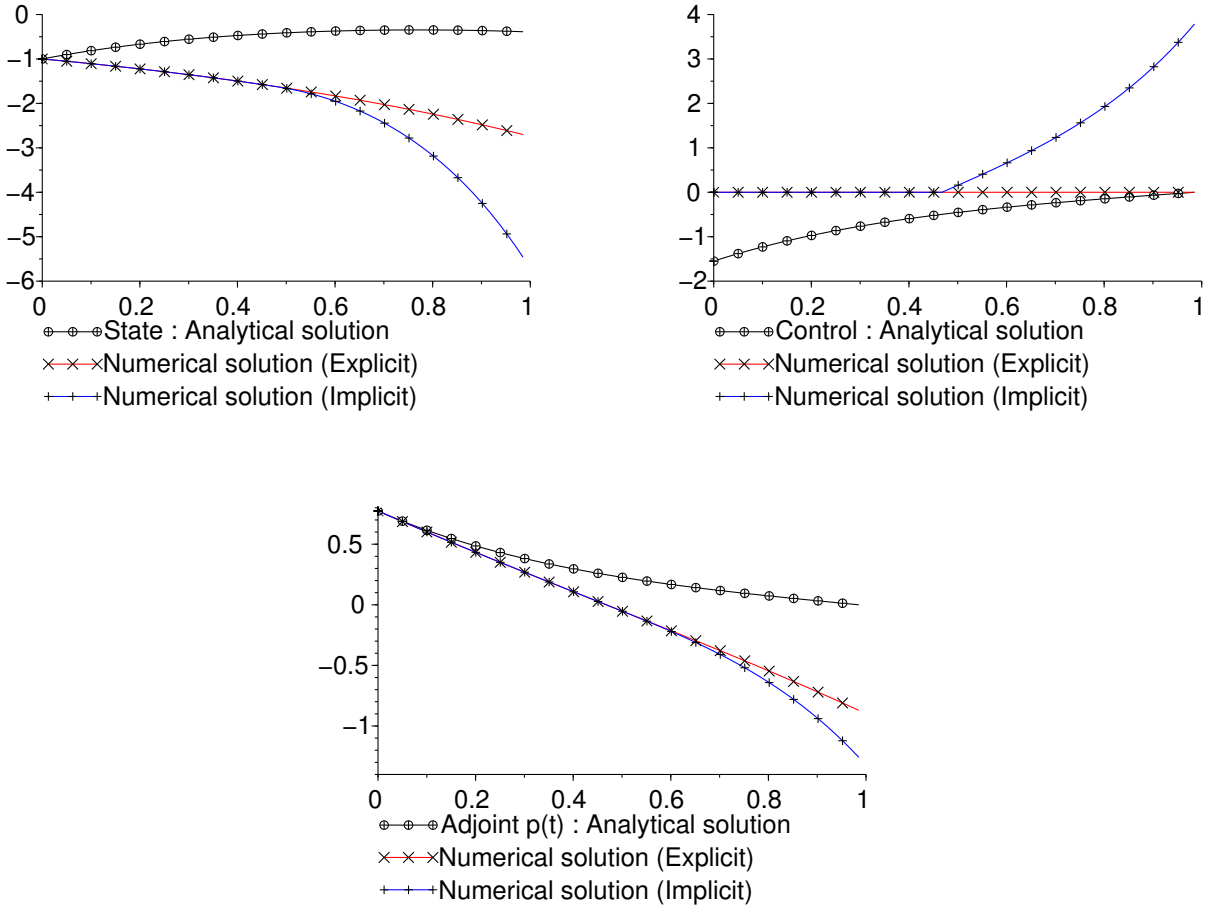


Figure 5.2: Numerical solution: indirect approach - $a = 1$, $b = -0.5$, $d = 1$, $e = -2$, $f = 3$, $x(0) = -1$, $N = 60$

Part III

Quadratic optimal control of LCS: the general case

Chapter 6

Derivation of the first order conditions

Abstract. In this chapter, we will derive necessary conditions for the quadratic optimal control of a general class of LCS. The first results involve index sets that are not convenient. These conditions are therefore re-expressed in the form of an LCS. In a second part, we will also show that the weakest form of stationarity is actually a sufficient condition for optimality. This chapter has been submitted to IEEE Transactions on Automatic Control [108].

The objective of this chapter is to analyze the quadratic optimal control of LCS. More precisely, we wish to investigate properties and numerical resolution of the problem:

$$\min \; J(x,u,v) = \int_0^T (x(t)^\intercal Q x(t) + u(t)^\intercal U u(t) + v(t)^\intercal V v(t))\, dt, \tag{6.1}$$

$$\text{subject to } \begin{cases} \dot{x}(t) = Ax(t) + Bv(t) + Fu(t), \\ w(t) = Cx(t) + Dv(t) + Eu(t), \\ 0 \leq v(t) \perp w(t) \geq 0, \\ Mx(0) + Nx(T) = x_b, \end{cases} \tag{6.2}$$

where $T > 0$, $A, Q \in \mathbb{R}^{n\times n}$, $V, D \in \mathbb{R}^{m\times m}$, $U \in \mathbb{R}^{m_u \times m_u}$, $E \in \mathbb{R}^{m\times m_u}$, $B \in \mathbb{R}^{n\times m}$, $F \in \mathbb{R}^{n\times m_u}$, $C \in \mathbb{R}^{m\times n}$, $M, N \in \mathbb{R}^{2n\times n}$, $x : [0,T] \to \mathbb{R}^n$, $u, v : [0,T] \to \mathbb{R}^m$, $x_b \in \mathbb{R}^{2n}$. In order to avoid trivial cases, we assume that $(C,E) \neq (0,0)$ and x_b is in the image set of $(M \;\; N)$. Also, we choose U symmetric and positive-definite, and Q and V semi-positive definite.

The problem of existence of solutions of the Optimal Control Problem (6.1)(6.2) is actually twofold. First, the existence of a trajectory for the LCS (6.2) is not straightforward, even if the system is expressed as an Initial Value Problem (IVP), as it was explained in Section 1.2. Secondly and most importantly, the existence of solution for (6.1)(6.2) is still an open question. A famous result due to Filipov [34, Theorem 9.2i] states the existence of an optimal control under convexity of the so-called velocity set $\mathcal{V}(x)$. In our case, $\mathcal{V}(x) = \{(u,v) \in \mathbb{R}^{2m} | 0 \leq v \perp Cx + Dv + Eu \geq 0\}$ is clearly not convex, due to the complementarity. *Therefore throughout this chapter, we admit that an optimal solution exists (in the sense of Definition 6.1.2 below)*, and the focus is on necessary conditions this optimal solution must comply with (relying strongly on the seminal work in [59], presented in Section 3.2), together with their numerical computation which relies on MPEC algorithms (see Section 3.1.2).

This section uses normal cones that are defined in the Appendix A.

6.1 First-order necessary conditions for the optimal control problem (6.1)(6.2)

6.1.1 Preliminaries

Let us begin by recalling some definitions given in Sections 3.1 and 3.2 and particularized for the problem (6.1)(6.2). First, let us fix some notations and definitions related to MPEC:

Definition 6.1.1. *Let $m \in \mathbb{N}$.*

- *The* complementarity cone *is defined as* $\mathcal{C}^m = \{(v,w) \in \mathbb{R}^m \times \mathbb{R}^m : 0 \le v \perp w \ge 0\}$.
- *Three different index sets are defined from this complementarity constraint, called the active sets and the degenerate set:*

$$I^{+0}(\bar{v},\bar{w}) = \{i \in \overline{m} : \bar{v} > 0 = \bar{w}\},$$

$$I^{0+}(\bar{v},\bar{w}) = \{i \in \overline{m} : \bar{v} = 0 < \bar{w}\},$$

$$I^{00}(\bar{v},\bar{w}) = \{i \in \overline{m} : \bar{v} = 0 = \bar{w}\}.$$

The sets $I^{\bullet 0}(\bar{v},\bar{w})$ and $I^{0\bullet}(\bar{v},\bar{w})$ are defined as $I^{\bullet 0}(\bar{v},\bar{w}) = I^{+0}(\bar{v},\bar{w}) \cup I^{00}(\bar{v},\bar{w})$, $I^{0\bullet}(\bar{v},\bar{w}) = I^{0+}(\bar{v},\bar{w}) \cup I^{00}(\bar{v},\bar{w})$.

Definition 6.1.2. *Let $n, m, m_u \in \mathbb{N}$.*

- *We refer to any absolutely continuous function on $[0,T]$ as an arc, and to any measurable function on $[0,T]$ as a control.*
- *An* admissible pair *for (6.2) is a 3-tuple of functions (x,u,v) on $[0,T]$ for which u,v are controls and x is an arc, that satisfy all the constraints in (6.2).*
- *Let us define the constraint set S, by*

$$S = \{(x,u,v) \in \mathbb{R}^n \times \mathbb{R}^{m_u} \times \mathbb{R}^m \ : \ (v, Cx + Dv + Eu) \in \mathcal{C}^m\}\,.$$

- *Given a constant $R > 0$, we say that an admissible pair (x^*,u^*,v^*) is a* local minimizer of radius R *for (6.1)(6.2) if there exists $\varepsilon > 0$ such that for every pair (x,u,v) admissible for (6.2), which also satisfies $\|x(t) - x^*(t)\| \le \varepsilon$, $\left\|\binom{u}{v} - \binom{u^*(t)}{v^*(t)}\right\| \le R$ a.e. $t \in [0,T]$ and $\int_0^T \|\dot{x}(t) - \dot{x}^*(t)\|dt \le \varepsilon$, we have $J(x^*,u^*,v^*) \le J(x,u,v)$.*
- *For every given $t \in [0,T]$, and constant scalars $\varepsilon > 0$ and $R > 0$, we define the neighborhood of the point $(x^*(t),u^*(t),v^*(t))$ as*

$$S_*^{\varepsilon,R}(t) = \left\{(x,u,v) \in S : \ \|x - x^*(t)\| \le \varepsilon, \ \left\|\begin{pmatrix}u\\v\end{pmatrix} - \begin{pmatrix}u^*(t)\\v^*(t)\end{pmatrix}\right\| \le R\right\}.$$

- *The dependence on time of index sets is denoted as $I_t^{0+}(x,u,v) = \{i \in \overline{m} : v_i(t) > 0 = (Cx(t) + Dv(t) + Eu(t))_i\}$. The same definition follows for $I_t^{+0}(x,u,v)$, $I_t^{00}(x,u,v)$, $I_t^{\bullet 0}(x,u,v)$, $I_t^{0\bullet}(x,u,v)$.*

- *For a positive measurable function k_S defined for almost every $t \in [0,T]$,* the bounded slope condition *is defined as the following implication:*

$$(x,u,v) \in S_*^{\varepsilon,R}(t), (\alpha,\beta,\gamma) \in \mathcal{N}_S^P(x,u,v) \implies \|\alpha\| \leq k_S(t) \left\| \begin{pmatrix} \beta \\ \gamma \end{pmatrix} \right\|. \tag{6.3}$$

As exposed in Section 3.2, the bounded slope condition (6.3) does not trivially hold. The next Proposition gives a sufficient condition for the LCP (6.2) to comply with it.

Proposition 6.1.1. *Suppose $E \neq 0$ and $\mathrm{im}(C) \subseteq \mathrm{im}(E)$. Then the local error bound condition (see Definition 3.1.2) holds at every admissible point, and the bounded slope condition for (6.2) holds.*

In order to prove this, one needs first the following lemma:

Lemma 6.1.1. *Let $A \in \mathbb{R}^{n\times m}$, $B \in \mathbb{R}^{m_u \times m}$ such that $B \neq 0$ and $\ker(B) \subseteq \ker(A)$. Then there exists $\alpha > 0$ such that: $\forall x \in \mathbb{R}^m$, $\|Ax\| \leq \alpha \|Bx\|$.*

Proof. Let $x \in \mathbb{R}^m$. We decompose x in the following way:

$$x = x_B + x_A + x_\perp,\ x_B \in \ker B,\ x_A \in \ker A \backslash \ker B \cup \{0\},\ x_\perp \in (\ker A)^\perp$$

such that $Ax = Ax_\perp$, $Bx = B(x_A + x_\perp)$. Thus:

$$\begin{aligned} \|Ax\|^2 &\leq \|A\|^2 \|x_\perp\|^2 \\ &\leq \|A\|^2 \left[\|x_\perp\|^2 + \|x_A\|^2\right] \\ &\leq \|A\|^2 \|x_A + x_\perp\|^2 \text{ because } x_A \perp x_\perp \end{aligned}$$

However, $x_A + x_\perp \in (\ker B)^\perp$. Therefore, the linear application $B : (\ker B)^\perp \to \mathrm{im} B$ (we write the same way the linear application and the associated matrix in the canonic basis) is an isomorphism, which admits an inverse $B^\dagger$, which is the Moore-Penrose inverse. It yields: $\exists! y \in \mathrm{im} B;\ B(x_A + x_\perp) = y \iff x_A + x_\perp = B^\dagger y$, and thus

$$\|x_A + x_\perp\| \leq \|B^\dagger\| \|y\| = \|B^\dagger\| \|B(x_A + x_\perp)\|.$$

We can then prove that:

$$\begin{aligned} \|Ax\|^2 &\leq \|A\|^2 \|x_A + x_\perp\|^2 \\ &\leq \|A\|^2 \|B^\dagger\|^2 \|B(x_A + x_\perp)\|^2 \\ &\leq \|A\|^2 \|B^\dagger\|^2 \|Bx\|^2 \end{aligned}$$

It easily yields the desired result. □

Proof of Proposition 6.1.1. Since the MPEC Linear Condition holds (see Definition 3.1.1), the local error bound condition also holds at every admissible points (see [59, Proposition 2.3]). Applying Proposition 3.2.1, a sufficient condition for the bounded slope condition to hold is:

$$\forall \lambda^H, \lambda^G \in \mathbb{R}^m,\ \|C^\intercal \lambda^H\| \leq k_S(t) \left\| \begin{matrix} \lambda^G + D^\intercal \lambda^H \\ E^\intercal \lambda^H \end{matrix} \right\|$$

Since $\mathrm{im}(C) \subseteq \mathrm{im}(E)$ (or equivalently, $\ker(E^\intercal) \subseteq \ker(C^\intercal)$), applying Lemma 6.1.1 with $C^\intercal$ and $E^\intercal$, we prove that:

$$\exists \alpha > 0,\ \forall \lambda^H \in \mathbb{R}^m,\ \|C^\intercal \lambda^H\|^2 \leq \alpha^2 \|E^\intercal \lambda^H\|^2$$

Therefore, it easily proves:

$$\exists \alpha > 0,\ \forall \lambda^G, \lambda^H \in \mathbb{R}^m,\ \|C^\intercal \lambda^H\|^2 \leq \alpha^2 \left\| \begin{matrix} \lambda^G + D^\intercal \lambda^H \\ E^\intercal \lambda^H \end{matrix} \right\|^2$$

Therefore, the sufficient condition for the bounded slope condition holds. □

6.1.2 Necessary first-order conditions

Let us now apply [59, Theorem 3.2], recalled in Theorem 3.2.1, to the Problem (6.1)(6.2).

Proposition 6.1.2. *Let (x^*, u^*, v^*) be a local minimizer of constant radius $R > 0$ for (6.1)(6.2). Suppose $\mathrm{im}(C) \subseteq \mathrm{im}(E)$. Then there exist an arc $p : [0, T] \to \mathbb{R}^n$, a scalar $\lambda_0 \leq 0$ and measurable functions $\lambda^G : \mathbb{R} \to \mathbb{R}^m$, $\lambda^H : \mathbb{R} \to \mathbb{R}^m$ such that the following conditions hold:*

1. *The non-triviality condition: $(\lambda_0, p(t)) \neq 0$, $\forall t \in [0, T]$.*
2. *The transversality condition: $\begin{pmatrix} p(0) \\ -p(T) \end{pmatrix} \in \mathrm{im} \begin{pmatrix} M^\intercal \\ N^\intercal \end{pmatrix}$.*
3. *The Euler adjoint equation: for almost every $t \in [0, T]$,*

$$\begin{aligned} \dot{p}(t) &= -A^\intercal p - 2\lambda_0 Q x^* - C^\intercal \lambda^H \\ 0 &= F^\intercal p + 2\lambda_0 U u^* + E^\intercal \lambda^H \\ 0 &= B^\intercal p + 2\lambda_0 V v^* + \lambda^G + D^\intercal \lambda^H \\ 0 &= \lambda_i^G(t), && \forall i \in I_t^{+0}(x^*, u^*, v^*) \\ 0 &= \lambda_i^H(t), && \forall i \in I_t^{0+}(x^*, u^*, v^*). \end{aligned} \tag{6.4}$$

4. *The Weierstrass condition for radius R: for almost every $t \in [t_0, t_1]$,*

$$(x^*(t), u, v) \in S,\ \left\| \begin{pmatrix} u \\ v \end{pmatrix} - \begin{pmatrix} u^*(t) \\ v^*(t) \end{pmatrix} \right\| < R$$

$$\begin{aligned} \implies \langle p(t), Ax^*(t) + Bv^*(t) + Fu^*(t)) \rangle + \lambda_0 \left(x^*(t)^\intercal Q x^*(t) + u^*(t)^\intercal U u^*(t) + v^*(t)^\intercal V v^*(t) \right) \\ \geq \langle p(t), Ax^*(t) + Bv + Fu) \rangle + \lambda_0 \left(x^*(t)^\intercal Q x^*(t) + u^\intercal U u + v^\intercal V v \right). \end{aligned} \tag{6.5}$$

Proof. Let us check that the problem complies with [59, Assumption 3.1]. These assumptions are recalled in Assumption 3.2.1 in Section 3.2. We check these in the same order:

1. Let $t \in [0, T]$ and $(x_1, u_1, v_1), (x_2, u_2, v_2) \in S_*^{\varepsilon, R}(t)$. First, let us check (3.10)(a):

$$\|(Ax_1 + Bv_1 + Fu_1) - (Ax_2 + Bv_2 + Fu_2)\| \leq \|A\|\|x_1 - x_2\| + \|B\|\|v_1 - v_2\| + \|F\|\|u_1 - u_2\|.$$

Secondly, we must check the inequality concerning the cost in (3.10)(b). For that, remark first that:

$$\begin{aligned} |x_1^\intercal Q x_1 - x_2^\intercal Q x_2| &= |(x_1+x_2)^\intercal Q(x_1-x_2)| \\ &\leq \|x_1 - x^*(t) + x_2 - x^*(t) + 2x^*(t)\|\|Q\|\|x_1-x_2\| \\ &\leq (\|x_1-x^*(t)\| + \|x_2-x^*(t)\| + 2\|x^*(t)\|)\,\|Q\|\|x_1-x_2\| \\ &\leq 2\|Q\|\,(\|x^*(t)\|+\varepsilon)\,\|x_1-x_2\|. \end{aligned}$$

Similarly, one proves that $|u_1^\intercal U u_1 - u_2^\intercal U u_2| \leq 2\|U\|(\|u^*(t)\|+R)\|u_1-u_2\|$ and $|v_1^\intercal V v_1 - v_2^\intercal V v_2| \leq 2\|V\|(\|v^*(t)\|+R)\|v_1-v_2\|$. Therefore:

$$\begin{aligned} |(x_1^\intercal Q x_1 + u_1^\intercal U u_1) - (x_2^\intercal Q x_2 + u_2^\intercal U u_2)| &\leq |x_1^\intercal Q x_1 - x_2^\intercal Q x_2| + |u_1^\intercal U u_1 - u_2^\intercal U u_2| \\ &\leq k_x(t)\|x_1-x_2\| + k_u(t)\|u_1-u_2\|. \end{aligned}$$

where $k_x(t) = 2\|Q\|\,(\|x^*(t)\|+\varepsilon)$ and $k_u(t) = 2\|U\|(\|u^*(t)\|+R)$. k_x, k_u are measurable functions of time, and $\|A\|$, $\|B\|$ and $\|F\|$ are all constant and therefore measurable functions. Thus (3.10) holds true.

2. Since $\mathrm{im}(C) \subseteq \mathrm{im}(E)$, and using Lemma 6.1.1, the bounded slope condition and the error bound condition for the system (6.2) holds at $(x^*(t), u^*(t), v^*(t))$ for all $t \in [0,T]$ hold, with a positive constant k_S.

3. The terms $k_S[\|B\| + \|F\| + k_u]$, k_x and $\|A\|$ are all integrable on $[0,T]$, and there obviously exists a positive number η such that $R \geq \eta k_S$ on $[0,T]$ (just take $\eta = R/k_S$).

4. Since all involved functions are smooth, all conditions of measurability and differentiability are met.

Calculations of the non-triviality and Weierstrass conditions are straightforward. Since all functions are differentiable, the Clarke subdifferential in (3.14) contains only the gradient, i.e.

$$\nabla_{x,u,v}\left(\langle p(t), Ax + Bv + Fu\rangle - \lambda_0(x^\intercal Q x + u^\intercal U u)\right),$$

and $U(\cdot)$ is in our case the whole space $\mathbb{R}^{2m}$, so the normal cone reduces to $\{0\}$. Simple calculations from (3.14) yield the Euler equation (6.4). Concerning the transversality condition (3.13), notice first that we did not impose any boundary cost. Denote $P_b = \left\{\binom{x_0}{x_T} : Mx_0 + Nx_T = x_b\right\}$, then for $\binom{x_0}{x_T} \in P_b$, since P_b is an affine vector space,

$$\mathcal{N}_{P_b}\begin{pmatrix} x_0 \\ x_T \end{pmatrix} = -\left(P_b - \begin{pmatrix} x_0 \\ x_T \end{pmatrix}\right)^* = -\ker(M\ N)^* = \mathrm{im}\begin{pmatrix} M^\intercal \\ N^\intercal \end{pmatrix}.$$

□

Remark 6.1.1. • *The tuple consisting of a trajectory and the associated multipliers solution of (6.4) is called an extremal. The case $\lambda_0 = 0$ is often called the abnormal case [59], and the corresponding extremal an abnormal extremal. In this case, no information can be derived from these necessary conditions. In other cases, we can choose this value most conveniently, since the adjoint state p is defined up to a multiplicative positive constant. In the rest of this paper, λ_0 will always be chosen as $-\frac{1}{2}$. The optimal trajectory is normal when, for instance, the initial point $x(0)$ or the final point $x(T)$ are free.*

- *In (6.5), all parts containing $x^*(t)$ can actually be subtracted from each side of the inequality.*

The Weierstrass condition (6.5) can be re-expressed as searching a local maximizer of the following MPEC:

$$\begin{aligned} &\max_{u,v} \ \langle p(t), Bv + Fu\rangle + \lambda_0 \left(u^\intercal U u\right) \\ &\text{s.t. } \ 0 \le v \perp Cx^*(t) + Dv + Eu \ge 0. \end{aligned} \tag{6.6}$$

For each $t \in [0,T]$, this is an MPEC, as presented in Section 3.1. These programs admit first-order conditions, the weak and strong stationarity: this motivates the next definition.

Definition 6.1.3. *Let (x^*, u^*, v^*) be an admissible pair for (6.2). Then:*

- *The FJ-type W(eak)-stationarity holds at (x^*, u^*, v^*) if there exist an arc p, a scalar $\lambda_0 \le 0$ and measurable functions λ^G, λ^H such that Proposition 6.1.2 (1)-(4) hold.*
- *The FJ-type S(trong)-stationarity holds at (x^*, u^*, v^*) if (x^*, u^*, v^*) is FJ-type W-stationary with arc p and there exist measurable functions η^G, η^H such that, for almost every $t \in [0,T]$,*

$$\begin{aligned} 0 &= F^\intercal p + 2\lambda_0 U u^* + E^\intercal \eta^H \\ 0 &= B^\intercal p + 2\lambda_0 V v^* + \eta^G + D^\intercal \eta^H \\ 0 &= \eta_i^G(t), && \forall i \in I_t^{+0}(x^*, u^*, v^*) \\ 0 &= \eta_i^H(t), && \forall i \in I_t^{0+}(x^*, u^*, v^*), \end{aligned}$$

$$\eta_i^G(t) \ge 0, \ \ \eta_i^H(t) \ge 0, \ \ \forall i \in I_t^{00}(x^*, u^*, v^*).$$

- *We simply call W-stationarity or S-stationarity the FJ-type W- or S-stationarity with $\lambda_0 = -\frac{1}{2}$.*

The multipliers η^G, η^H can be different in measure from the corresponding λ^G, λ^H in Proposition 6.1.2. The next theorem, whose proof follows directly from [59, Theorem 3.6] as recalled in Theorem 3.2.3, addresses this problem:

Theorem 6.1.1. *Let (x^*, u^*, v^*) be a local minimizer of radius R for (6.1)(6.2). Suppose that for almost every $t \in [0,T]$, the MPEC LICQ holds at $(u^*(t), v^*(t))$ for problem (6.6), i.e., the family of gradients*

$$\left\{ \begin{pmatrix} 0 \\ e_i \end{pmatrix} : i \in I_t^{0\bullet}(x^*, u^*, v^*) \right\} \bigcup \left\{ \begin{pmatrix} (E_{i\bullet})^\intercal \\ (D_{i\bullet})^\intercal \end{pmatrix} : i \in I_t^{\bullet 0}(x^*, u^*, v^*) \right\}. \tag{6.7}$$

is linearly independent, where e_i is a vector such that its j-th component is equal to δ_i^j, the Kronecker delta. Then the S-stationarity holds at (x^*, u^*, v^*). Moreover, in this case, the multipliers η^G, η^H can be taken equal to λ^G, λ^H, respectively, almost everywhere.

We can now state the following result:

Corollary 6.1.1. *Suppose $m = m_u$ and E is invertible. Then the local minimum (x^*, u^*, v^*) is S-stationary, and the multipliers η^G, η^H can be chosen equal to λ^G, λ^H almost everywhere.*

Proof. Let us first show that E invertible is equivalent to $\operatorname{rank}\begin{pmatrix} 0_m & E^\intercal \\ I_m & D^\intercal \end{pmatrix} = 2m$. Notice that $\operatorname{rank}\begin{pmatrix} 0_m & E^\intercal \\ I_m & D^\intercal \end{pmatrix} = \operatorname{rank}\begin{pmatrix} 0_m & I_m \\ E & D \end{pmatrix}$. Since $\begin{pmatrix} 0 \\ E \end{pmatrix}$ and $\begin{pmatrix} I_m \\ D \end{pmatrix}$ are linearly independent, we have:

$$\operatorname{rank}\begin{pmatrix} 0_m & I_m \\ E & D \end{pmatrix} = \operatorname{rank}\begin{pmatrix} 0_m \\ E \end{pmatrix} + \operatorname{rank}\begin{pmatrix} I_m \\ D \end{pmatrix} = \operatorname{rank}(E) + m.$$

Thus, if E is invertible, then $\operatorname{rank}\begin{pmatrix} 0_m & E^\intercal \\ I_m & D^\intercal \end{pmatrix} = 2m$. Conversely, if $\operatorname{rank}\begin{pmatrix} 0_m & E^\intercal \\ I_m & D^\intercal \end{pmatrix} = 2m$, then $\operatorname{rank}(E) = m$, so E is invertible.

