

# Contrôle optimal et robuste de l'attitude d'un lanceur. Aspects théoriques et numériques

Olivier Antoine

▶ **To cite this version:**

Olivier Antoine. Contrôle optimal et robuste de l'attitude d'un lanceur. Aspects théoriques et numériques. Mathématiques générales [math.GM]. Sorbonne Université, 2018. Français. NNT : 2018SORUS196 . tel-02447692

**HAL Id: tel-02447692**
**https://theses.hal.science/tel-02447692**

Submitted on 21 Jan 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université Pierre et Marie Curie
CNES - Direction des Lanceurs

École doctorale **ED 386**

Unité de recherche **Laboratoire Jacques Louis Lions**

Thèse présentée par **Antoine Olivier**

Soutenue le **4 octobre 2018**

En vue de l'obtention du grade de docteur de l'Université Pierre et Marie Curie

Discipline **Mathématiques Appliquées**

Spécialité **Contrôle Optimal**

# Contrôle optimal et robuste de l'attitude d'un lanceur

## Aspects Théoriques et Numériques

**Thèse dirigée par**

| | |
|---|---|
| Emmanuel Trélat | directeur |
| Thomas Haberkorn | co-directeur |
| Éric Bourgeois | co-directeur |
| David-Alexis Handschuh | co-directeur |

**Composition du jury**

| | | |
|---|---|---|
| *Rapporteur* | Jean-Baptiste Caillau | |
| *Examinateurs* | Pascal Frey | président du jury |
| | Hasnaa Zidani | |
| *Directeurs de thèse* | Emmanuel Trélat | |
| | Thomas Haberkorn | |
| | Éric Bourgeois | |
| | David-Alexis Handschuh | |

Université Pierre et Marie Curie
CNES - Direction des Lanceurs

Doctoral School **ED 386**

University Department **Laboratoire Jacques Louis Lions**

Thesis defended by **Antoine Olivier**

Defended on **4th October, 2018**

In order to become Doctor from Université Pierre et Marie Curie

Academic Field **Applied Mathematics**

Speciality **Optimal Control**

# Optimal and robust attitude control of a launcher

## Theoretical and numerical aspects

**Thesis supervised by**

| | |
|---|---|
| Emmanuel Trélat | Supervisor |
| Thomas Haberkorn | Co-Supervisor |
| Éric Bourgeois | Co-Supervisor |
| David-Alexis Handschuh | Co-Supervisor |

**Committee members**

| | | |
|---|---|---|
| *Referee* | Jean-Baptiste Caillau | |
| *Examiners* | Pascal Frey | Committee President |
| | Hasnaa Zidani | |
| *Supervisors* | Emmanuel Trélat | |
| | Thomas Haberkorn | |
| | Éric Bourgeois | |
| | David-Alexis Handschuh | |

**Mots clés :** contrôle optimal, contrôle d'attitude, phase balistique, méthode de continuation, méthodes directes, méthodes indirectes, contrôle robuste, contrainte intermédiaire

**Keywords:** optimal control, attitude control, ballistic phase, continuation method, direct methods, indirect methods, robust control, intermediate constraint

Cette thèse a été préparée au laboratoire suivant, avec le soutien financier de la FSMP.

**Laboratoire Jacques Louis Lions**
4 Place Jussieu
75005, PARIS
France



**Fondation Science Mathématiques de Paris**
11 rue Pierre et Marie Curie
75231F Paris Cedex 09
France



# Remerciements

## Contrôle optimal et robuste de l'attitude d'un lanceur
### Aspects Théoriques et Numériques

**Résumé**

L'objectif premier de cette thèse est d'étudier certains aspects du contrôle d'attitude d'un corps rigide, afin d'optimiser la trajectoire d'un lanceur au cours de sa phase balistique. Nous y développons un cadre mathématique permettant de formuler ce problème comme un problème de contrôle optimal avec des contraintes intermédiaires sur l'état. En parallèle de l'étude théorique de ce problème, nous avons mené l'implémentation d'un logiciel d'optimisation basé sur la combinaison d'une méthode directe et d'un algorithme de point intérieur, permettant à l'utilisateur de traiter une phase balistique quelconque. Nous entendons par là qu'il est possible de spécifier un nombre quelconque de contraintes intermédiaires, correspondant à un nombre quelconque de largages de charges utiles.
En outre, nous avons appliqué les méthodes dites indirectes, exploitant le principe du maximum de Pontryagin, à la résolution de ce problème de contrôle optimal. On cherche dans ce travail à trouver des trajectoires optimales du point de vue de la consommation en ergols, ce qui correspond à un coût $L^1$. Réputé difficile numériquement, ce critère peut être atteint grâce à une méthode de continuation, en se servant d'un coût $L^2$ comme intermédiaire de calcul et en déformant progressivement ce problème $L^2$. Nous verrons également d'autres exemples d'application des méthodes de continuation.
Enfin, nous présenterons également un algorithme de contrôle robuste, permettant de rejoindre un état cible à partir d'un état perturbé, en suivant une trajectoire de référence tout en conservant la structure bang-bang des contrôles. La robustesse d'un contrôle peut également être améliorée par l'ajout de variations aiguilles, et un critère qualifiant la robustesse d'une trajectoire à partir des valeurs singulières d'une certaine application entrée-sortie est déduit.