Let us now prove the corollary. Since E is invertible, $\operatorname{rank}\begin{pmatrix} 0_m & E^\intercal \\ I_m & D^\intercal \end{pmatrix} = 2m$.This rank condition ensures the fact that the family

$$\left\{ \begin{pmatrix} 0 \\ e_i \end{pmatrix} : 1 \leq i \leq m \right\} \bigcup \left\{ \begin{pmatrix} (E_{i\bullet})^\intercal \\ (D_{i\bullet})^\intercal \end{pmatrix} : 1 \leq i \leq m \right\}$$

is linearly independent. The family in (6.7) being a subfamily of this one, it is necessarily linearly independent. So the MPEC LICQ holds at $(u^*(t), v^*(t))$ for problem (6.6) (see Definition 3.1.1), and (x^*, u^*, v^*) is S-stationary. □

Remark 6.1.2. *It is actually sufficient to suppose that rank$(E) = m$. In this case, E is not necessarily invertible.*

Let us now state a result that allows us to reformulate the S-stationarity conditions through a complementarity system in order to remove the active sets. One can simply see it that way: for almost all $t \in [0, T]$, the conditions on the multipliers λ^H and λ^G are:

$$\begin{aligned} &\lambda_i^G(t) = 0, \ \forall i \in I_t^{+0}(x, u, v) \\ &\lambda_i^H(t) = 0, \ \forall i \in I_t^{0+}(x, u, v) \\ &\lambda_i^G(t) \geq 0, \ \lambda_i^H(t) \geq 0, \ \forall i \in I_t^{00}(x, u, v). \end{aligned} \tag{6.8}$$

The presence of the active and degenerate sets is tedious, since they depend on the optimal solution, not in a useful way. Nonetheless, the conditions in (6.8) look almost like a linear complementarity problem. The only thing missing is the sign of λ_i^G for $i \in I_t^{0+}(x, u, v)$ (and the same thing with λ_i^H on $I_t^{+0}(x, u, v)$). On these index sets, the multipliers could be negative. But we could for instance create new variables, say α and β, that will both be non-negative and comply with these conditions. This is the purpose of the next Proposition.

Proposition 6.1.3. *Suppose $m = m_u$ and (x, u, v) is an S-stationary trajectory. Then there exist measurable functions $\beta : [0, T] \to \mathbb{R}^m$, $\zeta : [0, T] \to \mathbb{R}$ such that:*

$$u(t) = U^{-1}\left(F^\intercal p(t) + E^\intercal \beta(t) - \zeta(t) E^\intercal v(t)\right)$$

and

$$\begin{pmatrix} \dot{x} \\ \dot{p} \end{pmatrix} = \begin{pmatrix} A & FU^{-1}F^\intercal \\ Q & -A^\intercal \end{pmatrix} \begin{pmatrix} x \\ p \end{pmatrix} + \begin{pmatrix} B - \zeta FU^{-1}E^\intercal \\ \zeta C^\intercal \end{pmatrix} v + \begin{pmatrix} FU^{-1}E^\intercal \\ -C^\intercal \end{pmatrix} \beta$$

$$0 \leq \begin{pmatrix} v \\ \beta \end{pmatrix} \perp \begin{pmatrix} D - \zeta EU^{-1}E^\intercal & EU^{-1}E^\intercal \\ D - \zeta EU^{-1}E^\intercal & EU^{-1}E^\intercal \end{pmatrix} \begin{pmatrix} v \\ \beta \end{pmatrix} + \begin{pmatrix} C & EU^{-1}F^\intercal \\ C & EU^{-1}F^\intercal \end{pmatrix} \begin{pmatrix} x \\ p \end{pmatrix} \geq 0$$

$$0 \leq v \perp \ \zeta(D + D^\intercal + V - \zeta EU^{-1}E^\intercal)v + (\zeta EU^{-1}E^\intercal - D^\intercal)\beta + (\zeta EU^{-1}F^\intercal - B^\intercal)p + \zeta Cx \geq 0. \tag{6.9}$$

To prove this Proposition, we first need the following Lemma.

Lemma 6.1.2. *Let (x, u, v) be an S-stationary trajectory, and λ^G, λ^H be the associated multipliers. Then there exists a measurable function $\zeta : [0, T] \to \mathbb{R}$ such that $\begin{pmatrix} \lambda^G(t) + \zeta(t)w(t) \\ \lambda^H(t) + \zeta(t)v(t) \end{pmatrix} \geq 0$, where w is defined in (6.2).*

Proof. First, remark that, for all $t \in [0, T]$, a candidate $\zeta(t)$ has been defined in (3.6), Theorem 3.1.2 in Section 3.1. Denote $F : [0, T] \times \mathbb{R} \to \mathbb{R}^{2m}$, $F(t, \zeta) = \begin{pmatrix} \lambda^G(t) + \zeta w(t) \\ \lambda^H(t) + \zeta v(t) \end{pmatrix}$. F is a Carathéodory mapping, since λ^G, λ^H, v and w are measurable, so $F(\cdot, \zeta)$ is measurable for each fixed $\zeta \in \mathbb{R}$, and $F(t, \cdot)$ is an affine function, and as such it is continuous, for each fixed t. By the Implicit Measurable Function Theorem [90, Theorem 14.16], there exists a measurable function $\zeta : [0, T] \to \mathbb{R}$ such that $F(t, \zeta(t)) \in \mathbb{R}^{2m}_+$, which is the intended result. □

Proof of Proposition 6.1.3. As proved in Lemma 6.1.2, there exists a measurable function $\zeta : [0, T] \to \mathbb{R}$ such that:

$$\begin{pmatrix} \lambda^G(t) + \zeta(t)w(t) \\ \lambda^H(t) + \zeta(t)v(t) \end{pmatrix} \geq 0.$$

Define $\alpha, \beta : [0, T] \to \mathbb{R}^m$ as $\alpha = \lambda^G + \zeta w$, $\beta = \lambda^H + \zeta v$. The variables α and β are, by construction, measurable and non-negative. From the fact that (x, u, v) is an S-stationary trajectory, we also have that, for almost every $t \in [0, T]$, $\lambda^G_i(t)v_i(t) = 0$ and $\lambda^H_i(t)w_i(t) = 0$ for all $i \in \overline{m}$. Therefore, we can deduce that:

$$\begin{cases} \lambda^G = \alpha - \zeta w \\ \lambda^H = \beta - \zeta v \\ 0 \leq \alpha \perp v \geq 0 \\ 0 \leq \beta \perp w \geq 0. \end{cases} \tag{6.10}$$

In (6.4), let us isolate u, using U symmetric positive definite. Inserting the redefinition of λ^H yields:

$$u(t) = U^{-1}(F^\intercal p(t) + E^\intercal \lambda^H(t)) = U^{-1}\left(F^\intercal p(t) + E^\intercal \beta(t) - \zeta(t)E^\intercal v(t)\right). \tag{6.11}$$

Recall that $w = Cx + Dv + Eu$. Inserting this u in (6.10), we obtain:

$$\lambda^G = \alpha - \zeta(Cx + Dv + Eu) = \alpha - \zeta(Cx + (D - \zeta EU^{-1}E^\intercal)v + EU^{-1}F^\intercal p + EU^{-1}E^\intercal \beta)).$$

Inserting (6.10) and (6.11) into (6.2) and (6.4) allows us to rewrite the differential equations defining x and p as :

$$\begin{cases} \dot{x} = Ax + Bv + Fu = Ax + FU^{-1}F^\intercal p + (B - \zeta FU^{-1}E^\intercal)v + FU^{-1}E^\intercal \beta \\ \dot{p} = -A^\intercal p + Qx - C^\intercal \lambda^H = -A^\intercal p + Qx + \zeta C^\intercal v - C^\intercal \beta. \end{cases}$$

The only equation left is the third equation in (6.4). Replacing λ^G and λ^H with the expressions (6.10) yields:

$$B^\intercal p - Vv + \alpha - \zeta\left(Cx + (D - \zeta EU^{-1}E^\intercal)v + EU^{-1}F^\intercal p + EU^{-1}E^\intercal \beta\right) + D^\intercal(\beta - \zeta v) = 0$$

$$\implies \alpha = (\zeta EU^{-1}F^\intercal - B^\intercal)p + \zeta Cx + (\zeta EU^{-1}E^\intercal - D^\intercal)\beta + \zeta(D + D^\intercal + V - \zeta EU^{-1}E^\intercal)v.$$

Replacing α and u in the complementarity conditions appearing in (6.2) and in (6.10) yields the complementarity conditions in (6.9). □

Remark 6.1.3.

- *The decomposition of (λ^G, λ^H) into (α, β, ζ) proposed in (6.10) is not unique, as it has been hinted in Remark 3.1.1. There actually is a single degree of freedom. Indeed, if this decomposition works for (α, β, ζ), then for any $\rho \geq 0$, we can decompose (λ^G, λ^H) as $(\alpha+\rho w, \beta+\rho v, \zeta+\rho)$. Therefore, for a fixed $t \in [0, T]$, any scalar greater than $\zeta(t)$ is suitable. Thus, if we can find an upper-bounded function ζ decomposing (λ^G, λ^H) into (α, β, ζ), then (λ^G, λ^H) can be decomposed into $(\bar{\alpha}, \bar{\beta}, \bar{\zeta})$, where $\bar{\zeta}$ is a constant along $[0, T]$ greater or equal to the supremum of ζ.*

- *A second remark concerns the three complementarity conditions defining β and v in (6.9). It is not written as a classical Variational Inequality (VI), since it involves $2m$ unknowns but $3m$ complementarity problems. The next proposition addresses this problem.*

Proposition 6.1.4. *Let r be any given positive scalar. Denote (P) the complementarity conditions appearing in (6.9), and denote (P_r) the problem:*

$$\begin{cases} 0 \leq \beta + rv \perp (D - \zeta E U^{-1} E^\intercal)v + E U^{-1} E^\intercal \beta + E U^{-1} F^\intercal p + Cx \geq 0 \\ 0 \leq v \perp \zeta(D + D^\intercal + V - \zeta E U^{-1} E^\intercal)v + (\zeta E U^{-1} E^\intercal - D^\intercal)\beta + (\zeta E U^{-1} F^\intercal - B^\intercal)p + \zeta Cx \geq 0 \\ \beta \geq 0. \end{cases} \tag{P_r}$$

Then (v, β) is a solution of (P) if and only if it is a solution of (P_r).

Proof. We rewrite more simply the two problems as follows:

$$\begin{cases} 0 \leq v \perp & \tilde{D}v + \tilde{U}\beta + q_1 \geq 0 \quad (P1) \\ 0 \leq \beta \perp & \tilde{D}v + \tilde{U}\beta + q_1 \geq 0 \quad (P2) \\ 0 \leq v \perp & \tilde{D}_2 v + \tilde{U}_2 \beta + q_2 \geq 0 \quad (P3) \end{cases} \qquad \begin{cases} 0 \leq \beta + rv \perp & \tilde{D}v + \tilde{U}\beta + q_1 \geq 0 \quad (P_r1) \\ \beta \geq 0 & \quad (P_r2) \\ 0 \leq v \perp & \tilde{D}_2 v + \tilde{U}_2 \beta + q_2 \geq 0 \quad (P_r3) \end{cases}$$

where $\tilde{D} = (D - \zeta E U^{-1} E^\intercal)$, $\tilde{U} = E U^{-1} E^\intercal$, $\tilde{D}_2 = \zeta(D + D^\intercal + V - \zeta E U^{-1} E^\intercal)$, $\tilde{U}_2 = (\zeta E U^{-1} E^\intercal - D^\intercal)$, $q_1 = E U^{-1} F^\intercal p + Cx$, $q_2 = (\zeta E U^{-1} F^\intercal - B^\intercal)p + \zeta Cx$.

- Let (v, β) be a solution of (P). Denote:

$$I^{0+} = \{i : v_i = \beta_i = 0,\ (\tilde{D}v + \tilde{U}\beta + q_1)_i > 0\},$$

$$I^{\bullet 0} = \{i : (\tilde{D}v + \tilde{U}\beta + q_1)_i = 0\}.$$

  These two sets form a partition of $\{1, ..., m\}$. Since CP $(P3)$ and (P_r3) are the same problem, (v, β) is also solution of (P_r3). Using $(P2)$, we find that β complies with (P_r2). We are just left with (P_r1). By assumption it follows that $\forall i \in I^{0+}, \beta_i + rv_i = 0,\ (\tilde{D}v + \tilde{U}\beta + q_1)_i > 0$ and $\forall i \in I^{\bullet 0}, \beta_i + rv_i \geq 0,\ (\tilde{D}v + \tilde{U}\beta + q_1)_i = 0$. So (v, β) is also a solution of (P_r1). This proves that (v, β) is a solution of (P_r).

- Conversely, let (v, β) be a solution of (P_r). Since it is a solution of (P_r1), denote $I_r^{0+} = \{i : \beta_i + rv_i = 0,\ (\tilde{D}v + \tilde{U}\beta + q_1)_i > 0\}$ and $I_r^{\bullet 0} = \{i : (\tilde{D}v + \tilde{U}\beta + q_1)_i = 0\}$. These two sets form a partition of $\{1, ..., m\}$. Since CP (P_r3) and $(P3)$ are the same problem, (v, β) is also solution of $(P3)$. For all $i \in I_r^{0+}$, $\beta_i + rv_i = 0$ and $(\tilde{D}v + \tilde{U}\beta + q_1)_i > 0$. Thanks to (P_r3) and (P_r2), we know that $\beta_i \geq 0$, $v_i \geq 0$. Since $r > 0$, we have a sum of positive terms that must equal 0, so $\beta_i = v_i = 0$. For all $i \in I_r^{\bullet 0}$, $(\tilde{D}v + \tilde{U}\beta + q_1)_i = 0$ and using (P_r3) and (P_r2), $\beta_i \geq 0$, $v_i \geq 0$. So (v, β) is also a solution of $(P1)$ and $(P2)$. It proves that (v, β) is a solution of (P).

□

Let us define $\tilde{\beta} = \beta + rv$ and replace β in (P_r). Thus we end up with the following LCP and inequality constraints:

$$\begin{cases} 0 \leq \begin{pmatrix} \tilde{\beta} \\ v \end{pmatrix} \perp \begin{pmatrix} EU^{-1}E^\intercal & D - (\zeta + r)EU^{-1}E^\intercal \\ \zeta EU^{-1}E^\intercal - D^\intercal & \zeta(D+V) + (\zeta + r)\left(D^\intercal - \zeta EU^{-1}E^\intercal\right) \end{pmatrix} \begin{pmatrix} \tilde{\beta} \\ v \end{pmatrix} \\ \qquad\qquad + \begin{pmatrix} EU^{-1}F^\intercal p + Cx \\ (\zeta EU^{-1}F^\intercal - B^\intercal)p + \zeta Cx \end{pmatrix} \geq 0 \\ \tilde{\beta} \geq rv. \end{cases}$$

To sum up, by Propositions 6.1.2, 6.1.3 and 6.1.4, the following theorem holds:

Theorem 6.1.2. *Let (x^*, u^*, v^*) be a local minimizer of constant radius $R > 0$ for (6.1)(6.2). Suppose $m = m_u$, E is invertible and (x^*, u^*, v^*) is not the projection of an abnormal extremal. Then there exist an arc $p : [0, T] \to \mathbb{R}^n$, and measurable functions $\tilde{\beta} : [0, T] \to \mathbb{R}^m$, $\zeta : [0, T] \to \mathbb{R}$ such that, for an arbitrary scalar $r > 0$:*

$$u^*(t) = U^{-1}\left(F^\intercal p(t) + E^\intercal \tilde{\beta}(t) - (\zeta(t) + r)E^\intercal v^*(t)\right)$$

and the following conditions hold:

1. *The transversality condition:* $\begin{pmatrix} p(0) \\ -p(T) \end{pmatrix} \in \operatorname{im} \begin{pmatrix} M^\intercal \\ N^\intercal \end{pmatrix}$.

2. *The Euler adjoint equation: for almost every $t \in [0, T]$,*

$$\begin{pmatrix} \dot{x}^* \\ \dot{p} \end{pmatrix} = \begin{pmatrix} A & FU^{-1}F^\intercal \\ Q & -A^\intercal \end{pmatrix} \begin{pmatrix} x^* \\ p \end{pmatrix} + \begin{pmatrix} FU^{-1}E^\intercal & B - (\zeta + r)FU^{-1}E^\intercal \\ -C^\intercal & (\zeta + r)C^\intercal \end{pmatrix} \begin{pmatrix} \tilde{\beta} \\ v^* \end{pmatrix} \tag{6.12}$$

$$\begin{cases} 0 \leq \begin{pmatrix} \tilde{\beta} \\ v^* \end{pmatrix} \perp \begin{pmatrix} EU^{-1}E^\intercal & D - (\zeta + r)EU^{-1}E^\intercal \\ \zeta EU^{-1}E^\intercal - D^\intercal & \zeta(D+V) + (\zeta + r)\left(D^\intercal - \zeta EU^{-1}E^\intercal\right) \end{pmatrix} \begin{pmatrix} \tilde{\beta} \\ v^* \end{pmatrix} \\ \qquad\qquad + \begin{pmatrix} C & EU^{-1}F^\intercal \\ \zeta C & \zeta EU^{-1}F^\intercal - B^\intercal \end{pmatrix} \begin{pmatrix} x^* \\ p \end{pmatrix} \geq 0 \\ \tilde{\beta} \geq rv^*. \end{cases}$$

3. *The Weierstrass condition for radius R: for almost every $t \in [t_0, t_1]$,*

$$(x^*(t), u, v) \in S, \ \left\| \begin{pmatrix} u \\ v \end{pmatrix} - \begin{pmatrix} u^*(t) \\ v^*(t) \end{pmatrix} \right\| < R$$

$$\implies \langle p(t), Ax^*(t) + Bv^*(t) + Fu^*(t)) \rangle - \frac{1}{2}\left(x^*(t)^\intercal Q x^*(t) + u^*(t)^\intercal U u^*(t)\right)$$
$$\geq \langle p(t), Ax^*(t) + Bv + Fu) \rangle - \frac{1}{2}\left(x^*(t)^\intercal Q x^*(t) + u^\intercal U u\right).$$

The importance of this result is twofold. First, it gives a way to analyze the optimal trajectory using these necessary conditions. All results concerning the analysis of LCS can be used to prove some properties of possible trajectories of (6.12) and to derive results on continuity, jumps or sensitivity on parameters, and therefore to prove some properties of the optimal trajectory. The analysis of LCS relies heavily on the matrix appearing in front of $\binom{\tilde{\beta}}{v}$ in the complementarity conditions of (6.12). However, with no more hypothesis on matrices appearing in (6.12), we were not able to derive sharper results.

6.2 Sufficiency of the W-stationarity

Surprisingly, the weakest form of stationarity for the problem (6.1)(6.2) turns to be also sufficient, in some sense. For this, we need to define trajectories with the same *history*. The development shown here is directly inspired by [46, Proposition 3.1] and by [101].

Definition 6.2.1. *Let (x,u,v) and (x^*,u^*,v^*) be two admissible trajectories for (6.2) (associated with $w = Cx + Dv + Eu$ and w^*, defined the same way). We say that they have the same history on $[0,T]$ if the following condition holds for almost every $t \in [0,T]$:*

$$[v_i(t) = 0 \iff v_i^*(t) = 0] \text{ and } [w_i(t) = 0 \iff w_i^*(t) = 0]$$

From the point of view of the switching systems, two trajectories have the same history on $[0,T]$ if they visit the same modes at the same time along $[0,T]$. In the following sufficient condition for optimality, the comparison of the different trajectories is done with respect to this history condition.

Theorem 6.2.1. *Suppose that (x^*,u^*,v^*) is an admissible W-stationary trajectory (with $\lambda_0 = -\frac{1}{2}$). Then, (x^*,u^*,v^*) minimizes (6.1)(6.2) among all admissible trajectories for (6.2) having the same history.*

Proof. Since (x^*,u^*,v^*) is a W-stationary trajectory, there exist an arc p and measurable functions λ^G and λ^H satisfying (6.4). Notice that (6.4) implies, for almost all $t \in [0,T]$ and all $i \in \overline{m}$,

$$\lambda_i^G(t)v_i^*(t) = 0 \text{ and } \lambda_i^H(t)w_i^*(t) = 0. \tag{6.13}$$

Let (x,u,v) be a second admissible trajectory for (6.2) with the same history as (x^*,u^*,v^*). Since they both have the same history, it also satisfies, for almost all $t \in [0,T]$ and all $i \in \overline{m}$:

$$\lambda_i^G(t)v_i(t) = 0 \text{ and } \lambda_i^H(t)w_i(t) = 0. \tag{6.14}$$

Denote $L(x,u,v) = \frac{1}{2}\left(x^\intercal Q x + u^\intercal U u + v^\intercal V v\right)$. The goal is to prove:

$$\Delta = \int_0^T \left(L(x(t),u(t),v(t)) - L(x^*(t),u^*(t),v^*(t))\right) dt \geq 0$$

For this, let us first transform the expression of Δ.

$$\begin{aligned}
\Delta &= \int_0^T \left(L(x(t),u(t),v(t)) - L(x^*(t),u^*(t),v(t))\right) dt \\
&\qquad + \int_0^T p(t)^\intercal(\dot{x}(t) - Ax(t) - Bv(t) - Fu(t) - (\dot{x}^*(t) - Ax^*(t) - Bv^*(t) - Fu^*(t)))dt \\
&= \int_0^T \left(L(x(t),u(t),v(t)) - L(x^*(t),u^*(t),v^*(t))\right) dt + \int_0^T \frac{d}{dt}\left[p(t)^\intercal(x(t) - x^*(t))\right] dt \\
&\qquad\qquad - \int_0^T \dot{p}(t)^\intercal(x(t) - x^*(t))dt \\
&\qquad\qquad - \int_0^T p(t)^\intercal \left(A(x(t) - x^*(t)) + B(v(t) - v^*(t)) + F(u(t) - u^*(t))\right) dt
\end{aligned}$$

The last equality is obtained by integration by parts of $\int_0^T p^\intercal(\dot{x}-\dot{x}^*)$.

$$\begin{aligned}
\Delta &= \int_0^T (L(x(t),u(t),v(t)) - L(x^*(t),u^*(t),v^*(t)))\,dt + \int_0^T \frac{d}{dt}\left[p(t)^\intercal(x(t)-x^*(t))\right]dt \\
&\quad - \int_0^T (\dot{p}(t) + A^\intercal p(t) + C^\intercal\lambda^H(t))^\intercal(x(t)-x^*(t))dt + \int_0^T (C^\intercal\lambda^H(t))^\intercal(x(t)-x^*(t))dt \\
&\quad - \int_0^T (F^\intercal p(t) + E^\intercal\lambda^H(t))^\intercal(u(t)-u^*(t))dt + \int_0^T (E^\intercal\lambda^H(t))^\intercal(u(t)-u^*(t))dt \\
&\quad - \int_0^T (B^\intercal p(t) + \lambda^G(t) + D^\intercal\lambda^H(t))^\intercal(v(t)-v^*(t))dt \\
&\quad + \int_0^T (\lambda^G(t) + D^\intercal\lambda^H(t))^\intercal(v(t)-v^*(t))dt \\
&= \int_0^T \Bigg(L(x(t),u(t),v(t)) - L(x^*(t),u^*(t),v^*(t)) \\
&\qquad\qquad - \begin{pmatrix} \dot{p}(t) + A^\intercal p(t) + C^\intercal\lambda^H(t) \\ F^\intercal p(t) + E^\intercal\lambda^H(t) \\ B^\intercal p(t) + \lambda^G(t) + D^\intercal\lambda^H(t) \end{pmatrix}^\intercal \begin{pmatrix} x(t)-x^*(t) \\ u(t)-u^*(t) \\ v(t)-v^*(t) \end{pmatrix} \Bigg) dt \\
&\quad + \int_0^T \lambda^H(t)^\intercal(C(x(t)-x^*(t)) + D(v(t)-v^*(t)) + E(u(t)-u^*(t)))dt \\
&\quad + \int_0^T \lambda^G(t)^\intercal(v(t)-v^*(t))dt + \int_0^T \frac{d}{dt}\left[p(t)^\intercal(x(t)-x^*(t))\right]dt
\end{aligned}$$

As it is proved in Proposition 6.1.2, $\begin{pmatrix} \dot{p}(t) + A^\intercal p(t) + C^\intercal\lambda^H(t) \\ F^\intercal p(t) + E^\intercal\lambda^H(t) \\ B^\intercal p(t) + \lambda^G(t) + D^\intercal\lambda^H(t) \end{pmatrix} = \nabla L(x^*(t),u^*(t),v^*(t))$. Since L is a convex function, it yields for almost all t in $[0,T]$:

$$L(x(t),u(t),v(t)) - L(x^*(t),u^*(t),v^*(t)) - (\nabla L(x^*(t),u^*(t),v^*(t)))^\intercal \begin{pmatrix} x(t)-x^*(t) \\ u(t)-u^*(t) \\ v(t)-v^*(t) \end{pmatrix} \geq 0 \quad (6.15)$$

Therefore, this proves:

$$\begin{aligned}
\Delta &\geq \int_0^T \lambda^H(t)^\intercal(w(t)-w^*(t))dt + \int_0^T \lambda^G(t)^\intercal(v(t)-v^*(t))dt + \int_0^T \frac{d}{dt}\left[p(t)^\intercal(x(t)-x^*(t))\right]dt \\
&\geq \int_0^T \frac{d}{dt}\left[p(t)^\intercal(x(t)-x^*(t))\right]dt,
\end{aligned}$$

the last inequality being obtained with (6.13) and (6.14). Furthermore:

$$\int_0^T \frac{d}{dt}\left[p(t)^\intercal(x(t)-x^*(t))\right]dt = -\begin{pmatrix} p(0) \\ -p(T) \end{pmatrix}^\intercal \begin{pmatrix} x(0)-x^*(0) \\ x(T)-x^*(T) \end{pmatrix}$$

But the boundary conditions in (6.2) yield $(M\ N)\begin{pmatrix} x(0)-x^*(0) \\ x(T)-x^*(T) \end{pmatrix} = 0$, such that

$$\begin{pmatrix} x(0)-x^*(0) \\ x(T)-x^*(T) \end{pmatrix} \in \ker(M\ N) = \left(\operatorname{im}\begin{pmatrix} M^\intercal \\ N^\intercal \end{pmatrix}\right)^\perp.$$

And since $\begin{pmatrix} p(0) \\ -p(T) \end{pmatrix} \in \text{im} \begin{pmatrix} M^\intercal \\ N^\intercal \end{pmatrix}$, it proves $\int_0^T \frac{d}{dt}\left[p(t)^\intercal(x(t)-x^*(t))\right]dt = 0$. Finally, we conclude that $\Delta \geq 0$. □

Remark 6.2.1. *One could want to get rid of the history hypothesis, since it "fixes" the switching times and does not render optimality according to these times. Very formally, it is easy to see where the problem has some leeway. Without the history hypothesis, one still can prove:*

$$\begin{aligned}\Delta \;=\; \int_0^T \Bigg(& L(x(t),u(t),v(t)) - L(x^*(t),u^*(t),v^*(t)) \\ & - \begin{pmatrix} \dot{p}(t) + A^\intercal p(t) + C^\intercal \lambda^H(t) \\ F^\intercal p(t) + E^\intercal \lambda^H(t) \\ B^\intercal p(t) + \lambda^G(t) + D^\intercal \lambda^H(t) \end{pmatrix}^\intercal \begin{pmatrix} x(t)-x^*(t) \\ u(t)-u^*(t) \\ v(t)-v^*(t) \end{pmatrix} \Bigg) dt \\ + \int_0^T & (\lambda^H(t)^\intercal w(t) + \lambda^G(t)^\intercal v(t))dt\end{aligned}$$

Suppose that $u(t) \neq u^*(t)$ on a measurable subset J of $[0,T]$. Then, by strict convexity of L in variable u, for almost all t in J:

$$L(x(t),u(t),v(t)) - L(x^*(t),u^*(t),v^*(t)) - (\nabla L(x^*(t),u^*(t),v^*(t)))^\intercal \begin{pmatrix} x(t)-x^*(t) \\ u(t)-u^*(t) \\ v(t)-v^*(t) \end{pmatrix} > 0$$

Therefore, using again (6.15), one proves:

$$\Delta > \int_0^T \lambda^H(t)^\intercal w(t)dt + \int_0^T \lambda^G(t)^\intercal v(t)dt$$

In order to simplify the problem, suppose that v and w have a different history than v^* and w^* in the neighbourhood of a single switching point t^*. Then, the inequality becomes, for some $\varepsilon > 0$:

$$\Delta > \int_{t^*-\varepsilon}^{t^*+\varepsilon} \left(\lambda^H(t)^\intercal w(t) + \lambda^G(t)^\intercal v(t)\right) dt$$

If for some $\varepsilon > 0$ small enough, $\left|\int_{t^*-\varepsilon}^{t^*+\varepsilon} \left(\lambda^H(t)^\intercal w(t) + \lambda^G(t)^\intercal v(t)\right) dt\right|$ is small enough, then $\Delta \geq 0$. Therefore, the first order conditions also render optimality according to small variations of the switching times.

Remark 6.2.2. *All these considerations about sufficiency of the W-stationarity still hold true if L is replaced by any other convex function, possibly non differentiable. Also, Remark 6.2.1 also holds the same way as long as L is strictly convex in one of its variable.*

Conclusion

Using the results exposed in Section 3.2, we were able to derive necessary conditions for optimality; these conditions were then re-expressed in the more suitable form of a linear system defined via a complementarity problem. As it has been proved, these stationarity conditions turns to be also sufficient. These results can therefore be used to develop an efficient numerical method: this is the topic of the next chapter.

Chapter 7

Numerical implementations

Abstract. Using the results of Chapter 6, we will describe two numerical schemes used for approximating the optimal trajectory: the direct method and the indirect method. The code developed for these methods is presented, along with several numerical results. The performances of both methods are tested, and eventually we show the advantages of both methods. This chapter has been submitted to IEEE Transactions on Automatic Control [108].

This chapter aims at providing numerical schemes that compute an approximation of a solution of (6.1)(6.2). As it has been exposed in Section 2.2, two approaches are followed in this manuscript:

- the direct approach, designed in Section 7.1;
- the hybrid direct/indirect approach, designed in Section 7.2.

7.1 Direct method

7.1.1 Description and properties

The direct method consists in discretizing directly the problem (6.1)(6.2) in order to solve a finite-dimensional optimization problem. To this aim let us propose the following explicit Euler discretization:

$$\begin{aligned} \min \ & \sum_{k=0}^{L} x_k^\intercal Q x_k + u_k^\intercal U u_k + v_k^\intercal V v_k \\ \text{s.t.} \ & \begin{cases} \dfrac{x_{k+1}-x_k}{h} = Ax_k + Bv_k + Fu_k, \ k = 0, ..., L-1 \\ 0 \leq v_k \perp Cx_k + Dv_k + Eu_k \geq 0, \ k = 0, ..., L-1 \\ Mx_0 + Nx_L = x_b, \end{cases} \end{aligned} \tag{7.1}$$

where $h = \frac{T}{L}$ is the time-step, considered constant. By simple application of [75, Theorem 1.4.3], one easily prove the following Proposition:

Proposition 7.1.1. *For all fixed positive scalar h, (7.1) admits a global minimum.*

The discretization of the complementarity conditions, appearing in (7.1), differs from the implicit Euler methods found in [1, 2, 27, 61, 71]. For this optimal control problem, the complementarity should not be seen as a way to express the variable v_k, but as a mixed constraint. Therefore, its

discretization must hold at all discrete times t_k, and the trajectory, solution of this discretized LCS, will be computed not step by step but for all k in one shot.
The problem is then to solve the program (7.1) numerically. To this end, we use one of the two Algorithms 1 or 2 described in Section 3.1.2. The reason to use these algorithms is that, under some hypothesis, they converge to M- or S-stationary points, and the analysis of (6.1)(6.2) made in Chapter 6 shows that the optimal trajectories are M- or S-stationary.

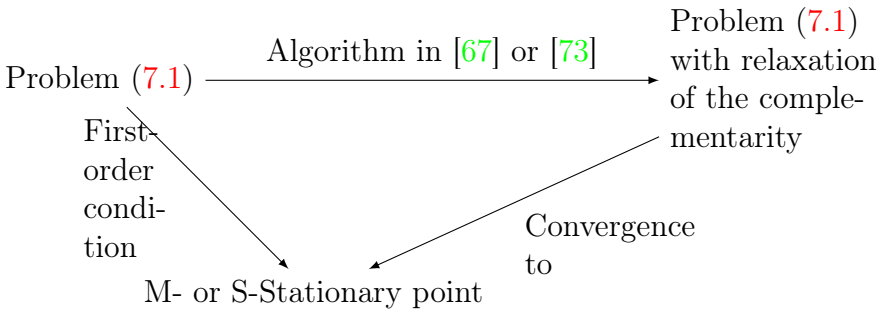


Figure 7.1: Sketch of the direct method for problem (6.1)(6.2).

Consistency of the scheme

Let us first compute the stationarity conditions for problem (7.1). Since the MPEC Linear Condition holds and if we suppose E invertible, according to Theorem 3.1.1, a local minimizer must be S-stationary. We denote $\{p_i\}_{i=0}^{L-1}$ the multipliers for the discretized differential equations, $\{\theta_i\}_{i=1}^{L}$ and $\{\nu_i\}_{i=1}^{L}$ the multipliers for each side of the complementarity constraints. The stationarity conditions for the MPEC (7.1) read as:

$$\begin{aligned}
\frac{x_{i+1}-x_i}{h} - Ax_i - Bv_i - Fu_i &= 0 \\
Qx_i - \left(A + \frac{1}{h}I\right)^{\intercal} p_i + \frac{1}{h}p_{i-1} - C^{\intercal}\nu_i &= 0 \\
Uu_i - F^{\intercal}p_i - E^{\intercal}\nu_i &= 0 \\
-B^{\intercal}p_i - \theta_i - D^{\intercal}\nu_i &= 0 \\
\theta_i = 0 \qquad & \forall i \in I^{+0}(x,u,v) \\
\nu_i = 0 \qquad & \forall i \in I^{0+}(x,u,v) \\
\nu_i \geq 0,\ \lambda_i \geq 0, \qquad & \forall i \in I^{00}(x,u,v),
\end{aligned} \tag{7.2}$$

for all $i \in \{1, ..., L-1\}$, $h = \frac{T}{L}$, L being a fixed positive integer.

Proposition 7.1.2. *The stationarity conditions (7.2) define a scheme consistent with the Euler adjoint equation of an S-stationary trajectory.*

Proof. Let us check that the consistency error goes to 0 when h goes to 0. For this, we take the solutions $(x, u, v, p, \lambda^H, \lambda^G)$ at discretisation times t_i. For $k = 1, .., L-1$, let us denote ε_k^h the

consistency error at time t_k :

$$\varepsilon_k^h = \begin{pmatrix} \dfrac{-x(t_{k+1})}{h} + \left(A + \dfrac{1}{h}I\right) x(t_k) + Bv(t_k) + Fu(t_k) \\ Qx(t_k) - \left(A + \dfrac{1}{h}I\right)^\intercal p(t_k) + \dfrac{1}{h}p(t_{k-1}) - C^\intercal \lambda^H(t_k) \\ Uu(t_k) - F^\intercal p(t_k) - E^\intercal \lambda^H(t_k) \\ - B^\intercal p(t_k) - \lambda^G(t_k) - D^\intercal \lambda^H(t_k) \end{pmatrix} \tag{7.3}$$

$$= \begin{pmatrix} Ax(t_k) + \dot{x}(t_k) + Bv(t_{k-1}) + Fu(t_{k-1}) + O(h) \\ Qx(t_k) - A^\intercal p(t_k) - \dot{p}(t_k) - C^\intercal \lambda^H(t_k) + O(h) \\ Uu(t_k) - F^\intercal p(t_k) - E^\intercal \lambda^H(t_k) \\ - B^\intercal p(t_k) - \lambda^G(t_k) - D^\intercal \lambda^H(t_k) \end{pmatrix} = \begin{pmatrix} O(h) \\ O(h) \\ 0 \\ 0 \end{pmatrix}.$$

It follows that $\lim_{h\to 0} \left(\max_{k=1,\dots,L} \|\varepsilon_k^h\|_\infty\right) = 0$. In addition, discrete multipliers ν and θ respect the same equality and inequality conditions as the multipliers λ^H and λ^G of a S-stationary trajectory at discrete times t_k. □

Remark 7.1.1. *It should be noted that this result holds thanks to the discretization made in* (7.1) *that admits some* asymmetry. *Indeed, the discretization made for x is not the same for discretizing the cost (where x is discretized as a piecewise constant function), and the dynamics (where x is discretized as a piecewise affine function). If one chooses to discretize x in the running cost also as a piecewise affine function, then the proof of Proposition 7.1.2 is more involved.*

7.1.2 Description of the code

A code in Python has been written in order to implement this method. The codes were designed using the library CasADi [7], which offers a framework for symbolic computation and most importantly a convenient way for interfacing with optimization solvers. The optimization solver used is IPOPT [111]. All the codes produced in this thesis are available for test at `https://gitlab.inria.fr/avieira/optLCS`. A class diagram showing the architecture of the code is presented Figure 7.2. The direct method was implemented in the class `OptLCSDirect`. Concerning the direct method, the main focus was on two features:

1. The code has to be easily launched, needing only the constants of the model. The constructor of the class only needs the matrices A, B, C, D, E, F, Q and U, and initial and final times t_0 and t_f. The code also automatically identifies the dimensions n and m, and treats initial, final and/or mixed boundary constraints. The resolution is typified by the method `compute_optimal`, which already has default values and can be launched directly.

2. The Algorithms 1 and 2 need an NLP solver, and the optimization problem has to be accordingly defined. This has been done in `optimLoop_augment` and `optimLoop_relax`.

7.1.3 Numerical examples

Order for 1D examples

Let us apply the direct method on the following example:

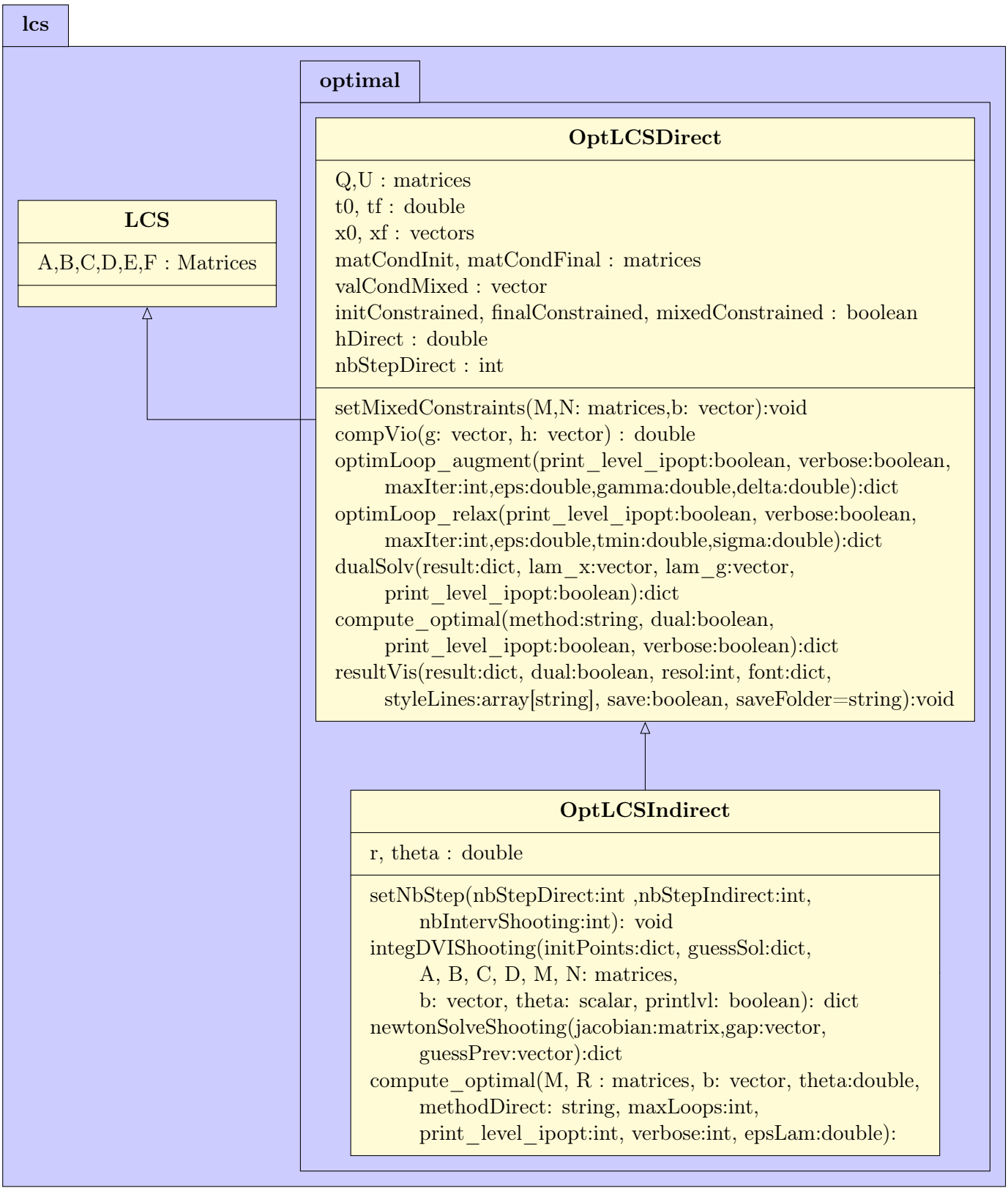


Figure 7.2: Class diagram for `optLCS`.

Example 7.1.1.

$$\begin{aligned} &\textit{minimize} \int_0^T \left(x(t)^2 + u(t)^2\right) dt, \\ &\textit{such that:} \begin{cases} \dot{x}(t) = ax(t) + bv(t) + fu(t), \\ 0 \leq v(t) \perp dv(t) + eu(t) \geq 0, \ \textit{a.e. on } [0,T] \\ x(0) = x_0, \ x(T) \textit{ free}, \end{cases} \end{aligned} \tag{7.4}$$

where all variables are scalars, $d > 0$, $b, e \neq 0$. *Using the results in [31], the constants in the LCS are chosen such that the system is completely controllable. As proved in Section 4.2, the only stationary trajectory is given by:*

$$\begin{cases} x^*(t) = \dot{p}(t) + ap(t) \\ u^*(t) = \begin{cases} fp(t) & \textit{if } efx(0) \leq 0, \\ \left(f - \frac{eb}{d}\right) p(t) & \textit{if } efx(0) \geq 0. \end{cases} \\ v^*(t) = \dfrac{1}{d} \max(0, -eu^*(t)) \end{cases} \tag{7.5}$$

where:

$$p(t) = \frac{1}{2\sqrt{\gamma}} \left[\left((\sqrt{\gamma} - a)e^{\sqrt{\gamma}t} + (\sqrt{\gamma} + a)e^{-\sqrt{\gamma}t} \right) p(0) + \left(e^{\sqrt{\gamma}t} - e^{-\sqrt{\gamma}t} \right) x(0) \right]$$

$p(0) = -\frac{x(0)\left(e^{2\sqrt{\gamma}T} - 1\right)}{(\sqrt{\gamma} - a)e^{2\sqrt{\gamma}T} + \sqrt{\gamma} + a}$ *and* $\gamma = \begin{cases} (a^2 + f^2) & \textit{if } efx(0) \leq 0, \\ \left(a^2 + \left(f - \frac{be}{d}\right)^2\right) & \textit{if } efx(0) \geq 0. \end{cases}$. *Figures 7.3-7.5 show the evolution of error with time-step in log-log scales, using the two different algorithms presented in Section 3.1.2, with the parameters* $a = 3$, $b = -0.5$, $d = 1$, $e = -2$, $f = 3$, $T = 1$, *and either* $x_0 = 1$ *or* $x_0 = -1$.

In these examples, we clearly see convergence of both algorithms, with an order close to 1. However, Figures 7.3b and 7.5c suggest that in some cases, the algorithms face difficulties when the time-step is decreasing. This is actually something known with direct methods: often they fail to be precise. We can simply understand it, since decreasing the time-step increases the dimension of the optimization problem to solve. In order to tackle such problems, one has to choose a different method presented in Section 7.2.

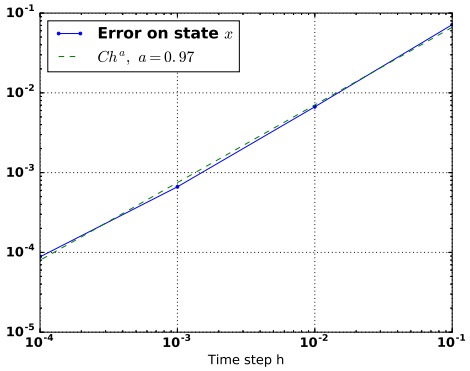
Example with $D = 0$

As alluded to in Section 1.2, a crucial parameter in LCS is the relative degree between w and v, which determines the solution set as a subclass of Schwarz' distributions [3]. Let us consider now a case with $D = 0$ and relative degree one.

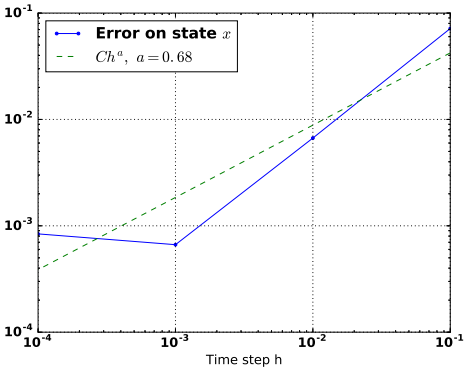
Example 7.1.2.

$$\begin{aligned} &\textit{minimize} \int_0^T \left(\|x(t)\|_2^2 + u(t)^2\right) dt \\ &\textit{such that:} \begin{cases} \dot{x}(t) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} x(t) + \begin{pmatrix} -1 \\ 1 \end{pmatrix} v(t) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u(t), \\ 0 \leq v(t) \perp \begin{pmatrix} -1 & 1 \end{pmatrix} x(t) + u(t) \geq 0, \\ x(0) = (-0.5, 1), \ x(T) \textit{ free}. \end{cases} \end{aligned} \tag{7.6}$$

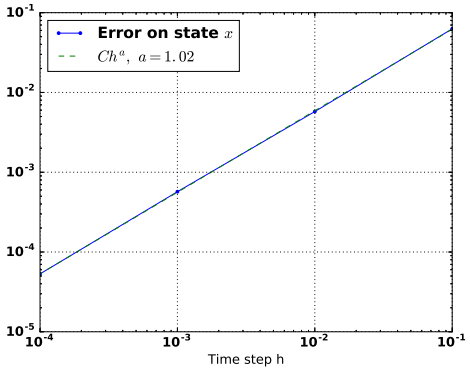
The numerical results for Example 7.1.2 are shown in Figure 7.6. They demonstrate that the direct method can also succeed when D *is not a P-matrix.*



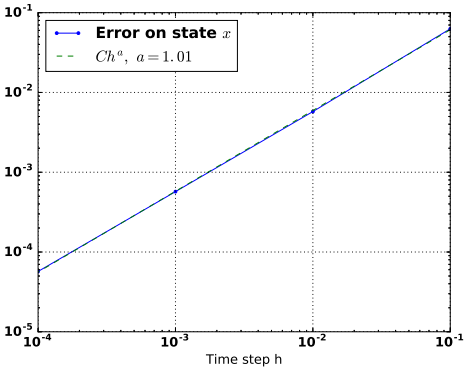
(a) Algorithm in [67], $x_0 = -1$



(b) Algorithm in [73], $x_0 = -1$



(c) Algorithm in [67], $x_0 = 1$



(d) Algorithm in [73], $x_0 = 1$

Figure 7.3: Error on state x when using Algorithms in [67] or in [73] with $a = 3$, $b = -0.5$, $d = 1$, $e = -2$, $f = 3$, $T = 1$ and $x_0 = -1$ or $x_0 = 1$ in Example 7.1.1.

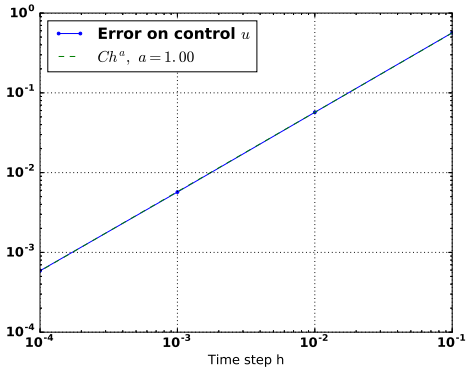
Higher dimensional examples

For higher dimension or when $C \neq 0$, we do not have an analytical solution to compare with the numerical one, but still we can check if the multipliers comply with an S-stationary trajectory. For this purpose, let us test them on a third example:

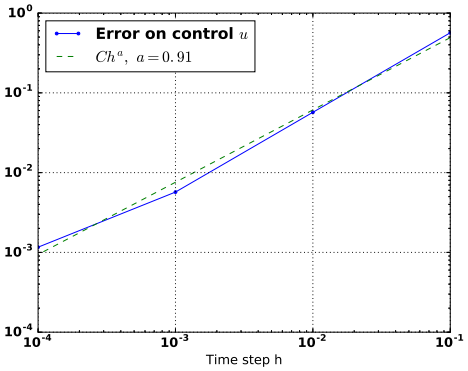
Example 7.1.3.

$$\begin{aligned} &\textit{minimize} \int_0^1 \left(\|x(t)\|_2^2 + 25\|u(t)\|_2^2\right) dt, \\ &\textit{such that:} \begin{cases} \dot{x}(t) = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} x(t) + \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix} v(t) + \begin{pmatrix} 1 & 3 \\ 2 & 1 \end{pmatrix} u(t), \\ 0 \leq v(t) \perp \begin{pmatrix} 3 & -1 \\ -2 & 0 \end{pmatrix} x(t) + v(t) + \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} u(t) \geq 0, \quad \textit{a.e. on } [0,1] \\ x(0) = \begin{pmatrix} -\frac{1}{2} \\ 1 \end{pmatrix}, \ x(T) \textit{ free}, \end{cases} \end{aligned} \tag{7.7}$$

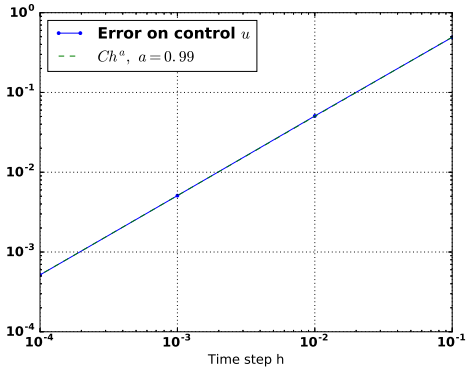
where x, u and v are functions $[0,1] \to \mathbb{R}^2$. As shown in Figure 7.7, the Algorithm in [73] seems to fail to respect the complementarity condition between v_2 and w_2 at the beginning. The Algorithm in [67] seems to behave better. Comparing first Figure 7.8b and Figure 7.7a, then Figure 7.8c and



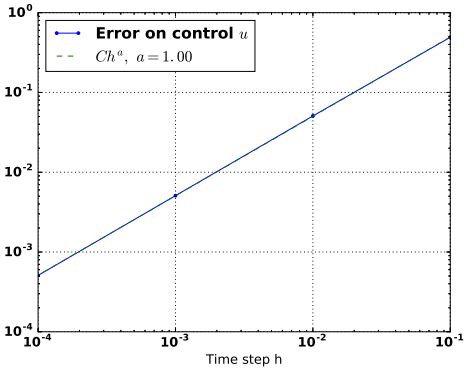
(a) Algorithm in [67], $x_0 = -1$



(b) Algorithm in [73], $x_0 = -1$



(c) Algorithm in [67], $x_0 = 1$



(d) Algorithm in [73], $x_0 = 1$

Figure 7.4: Error on control u when using Algorithms in [67] or in [73] with $a = 3$, $b = -0.5$, $d = 1$, $e = -2$, $f = 3$, $T = 1$ and $x_0 = -1$ or $x_0 = 1$ in Example 7.1.1.

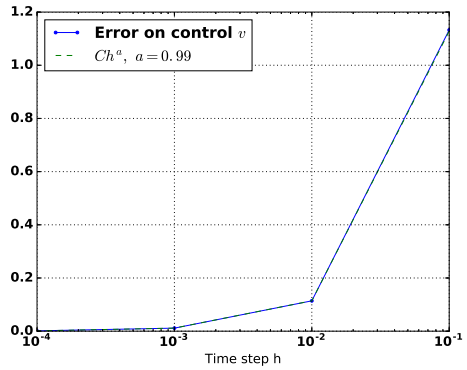
Figure 7.7c, results suggest that we retrieve an S-stationary trajectory (according to the sign of the multipliers, and their complementarity with v and w), as expected.

Since v is not upper-bounded nor present in the running cost in previous examples, the optimal trajectory may present big variations due to v. It is the case for the following Example 7.1.4, where x takes values in $\mathbb{R}^2$, u and v take values in $\mathbb{R}$.

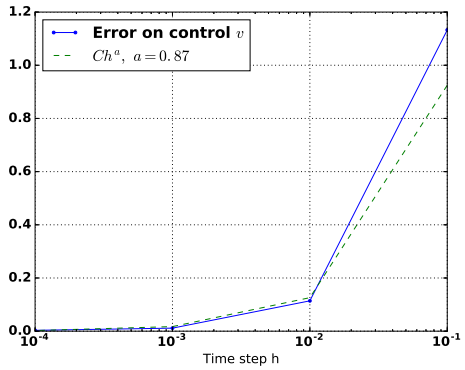
Example 7.1.4.

$$\begin{aligned} &\textit{minimize} \int_0^1 \left(\|x(t)\|_2^2 + u(t)^2\right) dt, \\ &\textit{such that:} \begin{cases} \dot{x}(t) = \begin{pmatrix} 5 & -6 \\ 3 & 9 \end{pmatrix} x(t) + \begin{pmatrix} 4 \\ 5 \end{pmatrix} v(t) + \begin{pmatrix} 0 \\ -4 \end{pmatrix} u(t), \\ 0 \le v(t) \perp \begin{pmatrix} -1 & 5 \end{pmatrix} x(t) + v(t) + 6u(t) \ge 0, \qquad \textit{a.e. on } [0,1] \\ x(0) = \begin{pmatrix} -\frac{1}{2} \\ -1 \end{pmatrix}, \ x(T) \textit{ free}, \end{cases} \end{aligned} \tag{7.8}$$

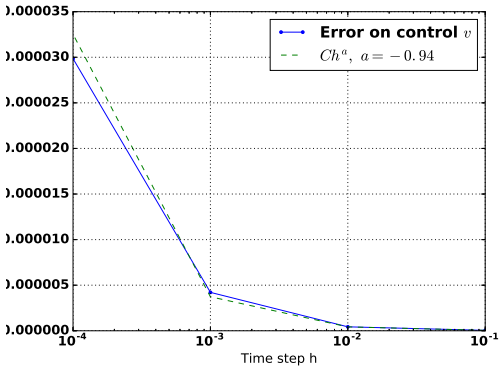
As shown in Figure 7.9, the optimal solution admits a peak on v at the very beginning of the interval. One could think that the state x admits a jump, which could mean that the solution of the LCS is distributional (in which case the dynamics in (6.2) has to be recast into measure differential



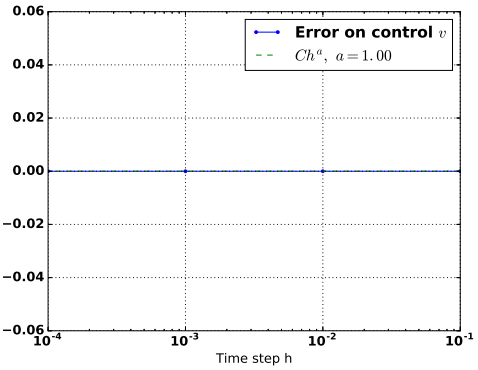
(a) Algorithm in [67], $x_0 = -1$



(b) Algorithm in [73], $x_0 = -1$



(c) Algorithm in [67], $x_0 = 1$



(d) Algorithm in [73], $x_0 = 1$

Figure 7.5: Error on v when using Algorithms in [67] or in [73] with $a = 3$, $b = -0.5$, $d = 1$, $e = -2$, $f = 3$, $T = 1$ and $x_0 = -1$ or $x_0 = 1$ in Example 7.1.1.