**Mots clés :** contrôle optimal, contrôle d'attitude, phase balistique, méthode de continuation, méthodes directes, méthodes indirectes, contrôle robuste, contrainte intermédiaire

---

## Optimal and robust attitude control of a launcher
### Theoretical and numerical aspects

**Abstract**

The first objective of this work is to study some aspects of the attitude control problem of a rigid body, in order to optimize the trajectory of a launcher during a ballistic flight. We state this problem in a general mathematical setting, as an optimal control problem with intermediate constraints on the state. Meanwhile, we also implement an optimization software that relies on the combination of a direct method and of an interior-point algorithm to optimize any given ballistic flight, with any number of intermediate constraints, corresponding to any number of satellite separations.
Besides, we applied the so-called indirect methods, exploiting Pontryagin maximum principle, to the resolution of this optimal control problem. In this work, optimal trajectories with respect to the consumption are looked after, which corresponds to a $L^1$ cost. Known to be numerically challenging, this criterion can be reached by performing a continuation procedure, starting from a $L^2$ cost, for which it is easier to provide a good initialization of the underlying optimization algorithm. We shall also study other examples of applications for continuation procedures.
Eventually, we will present a robust control algorithm, allowing to reach a target point from a perturbed initial point, following a nominal trajectory while preserving its bang-bang structure. The robustness of a control can be improved introducing needle-like variations, and a criterion to measure the robustness of a trajectory is designed, involving the singular value decomposition of some end-point mapping.

**Keywords:** optimal control, attitude control, ballistic phase, continuation method, direct methods, indirect methods, robust control, intermediate constraint

---

**Laboratoire Jacques Louis Lions**
4 Place Jussieu – 75005, PARIS – France

# Table des matières

**Remerciements** ix

**Résumé** xi

**Table des matières** xiii

**Table des figures** xvii

**Introduction générale au problème de contrôle d'attitude** 1
Positionnement du problème . . . 1
Géométrie générale d'un lanceur Ariane 5 . . . 1
Phase balistique . . . 5
Modélisation du problème de contrôle d'attitude . . . 7
Évolution de l'orientation d'un lanceur. . . . 7
Évolution de la vitesse angulaire . . . 7
Contrôle optimal. . . . 11
Structure du manuscrit et description des contributions . . . 12

**1 Controllability of the attitude for a rigid spacecraft** 17
1.1 Poisson stability of a vector field . . . 19
1.2 Lie algebra spanned by vector fields and controllability . . . 21
1.2.1 Lie bracket and Lie algebra . . . 22
1.2.2 Results in the non-symmetric case . . . 24
1.3 Controllability of the attitude of a rigid body . . . 26
1.3.1 Dimension of Lie$(Q, \vec{b})$ . . . 26
1.3.2 Dimension de Lie$(f_0, f_1)$ . . . 28
1.3.3 Controllability condition . . . 29
1.4 Conclusion of the chapter . . . 30

**2 Optimal control in finite dimension** 31
2.1 General setting . . . 32
2.2 Pontryagin Maximum Principle . . . 33
2.3 Numerical methods in optimal control . . . 36
2.3.1 Direct methods . . . 36
2.3.2 Indirect methods . . . 37
2.3.3 Comparison between the methods . . . 38
2.4 Application to the attitude control problem for a rigid body . . . 39
2.4.1 With an indirect method . . . 39