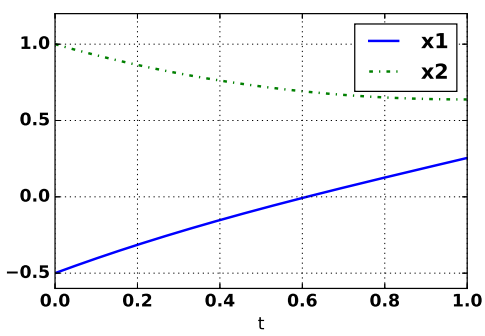
inclusions), but shrinking the time-step does not change this peak on v, which is always positive on a non-shrinking interval whatever the time-step h.

One could wonder what happens in Example 7.1.4 if a quadratic cost $v^\intercal V v$ (with V symmetric positive definite) is added in the running cost. This could prevent the initial huge peak on v. This is the investigation of Example 7.1.5. The code has been slightly changed in order to add a quadratic cost in v.

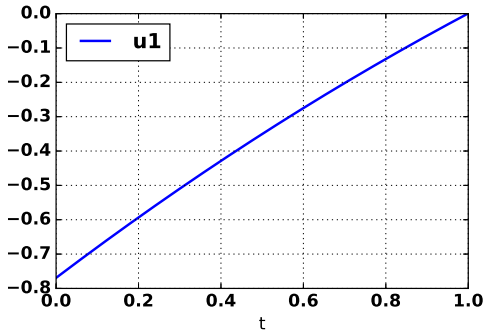
Example 7.1.5.

$$\begin{aligned} &\textit{minimize} \int_0^1 \left(\|x(t)\|_2^2 + u(t)^2 + \alpha v(t)^2\right) dt, \\ &\textit{such that:} \begin{cases} \dot{x}(t) = \begin{pmatrix} 5 & -6 \\ 3 & 9 \end{pmatrix} x(t) + \begin{pmatrix} 4 \\ 5 \end{pmatrix} v(t) + \begin{pmatrix} 0 \\ -4 \end{pmatrix} u(t), \\ 0 \le v(t) \perp \begin{pmatrix} -1 & 5 \end{pmatrix} x(t) + v(t) + 6u(t) \ge 0, \qquad \textit{a.e. on } [0,1] \\ x(0) = \begin{pmatrix} -\frac{1}{2} \\ -1 \end{pmatrix}, \ x(T) \textit{ free}, \end{cases} \end{aligned} \tag{7.9}$$

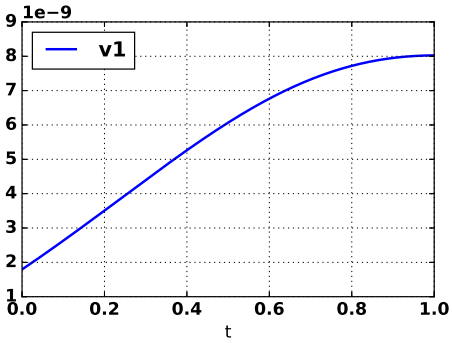
where $\alpha > 0$. The numerical results are shown Figure 7.10, and a special focus on v for $\alpha \in \{10, 5, 10^{-1}, 10^{-3}, 0\}$ is shown Figure 7.11 for $t \in [0, 0.1]$. We clearly see a continuity property of



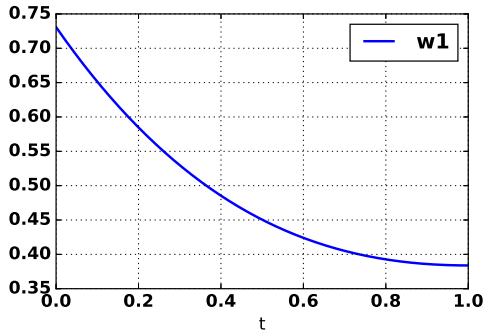
(a) State x



(b) Control u



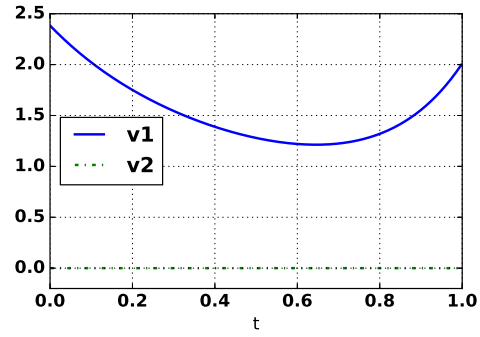
(c) Control v



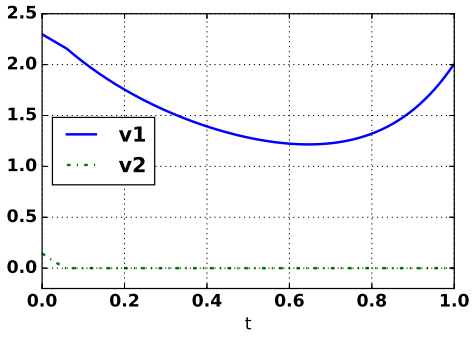
(d) Variable w

Figure 7.6: Numerical results for Example 7.1.2 using Algorithm in [67], $h = 10^{-3}$.

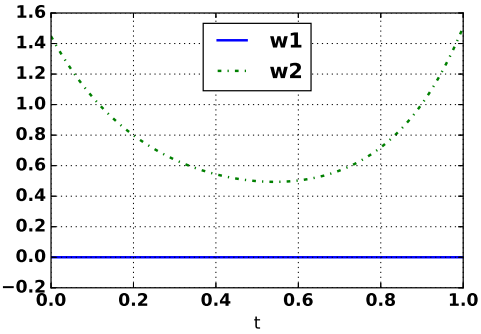
the solution with respect to α when it shrinks to 0. Adding this quadratic cost on v may then be a way to smoothen the solution, getting rid of the initial huge peak.



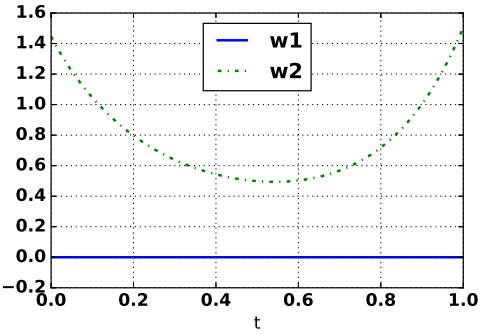
(a) v, Algorithm in [67]



(b) v, Algorithm in [73]

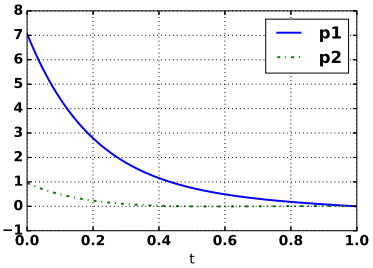


(c) w, Algorithm in [67]

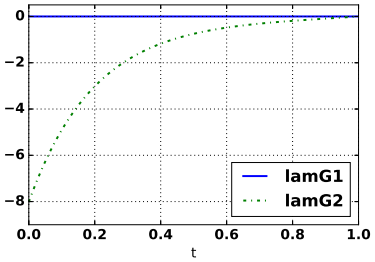


(d) w, Algorithm in [73]

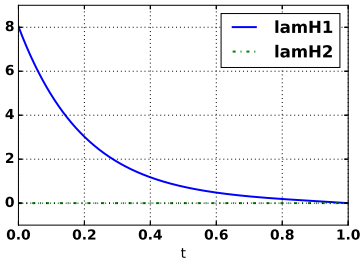
Figure 7.7: Numerical results for Example 7.1.3 using Algorithms in [67] and in [73], for comparison concerning complementarity. $h = 10^{-3}$.



(a) Adjoint state p

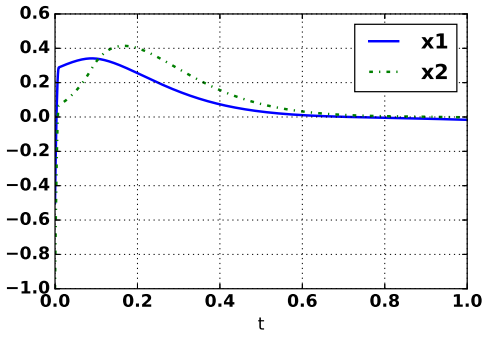


(b) Multiplier λ^G

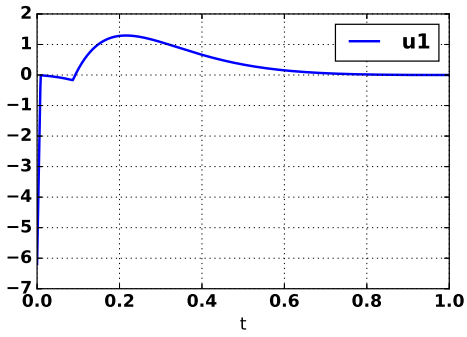


(c) Multiplier λ^H

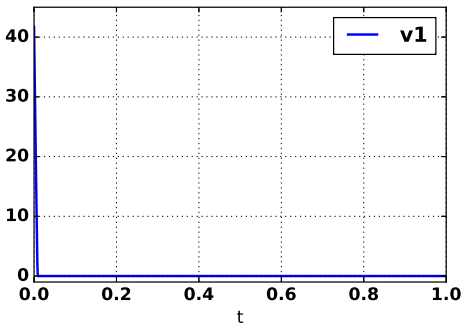
Figure 7.8: Computed multipliers for Example 7.1.3 using Algorithm in [67]. $h = 10^{-3}$.



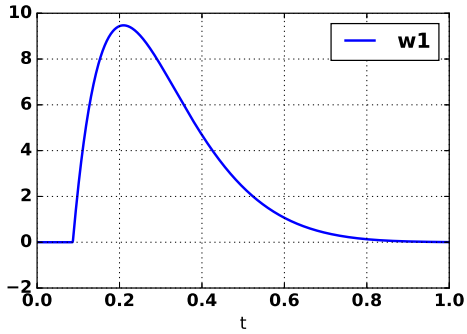
(a) State x



(b) Control u

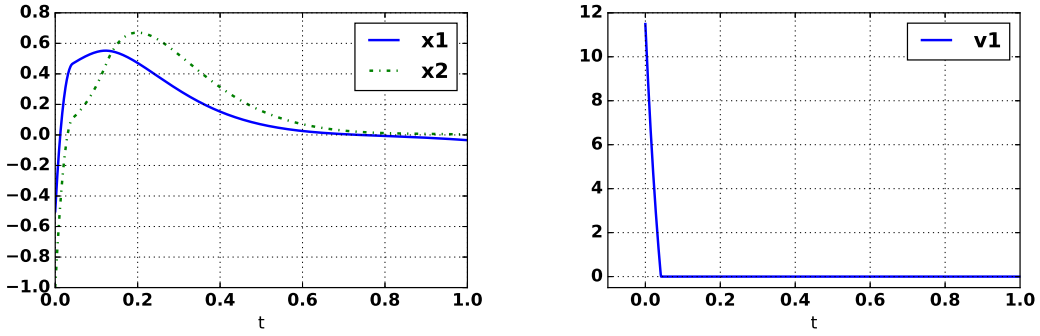


(c) Control v

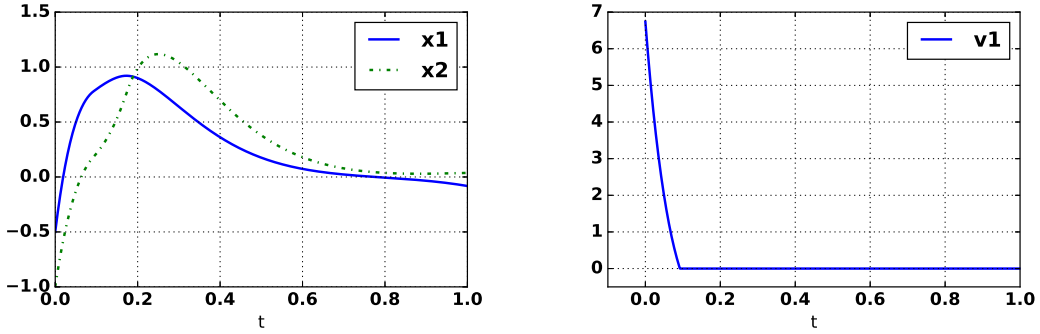


(d) Variable w

Figure 7.9: Numerical results for Example 7.1.4 using Algorithm in [67], $h = 10^{-3}$.



(a) $\alpha = 1$



(b) $\alpha = 10$

Figure 7.10: Numerical x and v found for Example 7.1.5 using Algorithm in [67], $h = 10^{-3}$, and different values of α.

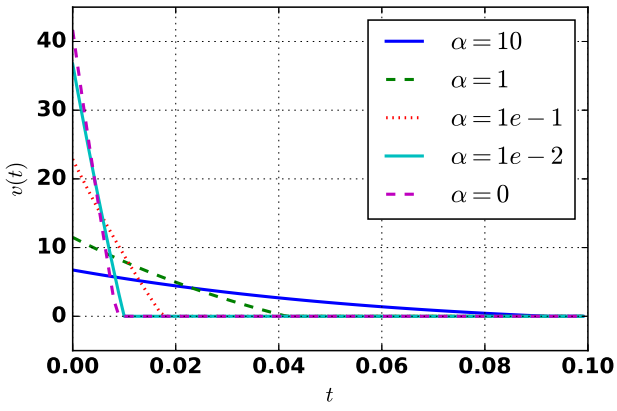


Figure 7.11: Numerical v on $[0, 0.1]$ found for Example 7.1.5 using Algorithm in [67], $h = 10^{-3}$, and different values of α.

In the previous numerical simulations, the optimal control seems always continuous. The next example suggests that the optimal control may jump. Let us consider the following problem, where for all t in $[0,1]$, $x(t) \in \mathbb{R}^2$ and $u(t) \in \mathbb{R}$.

Example 7.1.6. *[Discontinuous optimal controller]*

$$\begin{array}{l} \textit{minimize} \displaystyle\int_0^1 \left(\|x(t)\|_2^2 + u(t)^2\right) dt, \\ \textit{such that:} \begin{cases} \dot{x}(t) = \begin{pmatrix} 1 & -3 \\ -8 & 10 \end{pmatrix} x(t) + \begin{pmatrix} -3 \\ -1 \end{pmatrix} v(t) + \begin{pmatrix} 4 \\ 8 \end{pmatrix} u(t), \\ 0 \le v(t) \perp \begin{pmatrix} 1 & -3 \end{pmatrix} x(t) + 5v(t) + 3u(t) \ge 0, \qquad \textit{a.e. on } [0,1] \\ x(0) = \begin{pmatrix} -\frac{1}{2} \\ -1 \end{pmatrix}, \ x(T) \textit{ free}, \end{cases} \end{array} \tag{7.10}$$

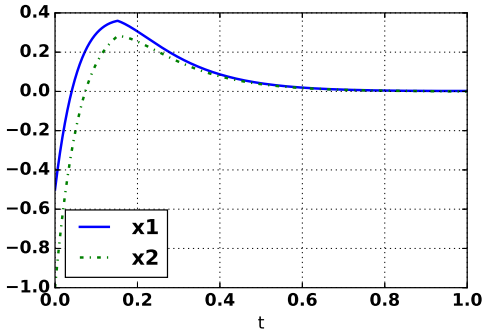
The numerical results are shown in Figure 7.12. The associated multipliers and adjoint state, retrieved from these calculations, are shown in Figure 7.13. The complementarity constraint is satisfied, and the associated multipliers suggest that the trajectory indeed is an S-stationary trajectory. It is clear that u admits a switch around $t_1 = 0.112$ and is not continuous (see Figure 7.12b). It is noteworthy to take a look at the different modes activated along the solution. In this case where the complementarity constraint is of dimension 1, we have three possible cases : $v = 0 < w$ (happening on $[0, t_1]$), $v > 0 = w$ (happening on $[t_1, 0.87]$ approximately), $v = 0 = w$ (happening on $[0.87, 1]$). It shows that, compared with some other methods for optimal control of switching systems (see for instance [86, 96]), this method does not require to guess a priori the number of switches nor the times of commutation in order to approximate the solution. The tracking of the switches is taken care of by the MPEC solver. This is a major advantage of the complementarity approach over event-driven, hybrid-like approaches.

Eventually, the class of solution considered may actually be too small, and the direct method may converge to a solution with the state admitting jumps. This is the main focus of the Example 7.1.7.

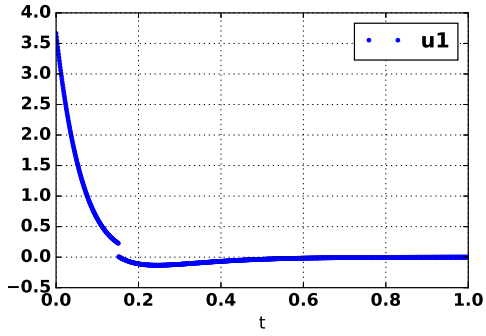
Example 7.1.7.

$$\begin{array}{l} \textit{minimize} \displaystyle\int_0^{10} \left(\|x(t)\|_2^2 + u(t)^2 + \alpha v(t)^2\right) dt, \\ \textit{such that:} \begin{cases} \dot{x}(t) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} x(t) + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} v(t) + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} u(t), \\ 0 \le v(t) \perp \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} x(t) + u(t) \ge 0, \qquad \textit{a.e. on } [0,10] \\ x(0) = \begin{pmatrix} -2 \\ 1 \\ -1 \end{pmatrix}, \ x(T) \textit{ free}, \end{cases} \end{array} \tag{7.11}$$

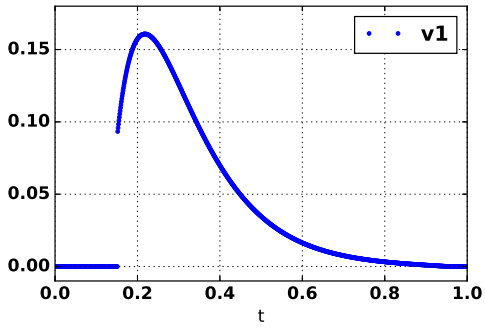
with $\alpha \in \{0, 1, 10\}$. As shown Figure 7.14, the solution with $\alpha = 0$ admits a huge peak around $t = 4.85$, that yields a jump on x_3. When $\alpha > 0$, this peak disappears, but a smaller at $t = 0$ is recovered. Even though adding v in the running cost smoothen the solution, it shows that the optimal solution still admits huge variation.



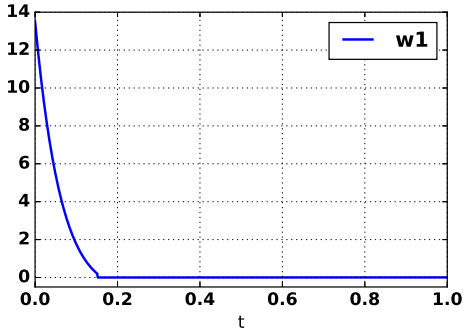
(a) State x



(b) Control u

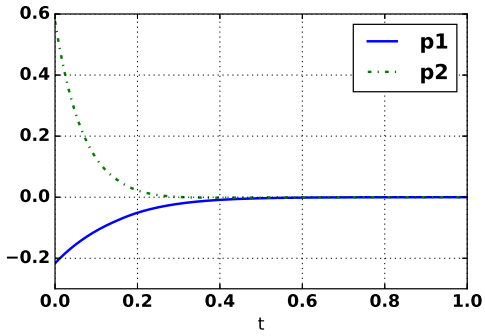


(c) Control v

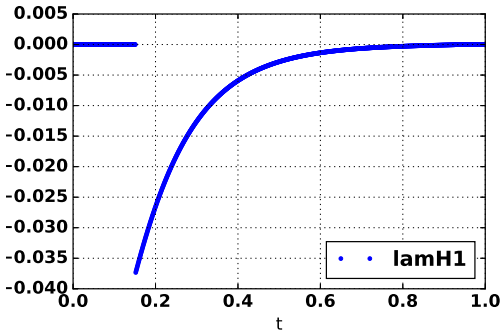


(d) Variable w

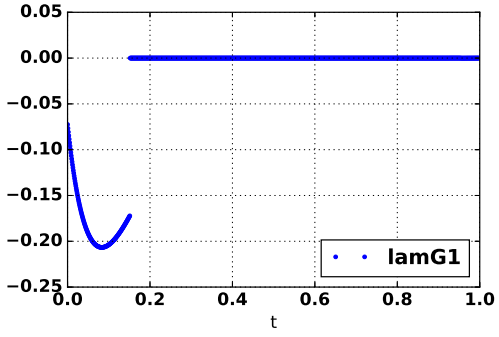
Figure 7.12: Numerical results for Example 7.1.6 using Algorithm in [67], $h = 10^{-3}$.



(a) Adjoint state p

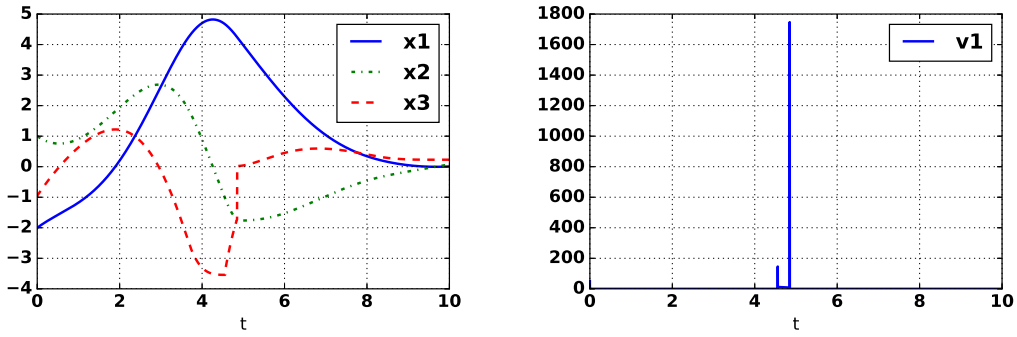


(b) Multiplier λ^H

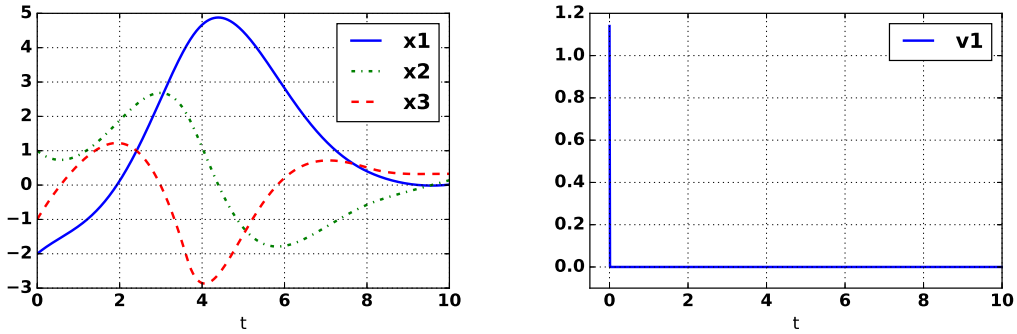


(c) Multiplier λ^G

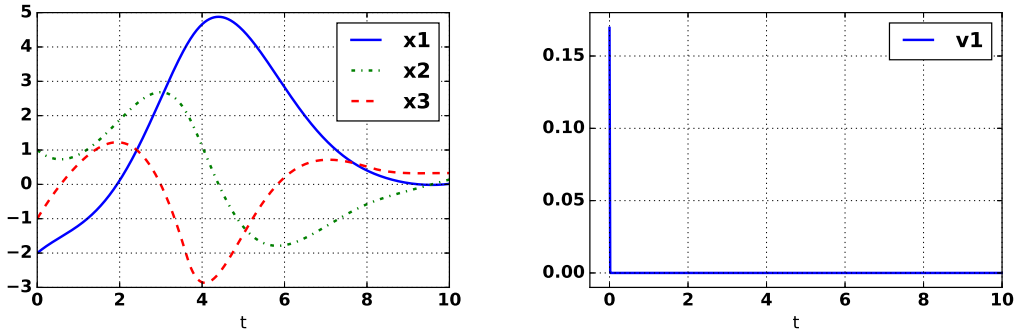
Figure 7.13: Numerical results for Example 7.1.6 using Algorithm in [67], $h = 10^{-3}$.



(a) $\alpha = 0$



(b) $\alpha = 1$



(c) $\alpha = 10$

Figure 7.14: Numerical results for x and v for Example 7.1.7 using Algorithm in [67], $h = 10^{-3}$, using different values for α.

7.1.4 A physical example

Let us now focus on a physical example, directly taken in [21].

Example 7.1.8. *Consider the electrical circuit of Figure 7.15, where D is an ideal diode, L the inductance, R the resistance, C the capacitor.*

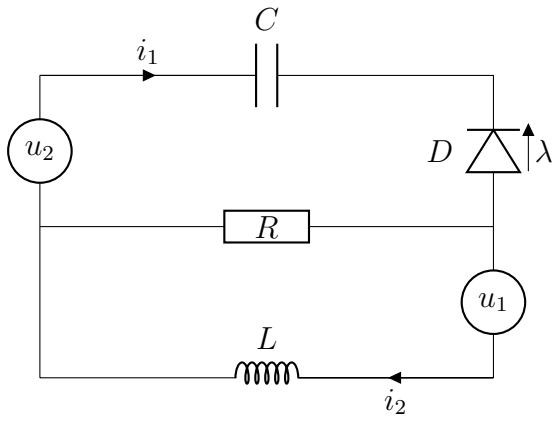


Figure 7.15: Circuit with an ideal diode and two voltage sources

Denote $x_1(t) = \int_0^t i_1(s)ds + x_1(0)$ (the charge of the capacitor, in coulomb) and $x_2(t) = i_2(t)$ (the electric current, in ampere). Then the evolution of this system is described as:

$$\begin{cases} \dot{x}_1(t) = \dfrac{-1}{RC}x_1(t) + x_2(t) - \dfrac{1}{R}\lambda(t) + \dfrac{1}{R}u_2(t), \\ \dot{x}_2(t) = \dfrac{-1}{LC}x_1(t) - \dfrac{1}{L}\lambda(t) + \dfrac{1}{L}(u_2(t) - u_1(t)), \\ 0 \leq \lambda(t) \perp w(t) = \dfrac{1}{RC}x_1(t) - x_2(t) + \dfrac{1}{R}\lambda(t) - \dfrac{1}{R}u_2(t) \geq 0, \end{cases}$$

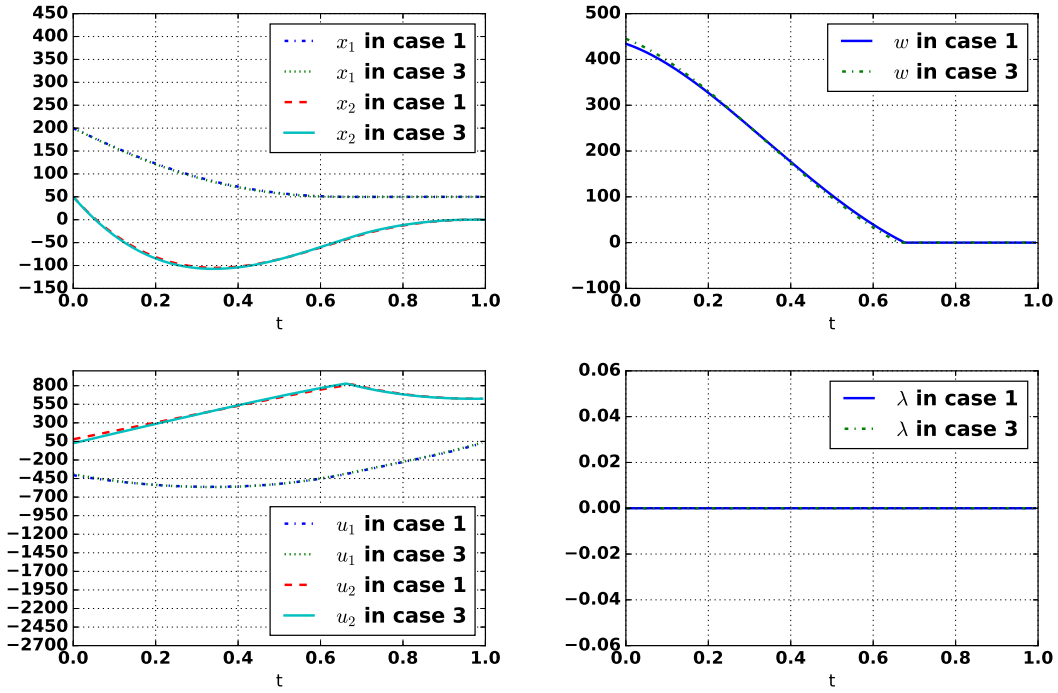
The constants are chosen as $R = 10\Omega$, $C = 80\ 000\mu F$, and $L = 2H$. Let us first remark that the complementarity condition impose in particular that $\dot{x}_1(t) \leq 0$ (the capacitor can not be charged). Suppose the initial state is $x(0) = (200C, 50A)$. Three different problems are solved:

1. *Minimizing the cost:*
$$\int_0^1 [Rx_2(t)^2 + u_1(t)^2 + u_2(t)^2]dt \tag{7.12}$$
with final state $x(1) = (50C, 0A)$.

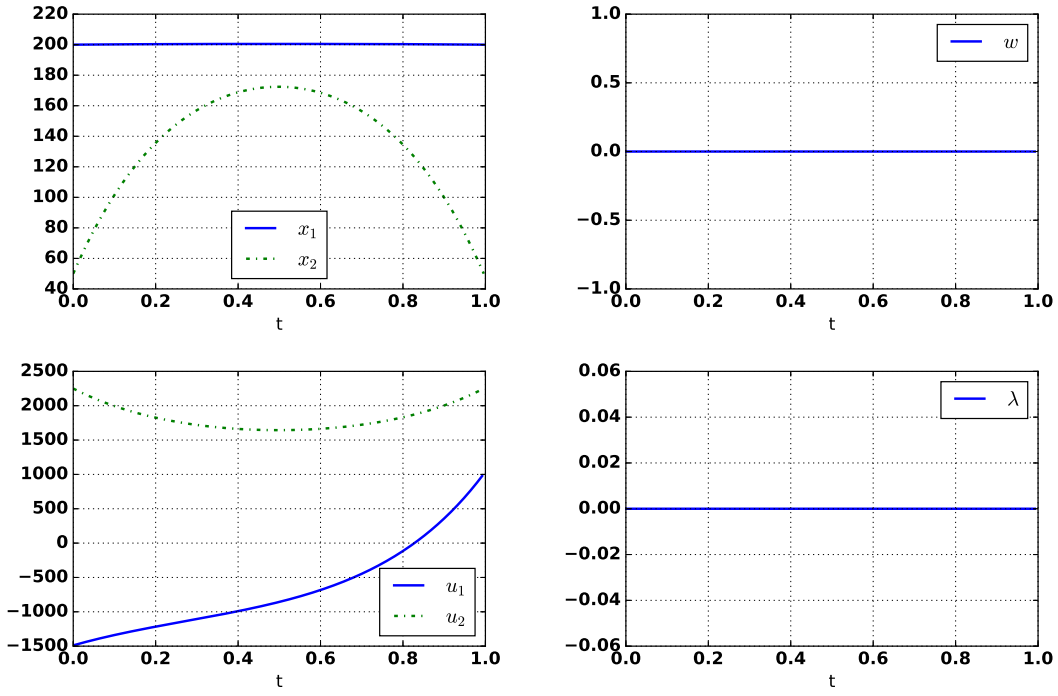
2. *Minimizing the cost (7.12) with final state $x(1) = x(0)$.*

3. *Minimizing the cost:*
$$\int_0^1 \left[\frac{1}{2C}x_1(t)^2 + Rx_2(t)^2 + u_1(t)^2 + u_2(t)^2\right] dt \tag{7.13}$$
with final state $x(1) = (50C, 0A)$.

The numerical results for this problem are plotted in Figure 7.16. One remarks that cases 1 and 3 gives similar trajectories. In case 2, it is interesting to see that since the control u_2 maintaining $\dot{x}_1 = 0$ is admissible, this path is followed. Actually, this is the only path available (since $\dot{x}_1$ must be nonnegative).



(a) Comparison between cases 1 and 3



(b) Case 2

Figure 7.16: Numerical results for Example 7.1.8 using Algorithm in [67], $h = 10^{-3}$.

7.2 Combining direct and indirect methods: the hybrid approach

The indirect method consists in solving the first-order necessary conditions derived in Chapter 6 in order to solve the optimal control problem. Since these conditions are necessary and sufficient, solving these equations is equivalent to solving the optimal control problem (6.1)(6.2). As pointed out in [104], it has the advantage that the numerical solutions are usually very precise, but the method suffers from a huge sensitivity on the initial guess. Indeed, if the initial guess is not close enough to a solution, then the method may fail to converge.
A natural approach is then to use both the direct and the indirect methods in order to obtain a very precise solution, taking advantage of both methods: this is called the hybrid approach. In our framework, we have to face two problems. First, the active index sets appearing in the Euler equations (6.4), used to impose conditions on the multipliers, are not useful as they are. This problem has been tackled by re-expressing these equations in Theorem 6.1.2.

7.2.1 Properties of the indirect approach

BVP solver for the indirect method

Solving the first order conditions boils down to solving a Boundary Value Problem (BVP). This is the case for instance when $x(T)$ is free; the transversality conditions impose in that case $p(T) = 0$. The problem is then to find a solution (x, p) of (6.12) such that $x(0) = x_0$ and $p(T) = 0$. Finding such a solution is not trivial, especially in this case since the dynamical system is an LCS.

Denote by $z = \binom{x}{p}$ and suppose we can rewrite boundary values on z as a linear equation $\tilde{M}z(0)+\tilde{N}z(T) = \tilde{x}_b$. Then Theorem 6.1.2 implies that the extremal is a solution of the Boundary Value Problem (BVP):

$$\begin{cases} (a)\ \dot{z} = \mathcal{A}z + \mathcal{B}\lambda \\ (b)\ 0 \leq \lambda \perp \mathcal{D}\lambda + \mathcal{C}z \geq 0 \\ (c)\ \mathcal{E}^\intercal\lambda \geq 0 \\ (d)\ \tilde{M}z(0) + \tilde{N}z(T) = \tilde{x}_b, \end{cases} \tag{7.14}$$

where the matrices $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$ and $\mathcal{E}$ are easily identifiable from (6.12), and $\lambda = \binom{\tilde{\beta}}{v}$. This is a Boundary Value Problem (BVP) formulated for an LCS with constraint (7.14)(c). The shooting method is usually employed to solve such a problem: roughly speaking, given $z_0 \in \mathbb{R}^{2n}$, we compute the solution $z(\cdot; z_0)$ of (7.14)(a)(b)(c) with initial data $z(0) = z_0$. Letting $F(z_0) = \tilde{M}z_0 + \tilde{N}z(T; z_0) - \tilde{x}_b$, the BVP becomes a root-search of F. In practice, we employ multiple shooting: we also take into account in F shooting nodes inside the interval $[0, T]$, where we make sure that $z(\cdot, z_0)$ is continuous. In the smooth case, we would use a Newton method, which needs the Jacobian $F'(z_0)$ to compute each iteration leading to the root of F. In our case, the dependence on z_0 of $z(T; z_0)$ is not smooth. Some properties concerning such dependence for LCS have been derived in [85], as recalled in Section 1.2.5. Broadly speaking, the authors built a linear Newton Approximation, which allow them to design a non-smooth Newton method for solving BVP for LCS. However, their result can not be directly applied here for two reasons. First, aside the complementarity conditions, we also have to take into account the inequality condition (7.14)(c). Secondly, their result relies on the fact that $\mathcal{B}$ $\mathrm{SOL}(\mathcal{C}z(t), \mathcal{D})$ is a singleton for all $t \in [0, T]$. However, this method still could work for (7.14), since the research will only be local. Section 7.2.1

shows numerical results where this non-smooth Newton method has been used successfully.

An efficient method to solve the LCS

In the first simulations we ran, we noticed that the integration step by step of the LCS IVP (7.14)(a)(b)(c), $z(0) = z_0$, admitted some numerical instability that multiple shooting could not solve. This problem was solved using the following proposition:

Proposition 7.2.1. *Let t_i, $t_{i+1} \in [0, T]$, $t_i < t_{i+1}$. (z, λ) is a solution of*

$$\begin{cases} \dot{z} = \mathcal{A}z + \mathcal{B}\lambda \\ 0 \leq \lambda \perp \mathcal{D}\lambda + \mathcal{C}z \geq 0 \\ \mathcal{E}^\intercal \lambda \geq 0 \\ z(t_i) = z_i, \end{cases} \tag{7.15}$$

on $[t_i, t_{i+1}]$, if and only if it is a global minimum of the optimal control problem:

$$\begin{aligned} \min \quad & \int_{t_i}^{t_{i+1}} \lambda(t)^\intercal (\mathcal{D}\lambda(t) + \mathcal{C}z(t))dt \\ s.t. \quad & \left.\begin{array}{l} \dot{z}(t) = \mathcal{A}z(t) + \mathcal{B}\lambda(t) \\ \lambda(t) \geq 0 \\ \mathcal{D}\lambda(t) + \mathcal{C}z(t) \geq 0 \\ \mathcal{E}\lambda(t) \geq 0 \end{array}\right\} \text{ a.e. on } [t_i, t_{i+1}] \\ & z(t_i) = z_i \end{aligned} \tag{7.16}$$

with minimum equal to 0.

Proof. ($\Longrightarrow$) Suppose (z, λ) is a solution of (7.15), then (z, λ) is obviously admissible for (7.16), and $\int_{t_i}^{t_{i+1}} \lambda(t)^\intercal (\mathcal{D}\lambda(t) + \mathcal{C}z(t))dt = 0$. Suppose there exists an admissible solution $(\tilde{z}, \tilde{\lambda})$ of (7.16) such that $\int_{t_i}^{t_{i+1}} \tilde{\lambda}(t)^\intercal (\mathcal{D}\tilde{\lambda}(t) + \mathcal{C}\tilde{z}(t))dt < 0$. It then means that there exists $\tau_1, \tau_2 \in [t_i, t_{i+1}]$ such that $[\tau_1, \tau_2]$ is of positive measure and $\tilde{\lambda}(t)^\intercal (\mathcal{D}\tilde{\lambda}(t) + \mathcal{C}\tilde{z}(t)) < 0$ a.e. on $[\tau_1, \tau_2]$. This contradicts the fact that $\tilde{\lambda} \geq 0$ and $\mathcal{D}\tilde{\lambda} + \mathcal{C}\tilde{z} \geq 0$ a.e. on $[t_i, t_{i+1}]$. Then the minimum is non-negative, and (λ, z) is a global minimum.
($\Longleftarrow$) Suppose (z, λ) is a solution of (7.16). Notice that $\lambda(t)^\intercal (\mathcal{D}\lambda(t) + \mathcal{C}z(t)) \geq 0$ a.e. on $[t_i, t_{i+1}]$, so $\int_{t_i}^{t_{i+1}} \lambda(t)^\intercal (\mathcal{D}\lambda(t) + \mathcal{C}z(t))dt = 0$ implies that $\lambda(t)^\intercal (\mathcal{D}\lambda(t) + \mathcal{C}z(t)) = 0$ a.e. on $[t_i, t_{i+1}]$. So (z, λ) is a solution of (7.15). □

Numerically, problem (7.16) will be solved for each interval $[t_i, t_{i+1}]$ using a classical direct method, where t_i is a node for the Multiple Shooting method. One could ask why this formulation is more stable than just discretizing directly equation (7.15). Intuitively, we can explain it as follows:

- Using for instance an implicit Euler a discretization of (7.15), one solves at each step the problem:

$$\begin{aligned} & z_{k+1} - z_k = h\left(\mathcal{A}z_{k+1} + \mathcal{B}\lambda_{k+1}\right) \\ & 0 \leq \lambda_{k+1} \perp \mathcal{D}\lambda_{k+1} + \mathcal{C}z_{k+1} \geq 0 \\ & \mathcal{E}^\intercal \lambda_{k+1} \geq 0, \end{aligned}$$

which takes the form of an LCP with unknown λ_{k+1} and an inequality constraint. But the exact solution $(z^*_{k+1}, \lambda^*_{k+1})$ will not be found. Instead, an approximated solution $(z_{k+1}, \lambda_{k+1}) = (z^*_{k+1} + \varepsilon_k, \lambda^*_{k+1} + \varepsilon_\lambda)$ will be sought. Then the error will propagate along the solution on $[0, T]$, causing instabilities.

- However, if one solves (7.16), all errors will appear under the integral sign. Since this integral is minimized (and we expect the result to be 0), the errors on the whole interval can also be expected to be minimized.

7.2.2 Description of the code

As for the direct method, a code in Python has been written in order to implement this method. The codes were once again designed using the library CasADi and IPOPT. A class diagram showing the architecture of the code is presented Figure 7.2. The indirect method was implemented in the class `OptLCSIndirect`. Concerning the indirect method, the main focus was on two features:

1. The code has to be as easily launched as the direct method. It only needs, aside what is needed for the direct method, the matrices M and N appearing in (7.14) and a number of steps and shooting nodes for the integration of this equation (via `setNbSetp`).

2. The multiple shooting has been implemented. There are two aspects to this feature:

   - the integration of the dynamics between each shooting node using (7.16). This has been solved using kind of a direct method with an Euler discretization (since (7.16) is seen as an optimal control problem). This is coded in the method `integDVIShooting`;
   - the non-smooth Newton method, presented in Section 7.2.1. The main focus was on solving the DI (1.15), and then solving a linear system in order to obtain a descent direction. It was implemented in `newtonSolveShooting`.

   The shooting algorithm was stopped as soon as the maximum gap at shooting nodes was of the order of the time-step (in order to assure continuity of the state and the adjoint state).

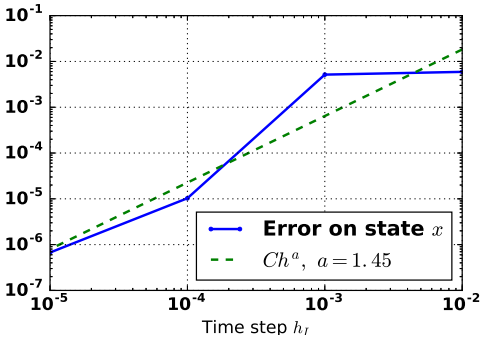
7.2.3 Numerical results

Analytical 1D example revisited

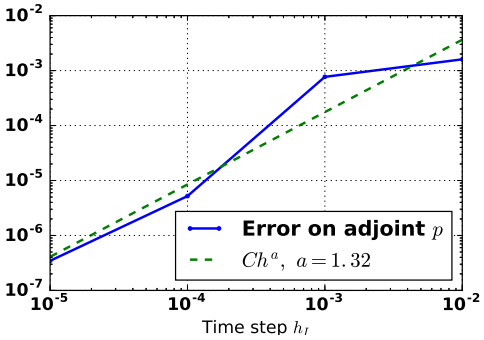
First, let us check the convergence of the method of Section 7.2.1 on the 1D Example 7.1.1. Since the Direct Method achieved to reach a satisfactory precision, one can expect also the Indirect method to converge. The results are presented in Figures 7.17 and 7.18. Overall, the method reaches the precision of the time step, even for very small precision. Concerning the state x and the adjoint state p, the convergence is even faster. Concerning λ^H, λ^G and v, it seems however harder to converge. But still, the desired precision is met, and it is often more precise than the Direct Method.

Example 7.1.3 revisited

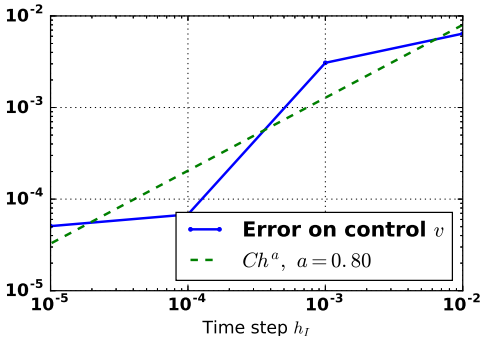
In order to compare the Hybrid Approach with the raw direct method, we ran simulations on Example 7.1.3 using different time-steps, and comparing the time spent for solving it at the desired precision. The results are shown in Tables 7.1 and 7.2. It appears clearly that, even though the



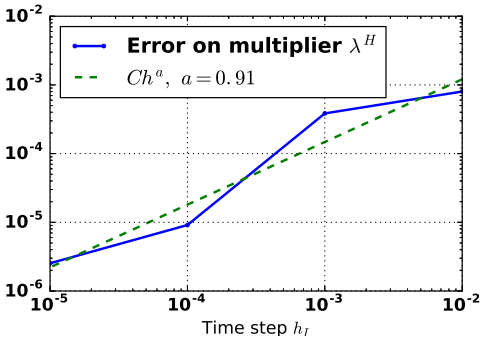
(a) Error on x



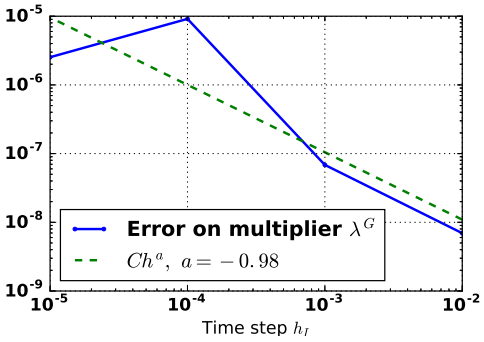
(b) Error on p



(c) Error on v



(d) Error on λ^H



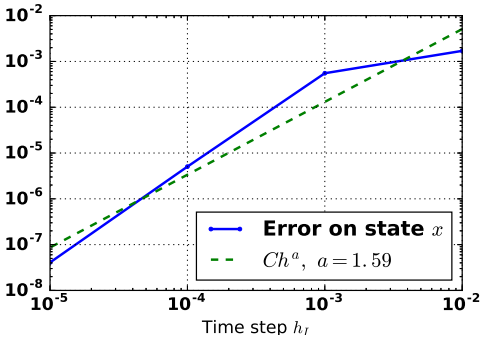
(e) Error on λ^G

Figure 7.17: Errors with the Hybrid Approach for Example 7.1.1 with $a = 3$, $b = -0.5$, $d = 1$, $e = -2$, $f = 3$, $T = 1$ and $x_0 = -1$.

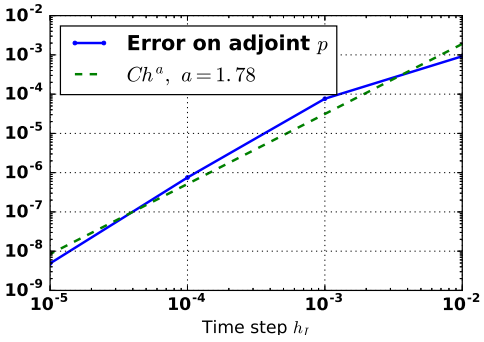
Indirect method is not that interesting for rough precisions, it becomes necessary for really high precisions. The Newton Method developed in this context is also satisfying, as shown in Figure 7.19, which shows the maximum gap left on x and p at shooting nodes. The program assumes to reach convergence as soon as the continuity on $\binom{x}{p}$ is met with precision h. As shown in this example, convergence is achieved in two iterations.

Conclusion

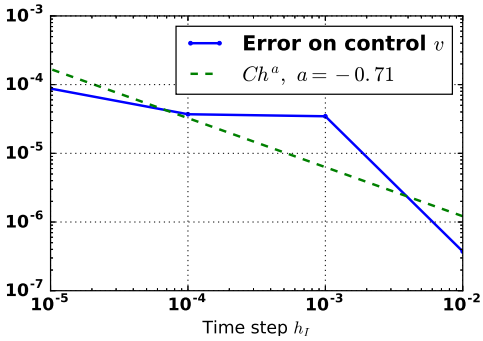
In this chapter, two codes for solving numerically the quadratic optimal control of LCS have been designed: a direct method, and a hybrid direct/indirect method. These results highlight some properties of the optimal solution and offer some new perspectives. The methods developed here



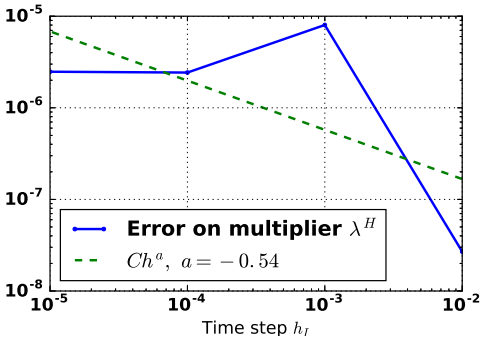
(a) Error on x



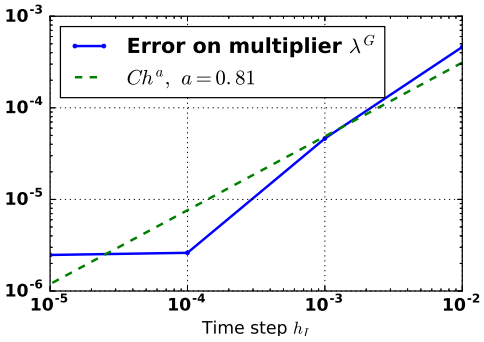
(b) Error on p



(c) Error on v



(d) Error on λ^H



(e) Error on λ^G

Figure 7.18: Errors with the Hybrid Approach for Example 7.1.1 with $a = 3$, $b = -0.5$, $d = 1$, $e = -2$, $f = 3$, $T = 1$ and $x_0 = 1$.

work well for the quadratic case. One could wonder what kind of solution one gets for a different cost. The next chapter focuses on a different problem: the minimal time optimal control problem.

h_D	Time spent (s)
10^{-2}	1.31
10^{-3}	37.50
10^{-4}	400.65
10^{-5}	∞
10^{-6}	∞

Table 7.1: Time spent for computing an approximate solution of Example 7.1.3 using the direct method, with different time steps h_D. ∞ means that the calculations did not end (segmentation fault).

Parameters	Time spent (s)
$h_D = 10^{-1}$, $h_I = 10^{-2}$, $n_S = 5$	1.39
$h_D = 10^{-1}$, $h_I = 10^{-3}$, $n_S = 10$	11.26
$h_D = 10^{-2}$, $h_I = 10^{-4}$, $n_S = 20$	97.56
$h_D = 10^{-3}$, $h_I = 10^{-5}$, $n_S = 50$	1 298.62
$h_D = 10^{-4}$, $h_I = 10^{-6}$, $n_S = 100$	32 163.36

Table 7.2: Time spent for computing an approximate solution of Example 7.1.3 using the Hybrid approach, with different time steps: h_D for the first guess, and final solution with h_I, using n_S intervals of shooting.

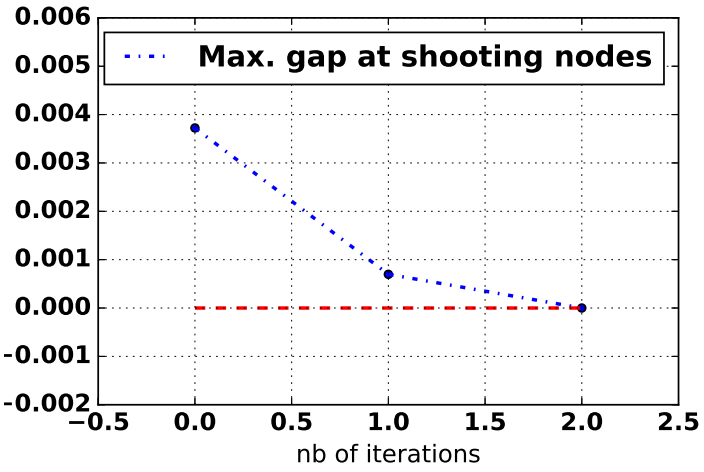


Figure 7.19: Maximum gaps on (x, p) at each iteration of the Newton Method used for the Indirect Method for computation of Example 7.1.3. $h_I = 10^{-5}$.

Part IV

Optimality conditions for the minimal time problem

Chapter 8

Extension of the nonlinear first order conditions

Abstract. In this chapter, we extend the results of Section 3.2 in order to tackle the minimal time problem for systems with complementarity constraints. A special focus is then made on LCS, and we investigate a bang-bang property.

This chapter focuses on finding first order conditions for the minimal time problem:

$$T^* = \min \quad T(x,u) \tag{8.1}$$

$$\text{s.t.} \quad \begin{cases} \dot{x}(t) = \phi(x(t),u(t)), \\ g(x(t),u(t)) \leq 0, \\ h(x(t),u(t)) = 0, & \text{a.e. on } [0,T(x,u)] \\ 0 \leq G(x(t),u(t)) \perp H(x(t),u(t)) \geq 0, \\ u(t) \in \mathcal{U} \\ (x(0),x(T(x,u))) = (x_0,x_f), \end{cases} \tag{8.2}$$

with $\phi : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^q$, $h : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^p$, $G, H : [t_0,t_1] \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^l$, $\mathcal{U} \subset \mathbb{R}^{m_u}$, $x_0, x_f \in \mathbb{R}^n$ given. We suppose that F and ϕ are $\mathcal{L} \times \mathcal{B}$-measurable, where $\mathcal{L} \times \mathcal{B}$ denotes the σ-algebra of subsets of appropriate spaces generated by product sets $M \times N$, where M is a Lebesgue ($\mathcal{L}$) measurable subset in $\mathbb{R}$, and N is a Borel ($\mathcal{B}$) measurable subset in $\mathbb{R}^n \times \mathbb{R}^m$.
We denote a solution of this problem by (T^*, x^*, u^*). If we don't bound u with the constraint $\mathcal{U}$, then the solution will most probably be $T^* = 0$ (i.e. an impulsive control, provided (8.2) is given a mathematical meaning).

The first order conditions given in [59] do not tackle this problem, since therein, the final time T^* is fixed beforehand. However, slight changes in the proof made for [109, Theorem 8.7.1] allow us to derive first order conditions for (8.1)(8.2). This chapter is organized as follows: first, the necessary conditions for (8.1)(8.2) will be derived. Then, we will show how these results are adapted to the problem of minimal time control for LCS, and some cases where the hypothesis are met. Finally, the result will be illustrated for a certain class of one dimensional LCS, deriving the analytical solution.

8.1 Necessary conditions

Since we have to compare different trajectories that are defined on different time-intervals, it should be understood that for $T > T^*$, a function w defined on $[0, T^*]$ is extended to $[0, T]$ by assuming constant extension: $w(t) = w(T^*)$ for all $t \in [T^*, T]$.

Definition 8.1.1.

- *We refer to any absolutely continuous function as an arc, and to any measurable function on $[0, T^*]$ as a control.*
- *An admissible pair for (8.1)(8.2) is a pair of functions (x, u) on $[t_0, t_1]$ for which u is a control and x is an arc, that satisfy all the constraints in (8.2).*
- *The complementarity cone is defined by*

$$\mathcal{C}^l = \{(v, w) \in \mathbb{R}^m \mid 0 \leq v \perp w \geq 0\}.$$

- *We define the set constraint by:*

$$S = \left\{(x, u) \in \mathbb{R}^n \times \mathcal{U} : g(x, u) \leq 0,\ h(x, u) = 0,\ (G(x, u), H(x, u)) \in \mathcal{C}^l\right\}.$$

- *We say that the local error bound condition holds (for the constrained system representing S) at $(\overline{x}, \overline{u}) \in S$ if there exist positive constants τ and δ such that:*

$$\mathrm{dist}_S(x, u) \leq \tau \left(\|\max\{0, g(x, u)\}\| + \|h(x, u)\| + \mathrm{dist}_{\mathcal{C}^l}(G(x, u), H(x, u))\right),\ \forall (x, u) \in \mathcal{B}_\delta(\overline{x}, \overline{u}).$$

- *For every given $t \in [t_0, t_1]$ and a positive constants R and ε, we define a neighbourhood of the point $(x^*(t), u^*(t))$ as:*

$$S_*^{\varepsilon,R}(t) = \{(x, u) \in S : \|x - x^*(t)\| \leq \varepsilon,\ \|u - u^*(t)\| \leq R\}. \tag{8.3}$$

(x^*, u^*) is a local minimizer of radius R if there exists ε such that for every pair (x, u) admissible for (8.1)(8.2) such that:

$$\|x^* - x\|_{W^{1,1}} = \|x^*(0) - x(0)\| + \int_0^{\min\{T(x,u), T(x^*,u^*)\}} \|\dot{x}^*(t) - \dot{x}(t)\| dt \leq \varepsilon,$$

$$\|u(t) - u^*(t)\| \leq R \text{ a.e. } [0, \min\{T(x, u), T(x^*, u^*)\}],$$

we have $T(x^*, u^*) \leq T(x, u)$.

We will have to do the following assumptions on the problem:

Assumption 8.1.1.

1. *There exist measurable functions k_x^ϕ, k_u^ϕ, such that for almost every $t \in [0, T^*]$ and for every (x^1, u^1), $(x^2, u^2) \in S_*^{\varepsilon,R}(t)$, we have:*

$$\|\phi(t, x^1, u^1) - \phi(t, x^2, u^2)\| \leq k_x^\phi(t)\|x^1 - x^2\| + k_u^\phi(t)\|u^1 - u^2\|. \tag{8.4}$$

2. *There exists a positive measurable function k_S such that for almost every $t \in [0, T^*]$, the bounded slope condition holds:*

$$(x, u) \in S_*^{\varepsilon,R}(t), (\alpha, \beta) \in \mathcal{N}_{S(t)}^P(x, u) \implies \|\alpha\| \leq k_S(t)\|\beta\|. \tag{8.5}$$

3. *The functions k_x^ϕ, and $k_S k_u^\phi$ are integrable, and there exists a positive number η such that $R \geq \eta k_S(t)$ a.e. $t \in [0, T^*]$.*

4. *ϕ is $\mathcal{L} \times \mathcal{B}$-measurable, g, h, G and H are strictly differentiable in variable (x, u).*

Let (x^*, u^*) be a local minimizer of (8.1)(8.2). In order to compute the first order condition of this problem, one introduces a new state variable (as inspired by [109]), absolutely continuous, which will represent time. For any $T > 0$, denote this variable $\tau : [0, T] \to [0, T^*]$, and let us introduce $\tilde{x} = x \circ \tau$, $\tilde{u} = u \circ \tau$. Then, for any $t \in [0, T]$:

$$\dot{\tilde{x}}(t) = \dot{\tau}(t)\dot{x}(\tau(t)) = \dot{\tau}(t)\phi(\tilde{x}(t), \tilde{u}(t)),$$

$$0 \leq G(\tilde{x}(t), \tilde{u}(t)) \perp H(\tilde{x}(t), \tilde{u}(t)) \geq 0.$$

This method is at the core of the proof for the following Theorem.

Define the sets

$$\begin{aligned} I_t^-(x, u) &= \{i \in \overline{q} \ : \ g_i(x(t), u(t)) < 0\}, \\ I_t^{+0}(x, u) &= \{i : G_i(x(t), u(t)) > 0 = H_i(x(t), u(t))\}, \\ I_t^{0+}(x, u) &= \{i : G_i(x(t), u(t)) = 0 < H_i(x(t), u(t))\}, \\ I_t^{00}(x, u) &= \{i : G_i(x(t), u(t)) = 0 = H_i(x(t), u(t))\}, \end{aligned}$$

and for any $(\lambda^g, \lambda^h, \lambda^G, \lambda^H) \in \mathbb{R}^{p+q+2m}$, denote:

$$\Psi(x, u; \lambda^g, \lambda^h, \lambda^G, \lambda^H) = g(x, u)^\intercal \lambda^g + h(x, u)^\intercal \lambda^h - G(x, u)^\intercal \lambda^G - H(x, u)^\intercal \lambda^H. \tag{8.6}$$

Theorem 8.1.1. *Suppose Assumption 8.1.1 holds. Let (x^*, u^*) be a local minimizer for (8.1)(8.2). If for almost every $t \in [0, T^*]$ the local error bound condition for the system representing S holds at $(x^*(t), u^*(t))$ (see Definition 8.1.1), then (x^*, u^*) is W-stationary; i.e. there exist an arc $p : [0, T^*] \to \mathbb{R}^n$, a scalar $\lambda_0 \in \{0, 1\}$ and multipliers $\lambda^g : [0, T^*] \to \mathbb{R}^q$, $\lambda^h : [0, T^*] \to \mathbb{R}^p$, $\lambda^G, \lambda^H : [0, T^*] \to \mathbb{R}^m$ such that:*

$$(\lambda_0, p(t)) \neq 0 \ \forall t \in [0, T^*], \tag{8.7a}$$

$$\begin{aligned} (\dot{p}(t), 0) \in \ & \partial^C \{\langle -p(t), \phi(\cdot, \cdot)\rangle\}(x^*(t), u^*(t)) \\ & + \nabla_{x,u} \Psi(x^*(t), u^*(t); \lambda^g(t), \lambda^h(t), \lambda^G(t), \lambda^H(t)) \\ & + \{0\} \times \mathcal{N}_{\mathcal{U}}^C(u^*(t)), \end{aligned} \tag{8.7b}$$

$$\lambda^g(t) \geq 0, \ \lambda_i^g(t) = 0, \ \forall i \in I_t^-(x^*, u^*), \tag{8.7c}$$

$$\lambda_i^G(t) = 0, \ \forall i \in I_t^{+0}(x^*, u^*), \tag{8.7d}$$

$$\lambda_i^H(t) = 0, \ \forall i \in I_t^{0+}(x^*, u^*), \tag{8.7e}$$

$$\lambda_0 = \langle p(t), \phi(x^*(t), u^*(t))\rangle. \tag{8.7f}$$

Moreover, the Weierstrass condition of radius R holds: for almost every $t \in [0, T^*]$:

$$(x^*(t), u) \in S, \ \|u - u^*(t)\| < R \implies \langle p(t), \phi(x^*(t), u)\rangle \leq \langle p(t), \phi(x^*(t), u^*(t))\rangle. \tag{8.7g}$$

Proof. Let $a \in \mathcal{C}^2([0,T^*],\mathbb{R}^n)$ be such that $a(0)=x^*(0)$ and $\|a-x^*\|_{W^{1,1}} = \int_0^{T^*} \|\dot{a}(t)-\dot{x}^*(t)\|dt < \frac{\varepsilon}{2}$. Let $b:[0,T^*]\to\mathbb{R}^m$ be a function such that $\|u^*(t)-b(t)\| < \frac{R}{2}$. Let us introduce the following fixed-end time optimal control problem:

$$\min \quad \tau(T^*) \tag{8.8}$$

$$\text{s.t.} \quad \begin{cases} \dot{\tilde{x}}(t) = \alpha(t)\phi(\tilde{x}(t),\tilde{u}(t)), \\ \dot{\tau}(t) = \alpha(t), \\ \dot{z}(t) = \alpha(t)\|\phi(\tilde{x}(t),\tilde{u}(t)) - \dot{a}(\tau(t))\| \\ g(\tilde{x}(t),\tilde{u}(t)) \le 0, \\ h(\tilde{x}(t),\tilde{u}(t)) = 0, \qquad \text{a.e. on } [0,T^*] \\ 0 \le G(\tilde{x}(t),\tilde{u}(t)) \perp H(\tilde{x}(t),\tilde{u}(t)) \ge 0, \\ \alpha(t) \in \left[\frac{1}{2},\frac{3}{2}\right], \\ \tilde{u}(t) \in \mathcal{U}, \\ \|\tilde{u}(t) - b(\tau(t))\| \le \frac{R}{2}, \\ (\tilde{x}(0),\tilde{x}(T^*)) = (x_0,x_f), \\ \tau(0)=0,\ |z(0)-z(T^*)| \le \frac{\varepsilon}{2}. \end{cases} \tag{8.9}$$

Denote by $(\tilde{x},\tau,z,\tilde{u},\tilde{v},\alpha)$ an admissible trajectory for (8.8)(8.9), where $(\tilde{x},\tau,z)$ are state variables and $(\tilde{u},\tilde{v},\alpha)$ are controls. We claim that a minimizer for this problem is

$$\left(x^*,\ \tau^*: t\mapsto t,\ z: t\mapsto \int_0^t \|\dot{x}^*(s)-\dot{a}(s)\|ds,\ u^*,\ v^*,\ \alpha^*\equiv 1\right) \tag{8.10}$$

with minimal cost $\tau^*(T^*)=T^*$. To prove this, let us assume that another admissible trajectory $(\tilde{x},\tau,z,\tilde{u},\tilde{v},\alpha)$ has a lower cost $T=\tau(T^*)<T^*$. Therefore, $\tau(0)=0$, $\tau(t)=\int_0^t \alpha(s)ds$, and since $\alpha>0$, τ is a continuous strictly increasing function from $[0,T^*]$ to $[0,T]$. Hence it admits an inverse τ^{-1}. Define on $[0,T]$:

$$x = \tilde{x}\circ\tau^{-1}, u = \tilde{u}\circ\tau^{-1}$$

and extend these functions to $[0,T^*]$ by assuming that $x(t)=x(T)$ for all $t\in[T,T^*]$ (the same goes for u). Obviously, (x,u) is an admissible trajectory for (8.1)(8.2), with minimal time T. Also, it is in the neighborhood of (x^*,u^*), since for almost all $t\in[0,T]$,

$$\begin{aligned} \|u^*(t)-u(t)\| &\le \|u^*(t)-b(t)\| + \|u(t)-b(t)\| \\ &\le \frac{R}{2} + \|\tilde{u}(\sigma)-b(\tau(\sigma))\| \text{ where } t=\tau(\sigma) \\ &\le R \end{aligned}$$

and:

$$\begin{aligned}
\|x - x^*\|_{W^{1,1}} &= \int_0^T \|\dot{x}(t) - \dot{x}^*(t)\| dt \\
&\leq \int_0^T \|\dot{x}(t) - \dot{a}(t)\| dt + \int_0^T \|\dot{x}^*(t) - \dot{a}(t)\| dt \\
&\leq \int_0^T \|\phi(x(t), u(t)) - \dot{a}(t)\| dt + \frac{\varepsilon}{2} \\
&\leq \int_0^{T^*} \|\phi(\tilde{x}(\sigma), \tilde{u}(\sigma)) - \dot{a}(\tau(\sigma))\| \dot{\tau}(\sigma) d\sigma + \frac{\varepsilon}{2} \\
&\leq \int_0^{T^*} \|\phi(\tilde{x}(\sigma), \tilde{u}(\sigma)) - \dot{a}(\tau(\sigma))\| \alpha(\sigma) d\sigma + \frac{\varepsilon}{2} \\
&\leq z(T^*) - z(0) + \frac{\varepsilon}{2} \\
&\leq \varepsilon.
\end{aligned}$$

Therefore, since they are in the same neighbourhood, the two trajectories can be compared. Since (x^*, u^*) is supposed to be a local minimizer for (8.1)(8.2), we should have $T^* \leq T$. This is a contradiction, so the claim that (8.10) is the minimizer was right.
Remark that since we supposed that Assumption 8.1.1 is verified, the same assumptions adapted for problem (8.8)(8.9) are also valid. Therefore, the results of [59, Theorem 3.2] for (8.8)(8.9) state that there exists an arc $p : [0, T^*] \to \mathbb{R}^n$, a scalar $\lambda_0 \in \{0, 1\}$ and multipliers $\lambda^g : [0, T^*] \to \mathbb{R}^q$, $\lambda^h : [0, T^*] \to \mathbb{R}^p$, $\lambda^G, \lambda^H : [0, T^*] \to \mathbb{R}^m$ such that (8.7a)-(8.7e) hold, along with the Weierstrass condition (8.7g).
Notice that since $z^*(T^*) - z^*(0) < \frac{\varepsilon}{2}$ and $\|u^*(t) - b(t)\| < \frac{R}{2}$ (the constraints are inactive), these inequalities do not appear in the first order conditions (the normal cone associated with these constraints reduces to $\{0\}$). Also, one could argue that there should be an adjoint state associated with z, but simple calculations show that it is identically 0.
Moreover, we should have another arc p_τ associated with τ, but it must comply with $\dot{p_\tau} \equiv 0$, $p_\tau(T^*) = -\lambda_0$, such that $p_\tau \equiv -\lambda_0$. Also, the stationary inclusion associated with α leads to $0 \in -\langle p, \phi(x^*(t), u^*(t))\rangle - p_\tau(t) + \mathcal{N}^C_{[\frac{1}{2}, \frac{3}{2}]}(\alpha(t))$. But since $\alpha \equiv 1$ and $1 \in \left]\frac{1}{2}, \frac{3}{2}\right[$, $\mathcal{N}^C_{[\frac{1}{2}, \frac{3}{2}]}(\alpha(t)) = \{0\}$ for almost every $t \in [0, T^*]$, and so, it yields (8.7f). □

8.2 Application to LCS

8.2.1 Sufficient condition for the bounded slope condition

These results still rely on assumptions, among which the bounded slope condition is a stringent, non-intuitive, and hard to verify condition. A sufficient condition for the bounded slope condition to hold is given by [59, Proposition 3.7], and recalled in Proposition 3.2.1. We give some cases for

which this condition holds when the underlying system is an LCS:

$$T^* = \min \quad T(x,u,v) \tag{8.11}$$

$$\text{s.t.} \quad \begin{cases} \dot{x}(t) = Ax(t) + Bv(t) + Fu(t), \\ 0 \leq v(t) \perp Cx(t) + Dv(t) + Eu(t) \geq 0, \quad \text{a.e. on } [0,T^*] \\ u(t) \in \mathcal{U} \\ (x(0), x(T^*)) = (x_0, x_f), \end{cases} \tag{8.12}$$

where $A \in \mathbb{R}^{n\times n}$, $D \in \mathbb{R}^{m\times m}$, $E \in \mathbb{R}^{m\times m_u}$, $B \in \mathbb{R}^{n\times m}$, $F \in \mathbb{R}^{n\times m_u}$, $C \in \mathbb{R}^{m\times n}$, $\mathcal{U} \subseteq \mathbb{R}^{m_u}$. Since u is constrained ($u(t) \in \mathcal{U}$), in contrast to Chapter 6 for the Linear Quadratic case, the MPEC LICQ cannot be used, and therefore we can not prove easily the existence of an S-stationary trajectory with the same multipliers λ^G, λ^H. However, a direct application of [59, Theorem 3.5(b)] to (8.8)(8.9) proves the following proposition.

Proposition 8.2.1. *Suppose Assumption 8.1.1 for the problem (8.11)(8.12) holds. Suppose also that $\mathcal{U}$ is a union of finitely many polyhedral sets. Let (x^*, u^*, v^*) be a local minimizer for (8.11)(8.12). Then, (x^*, u^*, v^*) is M-stationary, meaning it is W-stationary with arc p, and moreover, there exist measurable functions $\eta^G, \eta^H : [0, T^*] \to \mathbb{R}^m$ such that:*

$$0 = B^\intercal p + D^\intercal \eta^H + \eta^G,$$

$$0 \in -F^\intercal p - E^\intercal \eta^H + \mathcal{N}_{\mathcal{U}}^C(u^*(t)),$$

$$\eta_i^G(t) = 0, \ \forall i \in I_t^{+0}(x^*, u^*, v^*),$$

$$\eta_i^H(t) = 0, \ \forall i \in I_t^{0+}(x^*, u^*, v^*),$$

$$\eta_i^G \eta_i^H = 0 \text{ or } \eta_i^G > 0, \ \eta_i^H > 0, \ \forall i \in I_t^{00}(x^*, u^*, v^*).$$

Since the system is linear, [59, Proposition 2.3] asserts that the local error bound condition holds at every admissible point. There is one case for which one can check that the bounded slope condition hold: when $\mathcal{U} = \mathbb{R}^{m_u}$, as proved in Proposition 6.1.1. However, this case of unbounded $\mathcal{U}$ is rather unrealistic, since it could lead to $T^* = 0$ (in the sense that the target x_f can be reached from x_0 given any positive time $T^* > 0$; see for instance [74] for an example). When one attempts to add a constraint $\mathcal{U}$ to the previous proof, Proposition 3.2.1 adds a normal cone that prevents checking the inequality, unless one supposes that the optimal trajectory is inside an R-neighbourhood which lies in the interior of $\mathcal{U}$.

Nonetheless, there are two cases when (8.12) verifies the bounded slope condition, even with constraints on u.

Proposition 8.2.2. *Suppose:*

- *either $C = 0$,*
- *or D is a diagonal matrix with positive entries.*

Then the bounded slope condition for (8.12) holds.

Proof. The case when $C = 0$ is obvious, when one applies directly Proposition 3.2.1. Assume $D =$diag$(d_1, ..., d_m)$, where the $d_i > 0$, $i \in \overline{m}$, are the diagonal entries of D, and diag means that D is a diagonal matrix built with these entries. First of all, remark that:

$$\forall \lambda^H \in \mathbb{R}^m,\ \|C^\intercal \lambda^H\| \leq \|C^\intercal D^{-1}\| \|D\lambda^H\|.$$

Now, for $t \in [0, T^*]$ and $(x, u) \in S_*^{\varepsilon,R}(t)$, take λ^G and λ^H in $\mathbb{R}^m$ such that:

$$\lambda_i^G = 0,\ \forall i \in I_t^{+0}(x,u),\ \lambda_i^H = 0,\ \forall i \in I_t^{0+}(x,u),$$

$$\lambda_i^G > 0,\ \lambda_i^H > 0,\ \text{or } \lambda_i^G \lambda_i^H = 0,\ \forall i \in I_t^{00}(x,u).$$

It yields:

$$\|D\lambda^H\|^2 = \sum_{i \in I_t^{+0}(x,u)} (D_{ii}\lambda_i^H)^2 + \sum_{i \in I_t^{0+}(x,u)} (D_{ii}\lambda_i^H)^2 + \sum_{i \in I_t^{00}(x,u)} (D_{ii}\lambda_i^H)^2$$

One can easily see that:

$$\sum_{i \in I_t^{+0}(x,u)} (D_{ii}\lambda_i^H)^2 = \sum_{i \in I_t^{+0}(x,u)} (D_{ii}\lambda_i^H + \lambda_i^G)^2,$$

$$0 = \sum_{i \in I_t^{0+}(x,u)} (D_{ii}\lambda_i^H)^2 \leq \sum_{i \in I_t^{0+}(x,u)} (D_{ii}\lambda_i^H + \lambda_i^G)^2,$$

$$\forall i \in I_t^{00}(x,u),\ (D_{ii}\lambda_i^H)^2 \leq (D_{ii}\lambda_i^H)^2 + (\lambda_i^G)^2 + 2D_{ii}\lambda_i^H\lambda_i^G = (D_{ii}\lambda_i^H + \lambda_i^G)^2.$$

Therefore, it yields:

$$\|D\lambda^H\|^2 \leq \sum_{i=1}^{m} (D_{ii}\lambda_i^H + \lambda_i^G)^2 = \|D^\intercal \lambda^H + \lambda^G\|^2.$$

One finally proves: $\forall \zeta \in \mathcal{N}_{\mathcal{U}}(u)$,

$$\|C^\intercal \lambda^H\| \leq \|C^\intercal D^{-1}\| \|D^\intercal \lambda^H + \lambda^G\| \leq \|C^\intercal D^{-1}\| \left\| \begin{matrix} D^\intercal \lambda^H + \lambda^G \\ E^\intercal \lambda^H + \zeta \end{matrix} \right\|.$$

Using Proposition 3.2.1, we see that (8.12) complies with the bounded slope condition. □

8.2.2 A bang-bang property

Reachable set for linear systems

We turn ourselves to the reachability set of linear systems in order to state a result that will be useful in order to prove a bang-bang property for LCS. Consider the following system:

$$\begin{cases} \dot{x}(t) = Mx(t) + Nu(t), \\ u(t) \in \mathcal{V}, \end{cases} \tag{8.13}$$

for some matrices $M \in \mathbb{R}^{n \times n}$ and $N \in \mathbb{R}^{n \times m}$. We define the reachable (or accessible) set from $x_0 \in \mathbb{R}^n$ at time $t \geq 0$, with controls taking values in $\mathcal{V}$, denoted by $\text{Acc}_{\mathcal{V}}(x_0, t)$, the set of points $x(t)$, where $x : [0, t] \to \mathbb{R}^n$ is a solution of (8.13), with $u(s) \in \mathcal{V}$ for almost all $s \in [0, t]$ and $x(0) = x_0$. As stated in [103, Corollary 2.1.2], which is proved using Aumann's theorem (see for instance [39]), the following Proposition shows that the set of constraints $\mathcal{V}$ can be embedded in its convex hull:

Proposition 8.2.3. *[103, Corollary 2.1.2] Suppose that $\mathcal{V}$ is compact. Then:*

$$\mathrm{Acc}_{\mathcal{V}}(x_0,t) = \mathrm{Acc}_{\mathrm{conv}(\mathcal{V})}(x_0,t)$$

where $\mathrm{conv}(\mathcal{V})$ *denotes the convex hull of* $\mathcal{V}$.

Thanks to Krein-Milman's Theorem (see Appendix B), this justifies that minimal-time optimal controls can be searched as bang-bang controls (meaning, u only takes values that are extremal points of $\mathcal{V}$ if one supposes in addition that $\mathcal{V}$ is convex).

Extremal points for LCS

For this section, let us state the following Assumption:

Assumption 8.2.1. *In (8.12), $C = 0$, D is a $\boldsymbol{P}$-matrix, and $\mathcal{U}$ is a finite union of polyhedral compact convex sets.*

As it can be expected for a minimal time problem with linear dynamics, a bang-bang property can be proved, where the *bang-bang controls* have to be properly defined. Let us define first some notions. Denote by Ω the constraints on the controls (u,v) in (8.12), meaning:

$$\Omega = \{(u,v) \in \mathcal{U} \times \mathbb{R}^m | 0 \leq v \perp Dv + Eu \geq 0\}. \tag{8.14}$$

The set $\mathrm{Acc}_{\Omega}(x_0,t)$ denotes the reachable set from $x_0 \in \mathbb{R}^n$ at time $t \geq 0$ with controls with values in Ω. For a convex set $\mathfrak{C}$, a point $c \in \mathfrak{C}$ is called an extreme point if $\mathfrak{C}\backslash\{c\}$ is still convex. This is equivalent to say that:

$$c_1, c_2 \in \mathfrak{C},\ c = \frac{c_1+c_2}{2} \implies c = c_1 = c_2.$$

The set of extreme points of $\mathfrak{C}$ will be denoted $\mathrm{Ext}(\mathfrak{C})$.

Suppose Ω is compact (which is not necessarily the case: take for instance $D = 0$ with $0 \in \mathcal{U}$). Applying Proposition 8.2.3, one proves that:

$$\mathrm{Acc}_{\Omega}(x_0,t) = \mathrm{Acc}_{\mathrm{conv}(\Omega)}(x_0,t).$$

The set Ω is not convex and has empty interior; finding its boundary or extreme points is not possible in this case. However, Krein-Milman's Theorem (see Appendix B) proves that $\mathrm{conv}(\Omega)$ can be generated by its extreme points. In what follows, we will prove that the extreme points of $\mathrm{conv}(\Omega)$ are actually points of Ω that can be easily identified from the set $\mathcal{U}$.

For an index set $\alpha \subseteq \overline{m}$, denote by $\mathbb{R}^m_\alpha$ the set of points q in $\mathbb{R}^m$ such that $q_\alpha \geq 0$, $q_{\overline{m}\backslash\alpha} \leq 0$, and define $E^{-1}\mathbb{R}^m_\alpha = \{\tilde{u} \in \mathbb{R}^{m_u} | E\tilde{u} \in \mathbb{R}^m_\alpha\}$ (E is not necessarily invertible).

Lemma 8.2.1. *Suppose Assumption 8.2.1 holds true. For a certain $\alpha \subseteq \overline{m}$, denote by $\mathcal{P}_\alpha$ the set:*

$$\mathcal{P}_\alpha = \{(u,v) \in \left(\mathcal{U} \cap E^{-1}\mathbb{R}^m_\alpha\right) \times \mathbb{R}^m | v_\alpha = 0,\ D_{\overline{\alpha}\bullet}v + E_{\overline{\alpha}\bullet}u = 0,\ v \geq 0, Dv + Eu \geq 0\},$$

and by $\mathcal{E}_\alpha$ the set:

$$\mathcal{E}_\alpha = \{(u,v) \in \mathrm{Ext}\left(\mathcal{U} \cap E^{-1}\mathbb{R}^m_\alpha\right) \times \mathbb{R}^m | v_\alpha = 0,\ D_{\overline{\alpha}\bullet}v + E_{\overline{\alpha}\bullet}u = 0,\ v \geq 0, Dv + Eu \geq 0\}.$$

Then $Ext(\mathcal{P}_\alpha) = \mathcal{E}_\alpha$.

Proof. If $\mathcal{U} \cap E^{-1}\mathbb{R}^m_\alpha$ is empty, then the equality is obvious. Choose α such that $\mathcal{U} \cap E^{-1}\mathbb{R}^m_\alpha$ is not empty.

- $\mathcal{E}_\alpha \subseteq \text{Ext}(\mathcal{P}_\alpha)$: Let $(u, v) \in \mathcal{E}_\alpha$. Suppose that $(u, v) \notin \text{Ext}(\mathcal{P}_\alpha)$. Thus, there exist (u^1, v^1) and (u^2, v^2) in $\mathcal{P}_\alpha$, both different than (u, v), such that $(u, v) = \frac{1}{2}[(u^1, v^1) + (u^2, v^2)]$. But this implies that $u = \frac{1}{2}(u^1 + u^2)$, and since $u \in \text{Ext}(\mathcal{U} \cap E^{-1}\mathbb{R}^m_\alpha)$, $u = u_1 = u_2$. Therefore, since D is a **P**-matrix, $v = \text{SOL}(D, Eu) = \text{SOL}(D, Eu^i) = v^i$ for $i \in \{1, 2\}$. Therefore, $(u, v) = (u^1, v^1) = (u^2, v^2)$, and (u, v) is an extremal point of $\mathcal{P}_\alpha$. This is a contradiction.

- $\text{Ext}(\mathcal{P}_\alpha) \subseteq \mathcal{E}_\alpha$: Let $(u, v) \in \text{Ext}(\mathcal{P}_\alpha)$. Suppose that $u \notin \text{Ext}(\mathcal{U} \cap E^{-1}\mathbb{R}^m_\alpha)$. Therefore, there exists u^1 and u^2 in $\mathcal{U} \cap E^{-1}\mathbb{R}^m_\alpha$, different than u, such that $u = \frac{1}{2}(u^1 + u^2)$. Define for $i \in \{1, 2\}$ $v^i = \text{SOL}(D, Eu^i)$. Since Eu^1 and Eu^2 are members of $\mathbb{R}^m_\alpha$, for $i \in \{1, 2\}$:

  $$v^i_{\overline{\alpha}} = -(D_{\overline{\alpha}\overline{\alpha}})^{-1}(Eu^i)_{\overline{\alpha}}, \; v^i_\alpha = 0.$$

  So:

  $$v_{\overline{\alpha}} = -(D_{\overline{\alpha}\overline{\alpha}})^{-1}(Eu)_{\overline{\alpha}} = \frac{1}{2}\left(v^1_{\overline{\alpha}} + v^2_{\overline{\alpha}}\right),$$

  $$v_\alpha = 0 = \frac{1}{2}\left(v^1_\alpha + v^2_\alpha\right).$$

  So $(u, v) = \frac{1}{2}\left[(u^1, v^1) + (u^2, v^2)\right]$ with $(u^i, v^i) \in \mathcal{P}_\alpha$, $i \in \{1, 2\}$. But since $(u, v) \in \text{Ext}(\mathcal{P}_\alpha)$, $u = u^1 = u^2$. This is a contradiction.

□

Remark 8.2.1. *If* $\ker(E) = \{0\}$ *(and in particular, if* E *is invertible), it may be easier to search for extreme points of the set* $E\mathcal{U} \cap \mathbb{R}^m_\alpha$*, as one can prove easily that:*

$$\text{Ext}(\mathcal{U} \cap E^{-1}\mathbb{R}^m_\alpha) = E^{-1}\text{Ext}(E\mathcal{U} \cap \mathbb{R}^m_\alpha)$$

Proposition 8.2.4. *Suppose Assumption 8.2.1 holds true. Denote by* $\mathcal{E}$ *the set* $\mathcal{E} = \bigcup_{\alpha \subseteq \overline{m}} \mathcal{E}_\alpha$*, where* $\mathcal{E}_\alpha$ *is defined in Lemma 8.2.1. Then, for all* $t > 0$ *and all* $x_0 \in \mathbb{R}^n$*,*

$$\text{Acc}_\Omega(x_0, t) = \text{Acc}_\mathcal{E}(x_0, t),$$

where Ω *is defined in (8.14).*

Proof. The function $\text{SOL}(D, \cdot) : q \mapsto v = \text{SOL}(D, q)$ is piecewise linear and continuous, according to Proposition 1.1.2. The pieces of $\text{SOL}(D, \cdot)$ are the sets $\mathbb{R}^m_\alpha$, for α ranging over the subsets of $\overline{m}$. Therefore, for each $\alpha \subseteq \overline{m}$, the set $\mathcal{U} \cap E^{-1}\mathbb{R}^m_\alpha$ is the union of compact convex polyhedra (possibly empty), and therefore it admits a finite number of extreme points. Thus, each $\mathcal{P}_\alpha = \{(u, \text{SOL}(D, Eu)) \mid u \in \mathcal{U} \cap E^{-1}\mathbb{R}^m_\alpha\}$ in Lemma 8.2.1 is the union of compact convex polyhedra. In order to simplify the proof, suppose that $\mathcal{U}$ (and therefore, each non empty $\mathcal{P}_\alpha$) is a single compact convex polyhedron (and not a union of several; the proof would still be the same by reasoning on each of them). Therefore, by Lemma 8.2.1 and Krein-Milman's theorem (see Appendix B), $\mathcal{P}_\alpha = \text{conv}(\mathcal{E}_\alpha)$ for each subset α of $\overline{m}$. Since it can be shown that $\Omega = \bigcup_{\alpha \subseteq \overline{m}} \mathcal{P}_\alpha$, it proves:

$$\text{conv}(\Omega) = \text{conv}\left(\bigcup_{\alpha \subseteq \overline{m}} \mathcal{P}_\alpha\right).$$

Let us prove that $\text{conv}(\Omega) = \text{conv}(\mathcal{E})$:

- $\text{conv}(\Omega) \subseteq \text{conv}\big(\bigcup_{\alpha\subseteq\overline{m}} \mathcal{E}_\alpha\big)$: Remark that for all $\beta \subseteq \overline{m}$, $\mathcal{E}_\beta \subseteq \bigcup_{\alpha\subseteq\overline{m}} \mathcal{E}_\alpha$. Therefore, $\mathcal{P}_\beta = \text{conv}(\mathcal{E}_\beta) \subseteq \text{conv}\big(\bigcup_{\alpha\subseteq\overline{m}} \mathcal{E}_\alpha\big)$, and thus, since β was arbitrary, $\Omega = \bigcup_{\alpha\subseteq\overline{m}} \mathcal{P}_\alpha \subseteq \text{conv}\big(\bigcup_{\alpha\subseteq\overline{m}} \mathcal{E}_\alpha\big)$. It then leads to: $\text{conv}(\Omega) \subseteq \text{conv}\big(\bigcup_{\alpha\subseteq\overline{m}} \mathcal{E}_\alpha\big)$.
- $\text{conv}\big(\bigcup_{\alpha\subseteq\overline{m}} \mathcal{E}_\alpha\big) \subseteq \text{conv}(\Omega)$:

$$\begin{aligned} \forall \beta \subseteq \overline{m},\ \mathcal{E}_\beta \subseteq \text{conv}(\mathcal{E}_\beta) = \mathcal{P}_\beta \quad &\Longrightarrow \quad \bigcup_{\alpha\subseteq\overline{m}} \mathcal{E}_\alpha \subseteq \bigcup_{\alpha\subseteq\overline{m}} \mathcal{P}_\alpha = \Omega \\ &\Longrightarrow \quad \text{conv}\left(\bigcup_{\alpha\subseteq\overline{m}} \mathcal{E}_\alpha\right) \subseteq \text{conv}(\Omega). \end{aligned}$$

Applying now Proposition 8.2.3, it proves the following equalities:

$$\text{Acc}_{\mathcal{E}}(x_0, t) = \text{Acc}_{\text{conv}(\mathcal{E})}(x_0, t) = \text{Acc}_{\text{conv}(\Omega)}(x_0, t) = \text{Acc}_{\Omega}(x_0, t).$$

□

The interest of Proposition 8.2.4 is twofold: first, the complementarity constraints does not affect the bang-bang property that is shared with linear system (it is preserved even for this kind of piecewise linear system); secondly, it is actually sufficient to search for the extreme points of $\mathcal{U} \cap E^{-1}\mathbb{R}^m_\alpha$, as proved in Lemma 8.2.1 with the sets $\mathcal{E}_\alpha$. This result is illustrated in the next examples.

Example 8.2.1.

$$T^* = \min \quad T(x, u, v) \tag{8.15}$$

$$s.t. \quad \begin{cases} \dot{x}(t) = ax(t) + bv(t) + fu(t), \\ 0 \le v(t) \perp dv(t) + eu(t) \ge 0, \quad a.e.\ on\ [0, T^*] \\ u(t) \in \mathcal{U} = [-1, 1] \\ (x(0), x(T^*)) = (x_0, x_f), \end{cases} \tag{8.16}$$

where a, b, d, f, e are scalars, and we suppose $d > 0$ and $e \neq 0$. We suppose also that there exist at least one trajectory stirring x_0 to x_f.

In this case, there are two index sets α as described in Lemma 8.2.1: $\emptyset$ or $\{1\}$. Therefore, we should have a look at the extreme points of $\mathcal{U} \cap \mathbb{R}^1_\emptyset = \mathcal{U} \cap \mathbb{R}_- = [-1, 0]$ and of $\mathcal{U} \cap \mathbb{R}^1_{\{1\}} = \mathcal{U} \cap \mathbb{R}_+ = [0, 1]$. Thus, it is sufficient to look at input functions u with values in $\{-1, 0, 1\}$. Suppose that the constants in (8.16) (with $u(t)$ supposed unconstrained for the moment) is completely controllable, which means:

If $e > 0$ **:**
- *if $f < 0$, then $b \ge 0$ or $[b < 0$ and $b - \frac{fd}{e} > 0]$.*
- *if $f > 0$, then $b \le 0$ or $[b > 0$ and $b - \frac{fd}{e} < 0]$.*

If $e < 0$ **:** *the same cases as with $e > 0$ hold by inverting the sign of f.*

All other cases (like $f = 0$ or $e = 0$) are discarded.

Let us now deduce from Theorem 8.1.1 the only stationary solution. First of all, the equation (8.7b) tells us that the adjoint state complies with the ODE $\dot{p} = -ap$. Therefore there exists p_0 such that:

$$p(t) = p_0 e^{-at},\ \forall t \in [0, T^*].$$

Could we have $p_0 = 0$? It would imply that $p \equiv 0$ and then, $\lambda_0 = \langle p(t), ax(t) + bv(t) + fu(t) \rangle = 0$, so $(p(t), \lambda_0) = 0$ for almost all t in $[0, T^*]$. This is not allowed, so $p_0 \neq 0$. Moreover, there exist multipliers λ^G and λ^H such that, for almost all t in $[0, T^*]$:

$$\lambda^G(t) = -bp(t) - d\lambda^H(t), \tag{8.17}$$

$$\lambda^G(t) = 0 \text{ if } v(t) > 0 = dv(t) + eu(t), \tag{8.18}$$

$$\lambda^H(t) = 0 \text{ if } v(t) = 0 < dv(t) + eu(t), \tag{8.19}$$

$$fp(t) + e\lambda^H(t) \in \mathcal{N}_{[-1,1]}(u(t)) = \begin{cases} \{0\} & \text{if } |u(t)| \neq 1, \\ \mathbb{R}^+ & \text{if } u(t) = 1, \\ -\mathbb{R}^+ & \text{if } u(t) = -1, \end{cases} \tag{8.20}$$

$$\begin{aligned} \lambda_0 = \quad & \max \ \langle p(t), ax(t) + b\tilde{v} + f\tilde{u} \rangle, \\ & s.t. \ 0 \leq \tilde{v} \perp d\tilde{v} + e\tilde{u} \geq 0 \end{aligned} \tag{8.21}$$

(all these equations are derived from (8.7) with $g \equiv h \equiv 0$, $G \equiv v$, $H \equiv Cx + Dv + Eu$, $\phi \equiv Ax + Bv + Fu$). Since $d > 0$, one can easily prove that:

$$v = \frac{1}{d} \max(0, -eu).$$

Let us suppose for now that $e > 0$ (all subsequent work is easily adapted for $e < 0$ by replacing u by $-u$). In this case:

$$v = \begin{cases} 0 & \text{if } u \in [0, 1], \\ -\frac{eu}{d} & \text{if } u \in [-1, 0]. \end{cases} \tag{8.22}$$

Let us now discuss all possible cases for u. If $u(t) = 0$, then $v(t) = 0 = dv(t) + eu(t)$. We use stationarity conditions for (8.21): since the MPEC Linear Condition holds, there exists multipliers η^H and η^G such that $\eta^H = -\frac{f}{e}p(t)$, $\eta^G = \left(\frac{df}{e} - b\right)p(t)$, and

$$\eta^G \eta^H = 0 \text{ or } \eta^H > 0, \eta^G > 0$$

(one has M-stationarity, see Definition 3.1.3). However, $\eta^H \neq 0$ and $\eta^G \neq 0$. Furthermore, η^H has the same sign as $-fp_0$ and η^G has the same sign as fp_0. Therefore, the two have opposite signs, and $u(t) = 0$ can not be an M-stationary solution. Therefore, it proves that necessarily, the optimal control u^* complies with $|u^*(t)| = 1$ for almost all t on $[0, T^*]$.

- *If $u(t) = 1$, then $v(t) = 0 < dv(t) + eu(t)$, and by (8.19), $\lambda^H(t) = 0$. Then by (8.20), $fp_0 \geq 0$.*
- *If $u(t) = -1$, then $v(t) > 0 = dv(t) + eu(t)$, and by (8.17)(8.18), $\lambda^H(t) = -\frac{b}{d}p(t)$. Then by (8.20), $\left(f - \frac{eb}{d}\right)p_0 \leq 0$.*

It is impossible to have $fp_0 \geq 0$ and $\left(f - \frac{eb}{d}\right)p_0 \leq 0$ at the same time, since f and $f - \frac{eb}{d}$ have the same sign by the complete controllability conditions. Therefore, u^* take only one value along $[0, T^*]$: 1 or -1. Then we have two possible optimal state x^* starting from x_0:

if $a \neq 0$**:**

$$x^*(t) = \begin{cases} \left(x_0 + \frac{f}{a}\right)\exp(at) - \frac{f}{a} & \text{if } u^*(t) = 1, \\ \left(x_0 + \frac{be - fd}{ad}\right)\exp(at) - \frac{be - fd}{ad} & \text{if } u^*(t) = -1. \end{cases}$$

One must then find the solution that complies with $x(T^*) = x_f$. One can isolate the optimal time T^*:

$$T^* = \begin{cases} \frac{1}{a}\ln\left(\frac{ax_f+f}{ax_0+f}\right) & \text{if } ax_0+f \neq 0 \text{ and } \frac{ax_f+f}{ax_0+f} > 0, \\ \frac{1}{a}\ln\left(\frac{adx_f+be-fd}{adx_0+be-fd}\right) & \text{if } adx_0+be-fd \neq 0 \text{ and } \frac{adx_f+be-fd}{adx_0+be-fd} > 0. \end{cases} \tag{8.23}$$

Since we supposed that there exists at least one trajectory stirring x_0 to x_f, one of these two expressions of T^* must be positive. Therefore, one can infer that:

$$u^* \equiv \begin{cases} 1 \text{ if } ax_0+f \neq 0 \text{ and } \frac{ax_f+f}{ax_0+f} > 0, \\ -1 \text{ if } adx_0+be-fd \neq 0 \text{ and } \frac{adx_f+be-fd}{adx_0+be-fd} > 0. \end{cases}$$

if $a = 0$**:**

$$x^*(t) = \begin{cases} ft & \text{if } u^*(t) = 1, \\ \frac{be-fd}{d}t & \text{if } u^*(t) = -1. \end{cases}$$

With the same calculations made in the case $a \neq 0$, one proves that:

$$u^* \equiv \begin{cases} 1 \text{ if } f \neq 0 \text{ and } \frac{x_f}{f} > 0, \\ -1 \text{ if } be-fd \neq 0 \text{ and } \frac{dx_f}{be-fd} > 0. \end{cases}$$

The proof of Proposition 8.2.4 relies on the fact that when D is a **P**-matrix, Ω is the union of compact convex polyhedra. However, some examples show that even when D is not a **P**-matrix, then this property may hold.

Example 8.2.2.

$$T^* = \min \quad T(x,u,v) \tag{8.24}$$

$$s.t. \quad \begin{cases} \dot{x}(t) = ax(t) + bv(t) + fu(t), \\ 0 \leq v(t) \perp -v(t) + u(t) \geq 0, \quad a.e. \text{ on } [0,T^*], \\ u(t) \in \mathcal{U} = [-1,1], \\ (x(0), x(T^*)) = (x_0, x_f). \end{cases} \tag{8.25}$$

It is clear that there exists no solution to the LCP$(-1,u)$ appearing in (8.25) for $u \in [-1,0)$, so $\mathcal{U}$ can actually be restricted to $[0,1]$. A graphic showing the shape of Ω and its convex hull is shown in Figure 8.1.

It clearly appears that conv(Ω) is generated by three extreme points:

$$(u,v) \in \mathcal{E} = \{(0,0),(1,0),(1,1)\}.$$

Exactly the same way as in the proof of Proposition of 8.2.4, one can simply show that for all $t \geq 0$ and for all $x_0 \in \mathbb{R}$, $Acc_\Omega(x_0,t) = Acc_\mathcal{E}(x_0,t)$. Therefore, the optimal trajectory can be searched with controls (u,v) with values in $\mathcal{E}$. It is also interesting to note that this bang-bang property can be guessed from the condition of maximisation of the Hamiltonian in (8.7g) and from Figure 8.1. Indeed, (8.7g) state that at almost all time t, the linear function $\Lambda : (u,v) \mapsto \langle p(t), Bv+Fu \rangle$ must be maximized with variables (u,v) in Ω. When one tries to maximize Λ over conv(Ω) it becomes a Linear Program (LP) over a simplex. It is well known that linear functions reach their optimum over simplexes at extreme points; in this case, the extreme points are the points of $\mathcal{E}$.

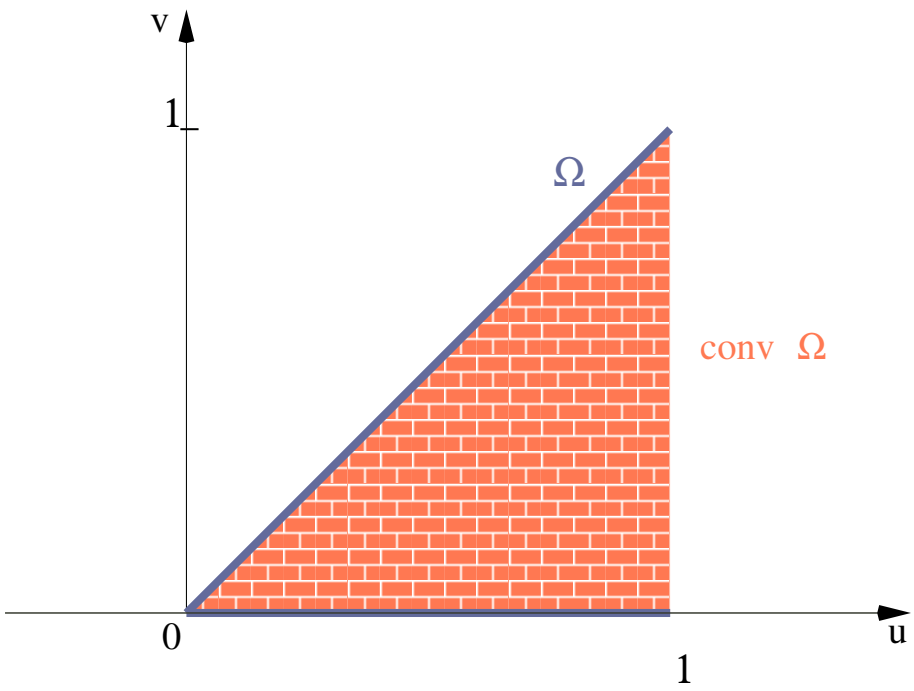


Figure 8.1: Ω and its convex hull for (8.25)

8.2.3 Characterisation through HJB equation

An other way to solve the minimal time optimal control problem is through the Dynamic Programming Principle and the Hamilton-Jacobi-Bellman (HJB) equation. The theory needs pure control constraints, but not convexity of the set of constraints. Therefore, the Assumption that $C = 0$ in (8.12) still holds. However, one doesn't need D to be a **P**-matrix anymore. The only necessary Assumption needed is an assumption of compactness.

Assumption 8.2.2. *In (8.12), $C = 0$, and the set Ω defined in (8.14) is a compact subset of $\mathbb{R}^{m_u} \times \mathbb{R}^m$.*

The HJB equation is a non-linear PDE that the objective cost must comply with. In this framework, the minimal time T^* is seen as a function of the target x_f. However, the equation will not be directly met by $T^*(x_f)$, but by a discounted version of it, called the Kružkov transform, and defined by:

$$z^*(x_f) = \begin{cases} 1 - e^{-T^*(x_f)} & \text{if } T^*(x_f) < +\infty \\ 1 & \text{if } T^*(x_f) = +\infty \end{cases} \tag{8.26}$$

This transformation comes immediately when one tries to solve this optimal control problem with the running cost:

$$C(t(x_f)) = \int_0^{t(x_f)} e^{-t} dt = 1 - e^{-t(x_f)}$$

where $t(x_f)$ is a free variable. Minimizing $C(t(x_f))$ amounts to minimizing T^*. Once one finds the optimal solution $z(x_f)$, it is easy to recover T^*, since $T^*(x_f) = -\ln(1 - z(x_f))$.

The concept of solution for the HJB equation needs the concept of viscosity solution. A reminder of the definitions of sub- and supersolutions appears in Appendix C. But the most useful definitions are recalled here. First of all, one needs the notion of lower semicontinuous envelope.

Definition 8.2.1. *Denote $z : X \to [-\infty, +\infty]$, $X \subseteq \mathbb{R}^n$. We call lower semicontinuous envelope of z the function $\underset{\sim}{z}$ defined pointwise by:*

$$\underset{\sim}{z}(x) = \liminf_{y \to x} z(y) = \lim_{r \to 0^+} \inf\{z(y) : y \in X,\ |y - x| \leq r\}$$

One can see easily that $\underset{\sim}{z} = z$ at every point where z is (lower semi-)continuous.
Secondly, one needs the definition of an envelope solution.

Definition 8.2.2. *Consider the Dirichlet problem*

$$\begin{cases} F(x, z(x), \nabla z(x)) = 0 & x \in \kappa \\ z(x) = g(x) & x \in \partial\kappa \end{cases} \tag{8.27}$$

with $\kappa \subseteq \mathbb{R}^n$ open, $F : \kappa \times \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ continuous, and $g : \partial\kappa \to \mathbb{R}$. Denote $\underline{\mathcal{S}} = \{$subsolutions of (8.27)$\}$ and $\overline{\mathcal{S}} = \{$supersolutions of (8.27)$\}$. Let $z : \overline{\kappa} \to \mathbb{R}$ be locally bounded.

1. *z is an envelope viscosity subsolution of (8.27) if there exists $\underline{\mathcal{S}}(z) \subseteq \underline{\mathcal{S}}$, $\underline{\mathcal{S}}(z) \neq \emptyset$, such that:*
$$z(x) = \sup_{w \in \underline{\mathcal{S}}(z)} w(x),\ x \in \overline{\kappa}$$
2. *z is an envelope viscosity supersolution of (8.27) if there exists $\overline{\mathcal{S}}(z) \subseteq \overline{\mathcal{S}}$, $\overline{\mathcal{S}}(z) \neq \emptyset$, such that:*
$$z(x) = \inf_{w \in \overline{\mathcal{S}}(z)} w(x),\ x \in \overline{\kappa}$$
3. *z is an envelope viscosity solution of (8.27) if it is an envelope viscosity sub- and supersolution.*

With these definitions, one can formulate the next Theorem, stating the HJB equation for z^*:

Theorem 8.2.1. *z^* is the envelope viscosity solution of the Dirichlet problem:*

$$\begin{cases} z + H(x, \nabla z) = 1 & \text{in } \mathbb{R}^n \backslash \{x_f\}, \\ z = 0 & \text{on } \{x_f\}, \end{cases} \tag{8.28}$$

where

$$H(x, p) = \sup_{(u,v) \in \Omega} \langle -p, Ax + Bv + Fu \rangle.$$

In case Assumption 8.2.1 is met, then H can be defined as:

$$H(x, p) = \sup_{(u,v) \in \mathcal{E}} \langle -p, Ax + Bv + Fu \rangle, \tag{8.29}$$

where $\mathcal{E}$ has been defined in Proposition 8.2.4.

Proof. By [10, Chapter V.3.2, Theorem 3.7], the lower semicontinuous envelope of z^*, $\underset{\sim}{z^*}$, is the envelope viscosity solution of the Dirichlet problem (8.28). Thanks to Proposition 8.2.3, one can analyse the problem equivalently on Ω or on convΩ. Reasoning on convΩ rather than on Ω, one can prove using [112, Proposition 2.6] that $T^*(\cdot)$ is a lower semicontinuous function; therefore, so is z^*. It proves that $z^* = \underset{\sim}{z^*}$ and therefore, z^* is the envelope viscosity solution of (8.28).
Finally, Proposition 8.2.4 justifies the expression of H in (8.29). □

Remark 8.2.2. *The target $\{x_f\}$ could be changed to any closed nonempty set $\mathcal{T}$ with compact boundary.*

Example 8.2.3. *Example 8.2.1 revisited.*
Let us check that the Kružkov transform of T^* found in (8.23) complies with (8.28). The verification will be carried in the case when $\frac{ax_f+f}{ax_0+f} > 0$, the other cases being treated with the same calculations. In this case, the Kružkov transform of T^* defined in (8.26) amounts to:

$$z^*(x_f) = 1 - \left(\frac{ax+f}{ax_0+f}\right)^{-\frac{1}{a}}.$$

Therefore, one must check that

$$1 - z^*(x_f) = -\left(\frac{ax_f+f}{ax_0+f}\right)^{-\frac{1}{a}} = \sup_{(u,v)\in\widetilde{\Omega}} \left\{(ax_f + bv + fu)\frac{dz^*}{dx}(x_f)\right\} \tag{8.30}$$

where Ω is defined as $\widetilde{\Omega} = \{(u,v) \in \{-1,0,1\} \times \mathbb{R} \mid 0 \leq v \perp dv + eu \geq 0\}$.

As it has been shown in Example 8.2.1, the sup in (8.30) is attained at $u = 1$, $v = 0$. Therefore:

$$\begin{aligned}
\sup_{(u,v)\in\Omega} \left\{(ax_f + bv + fu)\frac{dz^*}{dx}(x_f)\right\} &= (ax_f+f)\left(-\frac{1}{ax_0+f}\left(\frac{ax_f+f}{ax_0+f}\right)^{-\frac{1}{a}-1}\right) \\
&= -\left(\frac{ax_f+f}{ax_0+f}\right)^{-\frac{1}{a}} \\
&= 1 - z^*
\end{aligned}$$

Therefore, using the same definition of H made in (8.29), it is proven that z^* complies with the equation:

$$z^* + H\left(x_f, \frac{dz^*}{dx}\right) = 1,$$

which is the HJB Equation (8.28).

Conclusion

The necessary conditions for optimality exposed in Section 3.2 were extended to the case of minimal time problem. These results were precised for LCS, and some special properties that the optimum possesses in the case of LCS, were also shown. As future work, one could extend the class of LCS complying for the Bounded Slope Condition, and also prove the bang-bang property for a broader class of LCS, as Example 8.2.2 suggests. Finally, most results presented here do not present state in the constraints: this needs to be enhanced.

Conclusion

In this thesis, the optimal control of Linear Complementarity Systems (LCS) has been studied. Two problems were treated: the Linear Quadratic (LQ) optimal control problem with linear complementarity, and the minimal time problem for LCS. These two problems show that different aspects of the study of these systems produce useful results.

For the LQ problem, the main tool we studied are the necessary (and sufficient) conditions of optimality. The study of Mathematical Programming with Equilibrium Constraints (MPEC) hints us that the definition of the multipliers, that are needed for defining properly the stationary conditions, are not expressed in a convenient way, neither for analytical purposes nor for computational ones. The main goal in this thesis has been to re-express these conditions in a more numerically tractable way. Eventually, these conditions are expressed as an LCS (with an additional inequality), for which different results are already available. For instance, since these necessary conditions are often expressed as a Boundary Value Problem (BVP), some results let us construct a way for solving this problem. This, in turn, leads to the definition of two numerical methods for solving the LQ problem, based on the literature. These methods were implemented in a Python code, made in a way such that it is easy to use and to incorporate to a bigger library. The numerical approximations suggest some properties that the optimal solutions may have, and lead to further perspectives.

Concerning the minimal time problem, necessary conditions of optimality were firstly derived for a more general, non-linear system with complementarity constraints, that did not appear in the literature. However, these results, once applied to the case of LCS, seem not as fruitful and instructive as they are for the LQ case (at least if one want to use them for numerical resolution). A more geometrical approach is then used, in order to prove a bang-bang property for LCS in some cases. This, in turn, can be used in order to simplify the search for an optimal solution.

Perspectives

Of course, as it is customary with any topic of research, this study stimulates other questions.

- Concerning the LQ problem, the set of admissible solutions (the state absolutely continuous, the control in L^2) seems definitely too narrow: this is called the Lavrentiev effect. The necessary conditions obtained here should be extended to the case of RCBV solutions, as suggested in Section 1.2.3. Some results in [9] may help building these.

- Even in the case of absolutely continuous solutions, the optimality conditions for the LQ problem, written as a BVP where the underlying system is an LCS, is not fully analyzed. For instance, the Theorem 1.2.5 concerning the dependence of the BVP to the initial condition still needs to be fully adapted to this problem.

- The numerical methods work well for systems of low dimensional complementarity, but it seems harder to make them converge when the dimension is high. A work for enhancing the used algorithms is needed.

- Different assumptions were made across this manuscript, covering different aspects, like the existence of an optimal solution, or properties that the constraints must meet (like E invertible for the LQ problem, D **P**-matrix for the minimal time problem, or the somewhat restricting Bounded Slope Condition). Solving these questions or dropping some of these assumptions, will allow one to apply these results to a bigger range of systems, some of them already appearing in some engineering problems. A first approach may resemble to the one found in [13]; it can most certainly be extended to complementarity constraints.

Appendix

Appendix A

Non-smooth Analysis

Different tools of non-smooth analysis were used in this manuscript. First, let us define the different notions of normal cones.

Definition A.0.1. *Let $\Omega \subseteq \mathbb{R}^n$, and $x \in cl\ \Omega$.*

- *The proximal normal cone to Ω at x is defined as:*

$$\mathcal{N}_\Omega^P(x) = \{v \in \mathbb{R}^n \ : \ \exists \sigma > 0;\ \langle v, y - x\rangle \leq \sigma \|y - x\|^2\ \forall y \in \Omega\}.$$

- *The limiting normal cone to Ω at x is defined as:*

$$\mathcal{N}_\Omega^L(x) = \{v \in \mathbb{R}^n \ : \ \exists (x_k, v_k) \to (x, v);\ v^k \in \mathcal{N}_\Omega^P(x^k)\ \forall k\}.$$

- *The Clarke normal cone to Ω at x is defined as $\mathcal{N}_\Omega^C(x) = cl\ conv\ \mathcal{N}_\Omega^L(x)$.*

One can obtain the following inclusions:

$$\mathcal{N}_\Omega^P(x) \subseteq \mathcal{N}_\Omega^L(x) \subseteq \mathcal{N}_\Omega^C(x)\ \forall x \in \mathrm{cl}\Omega$$

Note that when Ω is convex, all these normal cones coincide with the normal cone of convex analysis (see [91]).

Aside normal cones, one can also define subdifferential of various functions.

Definition A.0.2. *Let $\phi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ a proper lower semicontinuous function, and x a point such that $\phi(x)$ is finite. The limiting subdifferential of ϕ at x is defined as:*

$$\partial^L \phi(x) = \left\{ v \in \mathbb{R}^n \ : \ \exists (v^k, x^k) \to (x, v);\ \lim_{y \to x^k} \frac{\phi(y) - \phi(x^k) - \langle v^k, y - x^k\rangle}{\|y - x^k\|} \geq 0,\ \forall k \right\}$$

If ϕ is Lipschitz near x, then the Clarke subdifferential of ϕ at x can be defined as $\partial^C \phi(x) = cl\ conv\ \partial^L \phi(x)$.

Appendix B

Krein-Milman Theorem

Since the Krein-Milman Theorem is heavily used in Chapter 8, it is worth recalling its statement. Let us start with a definition.

Definition B.0.1. *Let C be a convex compact subset of a Hausdorff locally convex set. Let $c \in C$. The point c is called an extremal point of C if $C \backslash \{c\}$ is still convex. Equivalently, c is an extreme point of C if the following implication holds:*

$$c_1, c_2 \in C,\ c = \frac{1}{2}(c_1 + c_2) \implies c = c_1 = c_2$$

The set of extreme points of C is denoted by $\mathrm{Ext}(C)$.

Theorem B.0.1 (Krein-Milman)**.** *Let C be a convex compact subset of a Hausdorff locally convex set. Then*

$$C = \text{cl conv } (\mathrm{Ext}(C))$$

Appendix C

Viscosity solutions

In order to understand some results concerning the HJB equation in Chapter 8, one needs to know some definitions related to the concept of viscosity solutions. The definitions given here, extracted from [10], are only the ones useful for this manuscript. In particular, the definitions given here are the ones useful to handle the concept of discontinuous viscosity solutions. The interested reader can find broader results in [10] and the references therein.

Let us first define the notion of subsolution and supersolution of a first order equation

$$F(x, u, \nabla u) = 0 \text{ in } \Omega, \tag{C.1}$$

with $\Omega \subseteq \mathbb{R}^n$ and $F : \Omega \times \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ continuous. For this, let us fix some notations: for $E \subseteq \mathbb{R}^n$, denote

$$\mathrm{USC}(E) = \{u : E \to \mathbb{R} \text{ upper semicontinuous}\},$$
$$\mathrm{LSC}(E) = \{u : E \to \mathbb{R} \text{ lower semicontinuous}\}.$$

Definition C.0.1. *A function $u \in USC(\Omega)$ (resp. $LSC(\Omega)$) is a viscosity subsolution (resp. supersolution) of (C.1) if, for any $\phi \in \mathscr{C}^1(\Omega)$ and $x \in \Omega$ such that $u - \phi$ has a local maximum (resp. minimum) at x,*

$$F(x, u(x), \nabla\phi(x)) \leq 0 \text{ (resp. } \geq 0).$$

Bibliography

[1] V. Acary, O. Bonnefon, and B. Brogliato. Time-stepping numerical simulation of switched circuits within the nonsmooth dynamical systems approach. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 29(7):1042–1055, 2010.

[2] V. Acary, O. Bonnefon, and B. Brogliato. *Nonsmooth Modeling and Simulation for Switched Circuits*, volume 69 of *Lecture Notes in Electrical Engineering*. Springer Science & Business Media, 2011.

[3] V. Acary, B. Brogliato, and D. Goeleven. Higher order Moreau's sweeping process: mathematical formulation and numerical simulation. *Mathematical Programming*, 113(1):133–217, 2008.

[4] L. Adam and J.V. Outrata. On optimal control of a sweeping process coupled with an ordinary differential equation. *Discrete & Continuous Dynamical Systems-Series B*, 19(9), 2014.

[5] S. Adly, T. Haddad, and L. Thibault. Convex sweeping process in the framework of measure differential inclusions and evolution variational inequalities. *Mathematical Programming*, 148(1-2):5–47, 2014.

[6] S. Adly, F. Nacry, and L. Thibault. Discontinuous sweeping process with prox-regular sets. *ESAIM: Control, Optimisation and Calculus of Variations*, 23(4):1293–1329, 2017.

[7] J. Andersson. *A General-Purpose Software Framework for Dynamic Optimization*. PhD thesis, Arenberg Doctoral School, KU Leuven, Department of Electrical Engineering (ESAT/SCD) and Optimization in Engineering Center, Kasteelpark Arenberg 10, 3001-Heverlee, Belgium, October 2013.

[8] P. Antsaklis and A. Nerode. Hybrid control systems: An introductory discussion to the special issue. *IEEE Transactions on Automatic Control*, 43(4):457–460, 1998.

[9] A.V. Arutyunov, D. Yu Karamzin, and F. L. Pereira. On constrained impulsive control problems. *Journal of Mathematical sciences*, 165(6):654–688, 2010.

[10] M. Bardi and I. Capuzzo-Dolcetta. *Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations*. Springer Science & Business Media, 2008.

[11] S.C. Bengea and R.A. DeCarlo. Optimal control of switching systems. *Automatica*, 41(1):11–27, 2005.

[12] J. Betts. *Practical methods for optimal control and estimation using nonlinear programming*, volume 19. Siam, 2010.

[13] A. Boccia, M.D.R. De Pinho, and R. Vinter. Optimal control problems with mixed and pure state constraints. *SIAM Journal on Control and Optimization*, 54(6):3061–3083, 2016.

[14] J. F. Bonnans and A. Hermant. Second-order analysis for optimal control problems with pure state constraints and mixed control-state constraints. In *Annales de l'Institut Henri Poincaré (C) Non Linear Analysis*, volume 26, pages 561–598. Elsevier, 2009.

[15] B. Bonnard, L. Faubourg, G. Launay, and E. Trélat. Optimal control with state constraints and the space shuttle re-entry problem. *Journal of Dynamical and Control Systems*, 9(2):155–199, 2003.

[16] B. Bonnard, L. Faubourg, and E. Trélat. Optimal control of the atmospheric arc of a space shuttle and numerical simulations with multiple-shooting method. *Mathematical Models and Methods in Applied Sciences*, 15(01):109–140, 2005.

[17] M. Branicky, V. Borkar, and S. K. Mitter. A unified framework for hybrid control: Model and optimal control theory. *IEEE Transactions on Automatic Control*, 43(1):31–45, 1998.

[18] L.M. Briceno-Arias, N. D. Hoang, and J. Peypouquet. Existence, stability and optimality for optimal control problems governed by maximal monotone operators. *Journal of Differential Equations*, 260(1):733–757, 2016.

[19] B. Brogliato. Some perspectives on the analysis and control of complementarity systems. *IEEE Transactions on Automatic Control*, 48(6):918–935, 2003.

[20] B. Brogliato. Absolute stability and the Lagrange–Dirichlet theorem with monotone multivalued mappings. *Systems & Control Letters*, 51(5):343–353, 2004.

[21] B Brogliato. *Nonsmooth Mechanics*. Springer, 2016.

[22] B. Brogliato and L. Thibault. Existence and uniqueness of solutions for non-autonomous complementarity dynamical systems. *Journal of Convex Analysis*, 17(3):961–990, 2010.

[23] M. Brokate and P. Krejčí. Optimal control of ode systems involving a rate independent variational inequality. *Discrete & Continuous Dynamical Systems-Series B*, 18(2), 2013.

[24] A.E. Bryson. *Applied Optimal Control: Optimization, Estimation and Control*. Halsted Press book'. Taylor & Francis, 1975.

[25] C. Büskens and H. Maurer. SQP-methods for solving optimal control problems with control and state constraints: adjoint variables, sensitivity analysis and real-time control. *Journal of computational and applied mathematics*, 120(1-2):85–108, 2000.

[26] K. Camlibel, L. Iannelli, A. Tanwani, and S. Trenn. Differential-algebraic inclusions with maximal monotone operators. In *Decision and Control (CDC), 2016 IEEE 55th Conference on*, pages 610–615. IEEE, 2016.

[27] M.K. Camlibel, W.P.M.H. Heemels, and J.M.H. Schumacher. Consistency of a time-stepping method for a class of piecewise-linear networks. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 49(3):349–357, 2002.

[28] T. H. Cao and B. S. Mordukhovich. Optimal control of a perturbed sweeping process via discrete approximations. *Disc. Cont. Dyn. Syst. Ser. B*, 21(10):3331–3358, 2015.

[29] C. G Cassandras, D.L. Pepyne, and Y. Wardi. Optimal control of a class of hybrid systems. *Automatic Control, IEEE Transactions on*, 46(3):398–415, 2001.

[30] C. Castaing and M.D.P. Monteiro Marques. Bv periodic solutions of an evolution problem associated with continuous moving convex sets. *Set-Valued Analysis*, 3(4):381–399, 1995.

[31] M.K. Çamlibel. Popov–Belevitch–Hautus type controllability tests for linear complementarity systems. *Systems & Control Letters*, 56(5):381–387, 2007.

[32] M.K. Çamlibel, W.P.M.H. Heemels, and J.M. Schumacher. On linear passive complementarity systems. *European Journal of Control*, 8(3):220 – 237, 2002.

[33] M.K. Çamlibel and J.M. Schumacher. On the Zeno behavior of linear complementarity systems. In *Decision and Control, 2001. Proceedings of the 40th IEEE Conference on*, volume 1, pages 346–351. IEEE, 2001.

[34] L. Cesari. *Optimization-Theory and Applications: Problems with Ordinary Differential Equations*, volume 17. Springer Science & Business Media, 2012.

[35] M. Chyba, E. Hairer, and G. Vilmart. The role of symplectic integrators in optimal control. *Optimal Control Applications and Methods*, 30(4):367–382, 2009.

[36] F. Clarke. *Optimization and Nonsmooth Analysis*, volume 5. SIAM, 1990.

[37] F. Clarke. *Necessary Conditions in Dynamic Optimization*. Number 816. American Mathematical Soc., 2005.

[38] F. Clarke and M.D.R. De Pinho. Optimal control problems with mixed constraints. *SIAM Journal on Control and Optimization*, 48(7):4500–4524, 2010.

[39] F. H. Clarke. A variational proof of Aumann's theorem. *Applied Mathematics and Optimization*, 7(1):373–378, 1981.

[40] F.H. Clarke. The maximum principle under minimal hypotheses. *SIAM Journal on Control and Optimization*, 14(6):1078–1091, 1976.

[41] G. Colombo and V.V. Goncharov. The sweeping processes without convexity. *Set-Valued Analysis*, 7(4):357–374, 1999.

[42] G. Colombo, R. Henrion, N.D. Hoang, and B.S. Mordukhovich. Discrete approximations of a controlled sweeping process. *Set-Valued and Variational Analysis*, 23(1):69–86, 2014.

[43] G. Colombo, R. Henrion, D. H. Nguyen, and B. S. Mordukhovich. Optimal control of the sweeping process over polyhedral controlled sets. *Journal of Differential Equations*, 260(4):3397–3447, 2016.

[44] G. Colombo and M. Palladino. The minimum time function for the controlled Moreau's sweeping process. *SIAM Journal on Control and Optimization*, 54(4):2036–2062, 2016.

[45] R. Cottle, J.-S. Pang, and R. Stone. *The Linear Complementarity Problem*. SIAM, 2009.

[46] M. d. R. de Pinho. Mixed constrained control problems. *Journal of Mathematical Analysis and Applications*, 278(2):293–307, 2003.

[47] M.D.R. De Pinho and J. Rosenblueth. Necessary conditions for constrained problems under Mangasarian–Fromowitz conditions. *SIAM Journal on Control and Optimization*, 47(1):535–552, 2008.

[48] M. Diehl, H. G. Bock, H. Diedam, and P-B Wieber. Fast direct multiple shooting algorithms for optimal robot control. In *Fast motions in Biomechanics and Robotics*, pages 65–93. Springer, 2006.

[49] A. V. Dmitruk and A. M. Kaganovich. The hybrid maximum principle is a consequence of Pontryagin maximum principle. *Systems & Control Letters*, 57(11):964–970, 2008.

[50] A.V. Dmitruk. Maximum principle for the general optimal control problem with phase and regular mixed constraints. *Computational Mathematics and Modeling*, 4(4):364–377, 1993.

[51] A. Dontchev and W. Hager. The Euler approximation in state constrained optimal control. *Mathematics of Computation*, 70(233):173–203, 2001.

[52] A. Dontchev, W. Hager, and V. Veliov. Second-order Runge–Kutta approximations in control constrained optimal control. *SIAM Journal on Numerical Analysis*, 38(1):202–226, 2000.

[53] G. Elnagar and M. Kazemi. Pseudospectral Chebyshev optimal control of constrained nonlinear dynamical systems. *Computational Optimization and Applications*, 11(2):195–217, 1998.

[54] F. Facchinei and J.-S. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Science & Business Media, 2007.

[55] A. Giua, C. Seatzu, and C. Van der Mee. Optimal control of switched autonomous linear systems. In *Decision and Control, 2001. Proceedings of the 40th IEEE Conference on*, volume 3, pages 2472–2477 vol.3, 2001.

[56] D. Goeleven and B. Brogliato. Stability and instability matrices for linear evolution variational inequalities. *IEEE Transactions on Automatic Control*, 49(4):521–534, 2004.

[57] Q. Gong, M. Ross, W. Kang, and F. Fahroo. Connections between the covector mapping theorem and convergence of pseudospectral methods for optimal control. *Computational Optimization and Applications*, 41(3):307–335, 2008.

[58] L. Guo, G.-H. Lin, and J. J. Ye. Solving mathematical programs with equilibrium constraints. *Journal of Optimization Theory and Applications*, 166(1):234–256, 2015.

[59] L. Guo and J. J. Ye. Necessary optimality conditions for optimal control problems with equilibrium constraints. *SIAM Journal on Control and Optimization*, 54(5):2710–2733, 2016.

[60] W. Hager. Runge–Kutta methods in optimal control and the transformed adjoint system. *Numerische Mathematik*, 87(2):247–282, 2000.

[61] L. Han, A. Tiwari, M.K. Çamlibel, and J.-S. Pang. Convergence of time-stepping schemes for passive and extended linear complementarity systems. *SIAM Journal on Numerical Analysis*, 47(5):3768–3796, 2009.

[62] M. Hautus and L. Silverman. System structure and singular control. *Linear Algebra and its Applications*, 50:369–402, 1983.

[63] W.P.M.H. Heemels, J. M. Schumacher, and S. Weiland. Linear Complementarity Systems. *SIAM Journal on Applied Mathematics*, 60(4):1234–1269, 2000.

[64] W.P.M.H. Heemels, J.M. Schumacher, and S. Weiland. Well-posedness of linear complementarity systems. In *Decision and Control, 1999. Proceedings of the 38th IEEE Conference on*, volume 3, pages 3037–3042. IEEE, 1999.

[65] O. Janin and C.-H. Lamarque. Comparison of several numerical methods for mechanical systems with impacts. *International Journal for Numerical Methods in Engineering*, 51(9):1101–1132, 2001.

[66] C. Kanzow and A. Schwartz. Mathematical programs with equilibrium constraints: enhanced Fritz John-conditions, new constraint qualifications, and improved exact penalty results. *SIAM Journal on Optimization*, 20(5):2730–2753, 2010.

[67] C. Kanzow and A. Schwartz. A new regularization method for mathematical programs with complementarity constraints with strong convergence properties. *SIAM Journal on Optimization*, 23(2):770–798, 2013.

[68] M. Kunze and M.D.P. Monteiro Marques. Yosida–Moreau regularization of sweeping processes with unbounded variation. *Journal of Differential Equations*, 130(2):292–306, 1996.

[69] Han L., Camlibel M.K., Pang J.-S., and Heemels W.P.M.H. A unified numerical scheme for linear-quadratic optimal control problems with joint control and state constraints. *Optimization Methods and Software*, 27(4-5):761–799, 2012.

[70] J. Lasserre, D. Henrion, C. Prieur, and E. Trélat. Nonlinear optimal control via occupation measures and lmi-relaxations. *SIAM Journal on Control and Optimization*, 47(4):1643–1666, 2008.

[71] D. Leenaerts. On linear dynamic complementarity systems. *IEEE Transactions on Circuits and Systems I*, 46(8):1022–1026, 1999.

[72] S. Leyffer. MacMPEC: AMPL collection of MPECs, https://wiki.mcs.anl.gov/leyffer/index.php/macmpec, August 2015.

[73] S. Leyffer, G. López-Calva, and J. Nocedal. Interior methods for mathematical programs with complementarity constraints. *SIAM Journal on Optimization*, 17(1):52–77, 2006.

[74] J. Lohéac, E. Trélat, and E. Zuazua. Minimal controllability time for finite-dimensional control systems under state constraints. February 2018. working paper or preprint.

[75] Z.-Q. Luo, J.-S. Pang, and D. Ralph. *Mathematical Programs with Equilibrium Constraints*. Cambridge University Press, 1996.

[76] K. Makowski and L. Neustadt. Optimal control problems with mixed control-phase variable equality and inequality constraints. *SIAM Journal on Control*, 12(2):184–228, 1974.

[77] B. Maury and J. Venel. A mathematical framework for a crowd motion model. *Comptes Rendus Mathématique*, 346(23-24):1245–1250, 2008.

[78] Michael McAsey, Libin Mou, and Weimin Han. Convergence of the forward-backward sweep method in optimal control. *Computational Optimization and Applications*, 53(1):207–226, 2012.

[79] B. Mordukhovich. *Variational Analysis and Generalized Differentiation II: Applications*, volume 330. Springer Science & Business Media, 2006.

[80] B. Mordukhovich and L. Wang. Optimal control of differential-algebraic inclusions. In *Optimal Control, Stabilization and Nonsmooth Analysis*, pages 73–83. Springer, 2004.

[81] J.J. Moreau. Evolution problem associated with a moving convex set in a Hilbert space. *Journal of Differential Equations*, 26(3):347–374, 1977.

[82] J. Outrata. Optimality conditions for a class of mathematical programs with equilibrium constraints. *Mathematics of Operations Research*, 24(3):627–644, 1999.

[83] J-S. Pang and M. Fukushima. Complementarity constraint qualifications and simplified b-stationarity conditions for mathematical programs with equilibrium constraints. *Computational Optimization and Applications*, 13(1-3):111–136, 1999.

[84] J.-S. Pang and D. E. Stewart. Differential variational inequalities. *Mathematical Programming*, 113(2):345–424, 2008.

[85] J-S. Pang and D. E. Stewart. Solution dependence on initial conditions in differential variational inequalities. *Mathematical Programming*, 116(1):429–460, 2009.

[86] B. Passenberg, P.E. Caines, M. Leibold, O. Stursberg, and M. Buss. Optimal control for hybrid systems with partitioned state space. *IEEE Transactions on Automatic Control*, 58(8):2131–2136, 2013.

[87] B. Piccoli. Hybrid systems and optimal control. volume 1, pages 13–18. IEEE, 1998.

[88] L. Pontryagin, V.G. Boltyanskii, R.V. Gamkrelidze, and E.F. Mishchenko. *The Mathematical Theory of Optimal Processes*. 1962.

[89] R. Pytlak. *Numerical Methods for Optimal Control Problems with State Constraints*. Springer, 2006.

[90] R. Rockafellar and R.J-B Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.

[91] R. T. Rockafellar. *Convex Analysis*. Princeton university press, 2015.

[92] M. Ross. A roadmap for optimal control: the right way to commute. *Annals of the New York Academy of Sciences*, 1065(1):210–231, 2005.

[93] M. Ross and F. Fahroo. Legendre pseudospectral approximations of optimal control problems. In *New Trends in Nonlinear Dynamics and Control and their Applications*, pages 327–342. Springer, 2003.

[94] J.M. Sanz-Serna. Symplectic Runge–Kutta schemes for adjoint equations, automatic differentiation, optimal control, and more. *SIAM Review*, 58(1):3–33, 2016.

[95] H. Scheel and S. Scholtes. Mathematical programs with complementarity constraints: Stationarity, optimality, and sensitivity. *Mathematics of Operations Research*, 25(1):1–22, 2000.

[96] M.S. Shaikh and P.E. Caines. On the hybrid optimal control problem: theory and algorithms. *IEEE Transactions on Automatic Control*, 52(9):1587–1603, 2007.

[97] J. Shen and J-S. Pang. Linear complementarity systems: Zeno states. *SIAM Journal on Control and Optimization*, 44(3):1040–1066, 2005.

[98] G. Smirnov. *Introduction to the Theory of Differential Inclusions*, volume 41. American Mathematical Soc., 2002.

[99] D. Stewart. A high accuracy method for solving odes with discontinuous right-hand side. *Numerische Mathematik*, 58(1):299–328, 1990.

[100] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*, volume 12. Springer Science & Business Media, 2013.

[101] H. J. Sussmann. A maximum principle for hybrid optimal control problems. In *Proceedings of the 38th IEEE Conference on Decision and Control (Cat. No.99CH36304)*, volume 1, pages 425–430 vol.1, Dec 1999.

[102] H. J. Sussmann. A nonsmooth hybrid maximum principle. In *Stability and Stabilization of Nonlinear Systems*, pages 325–354. Springer, 1999.

[103] E. Trélat. *Contrôle optimal : Théorie & Applications*. Vuibert, 2005. ISBN 2 7117 7175 X.

[104] E. Trélat. Optimal control and applications to aerospace: some results and challenges. *Journal of Optimization Theory and Applications*, 154(3):713–758, 2012.

[105] R. Vasudevan, H. Gonzalez, R. Bajcsy, and S. Shankar Sastry. Consistent approximations for the optimal control of constrained switched systems—part 1: A conceptual algorithm. *SIAM Journal on Control and Optimization*, 51(6):4463–4483, 2013.

[106] R. Vasudevan, H. Gonzalez, R. Bajcsy, and S. Shankar Sastry. Consistent approximations for the optimal control of constrained switched systems—part 2: An implementable algorithm. *SIAM Journal on Control and Optimization*, 51(6):4484–4503, 2013.

[107] A. Vieira, B. Brogliato, and C. Prieur. Preliminary results on the optimal control of linear complementarity systems. *IFAC-PapersOnLine*, 50(1):2977 – 2982, 2017.

[108] A. Vieira, B. Brogliato, and C. Prieur. Quadratic Optimal Control of Linear Complementarity Systems: First order necessary conditions and numerical analysis. https://hal.inria.fr/hal-01690400, January 2018.

[109] R. Vinter. *Optimal Control*. Springer Science & Business Media, 2010.

[110] G. Wachsmuth. On LICQ and the uniqueness of Lagrange multipliers. *Operations Research Letters*, 41(1):78–80, 2013.

[111] A. Wächter and L. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical programming*, 106(1):25–57, 2006.

[112] P.R. Wolenski and Y. Zhuang. Proximal analysis and the minimal time function. *SIAM Journal on Control and Optimization*, 36(3):1048–1072, 1998.

[113] J.J. Ye. Optimality conditions for optimization problems with complementarity constraints. *SIAM Journal on Optimization*, 9(2):374–387, 1999.

[114] J.J. Ye. Necessary and sufficient optimality conditions for mathematical programs with equilibrium constraints. *Journal of Mathematical Analysis and Applications*, 307(1):350–369, 2005.

[115] J.J. Ye, D.L. Zhu, and Q.J. Zhu. Exact penalization and necessary optimality conditions for generalized bilevel programming problems. *SIAM Journal on Optimization*, 7(2):481–507, 1997.