2.4.2 With a direct method . . . 44
2.5 Conclusion of the chapter . . . 49

**3 Optimal control with intermediate constraints 51**
3.1 Introduction of the chapter . . . 52
3.2 Optimal control formulation . . . 54
3.2.1 Hybrid maximum principle. . . . 54
3.2.2 PMP for $(\mathcal{P})_{via,s}$ and $(\mathcal{P})_{pen,\varepsilon}$ . . . 56
3.2.3 Shooting functions for $(\mathcal{P})_{via,s}$ and $(\mathcal{P})_{pen,\varepsilon}$ . . . 59
3.3 Application to the attitude control of a rigid body . . . 60
3.3.1 The attitude control problem . . . 60
3.3.2 Continuation procedure . . . 61
3.4 Numerical results . . . 64
3.5 Conclusion of this chapter . . . 66

**4 Redundancy implies robustness for bang-bang control strategies 67**
4.1 Introduction of the chapter . . . 68
4.1.1 Overview of the method . . . 68
4.1.2 State of the art on robust control design . . . 71
4.1.3 Structure of this chapter . . . 73
4.2 Tracking algorithm . . . 73
4.2.1 Reduced end-point mapping . . . 73
4.2.2 Absorbing perturbations . . . 75
4.3 Promoting robustness . . . 80
4.3.1 An auxiliary optimization problem . . . 81
4.3.2 Redundancy creates robustness . . . 81
4.4 Numerical results . . . 84
4.4.1 Computing the nominal trajectory . . . 84
4.4.2 Robustifying the nominal trajectory . . . 85
4.5 Proof of Proposition 4.1 . . . 87
4.6 Conclusion of the chapter and perspectives . . . 92

**5 Combination of direct methods and homotopy 93**
5.1 Introduction of the chapter . . . 94
5.2 Modeling Approach and Optimal Control Problem . . . 96
5.2.1 Modeling Approach . . . 96
5.2.2 The Optimal Control Problem . . . 97
5.2.3 Previous Results for $\lambda_0 = 0$ . . . 98
5.3 Resolution of a Simplified Model . . . 99
5.3.1 Simplified Model for one Population with no State Constraints . . . 99
5.3.2 A Maximum Principle in Infinite Dimension . . . 99
5.4 The Continuation Procedure . . . 102
5.4.1 General Principle . . . 102
5.4.2 From $(\mathbf{OCPPDE_1})$ to $(\mathbf{OCPPDE_0})$ . . . 103
5.4.3 General Algorithm . . . 104
5.5 Numerical Results . . . 105
5.6 Perspectives . . . 112
5.7 Conclusion of the chapter . . . 114

**Conclusion and perspectives 115**

**A A software to solve a complete ballistic phase** **117**

**B Liouville's theorem** **123**

**C Linear Algebra** **127**
C.1 Singular value decomposition and pseudoinverse . . . . . 127
C.2 A least-squares problem . . . . . 128
C.3 Condition number of a matrix . . . . . 129

**Bibliographie** **131**

# Table des figures


1 Vue éclatée d'un lanceur Ariane 5 (*Source : Manuel utilisateur Ariane 5*). . . . . 3
2 Représentation de la partie basse d'Ariane 5, avec les deux EAP et l'EPC. *Source : CNES*. . . . . 4
3 Représentation de la partie haute d'Ariane 5. *Source : CNES*. . . . . 4
4 Chronologie d'un vol Ariane 5 ECA. *Source : CNES*. . . . . 5
5 Schéma du Système de Contrôle d'Attitude . . . . . 6
6 Séquence des trois angles d'attitude. . . . . 8

1.1 Scheme of the SCA. Some pairs of thrusters create opposite torques. . . . . 18
1.2 Crochet de Lie $[X, Y]$ de deux champs de vecteurs $X$ et $Y$. . . . . 22

2.1 General principle of the continuation procedure. The resolution of $(\mathbf{OCP})_\alpha$ is used to initialize the resolution for $\alpha - \Delta\alpha$. . . . . 43
2.2 Principle of the continuation procdeure with linear prediction. The solutions of the problem $(\mathbf{OCP})_\alpha$ and $(\mathbf{OCP})_{\alpha+\Delta\alpha_1}$ are used to initialize the shooting problem for $\alpha - \Delta\alpha_2$, doing an affine extrapolation in order to get an approximation for $Z_{\alpha-\Delta\alpha_2}$. . . . . 44
2.3 Controls for the resolution of $(\mathbf{OCP})_1$. . . . . 45
2.4 Controls for the resolution of $(\mathbf{OCP})_\alpha$ with $\alpha = 0.36$. . . . . 45
2.5 Controls for the resolution of $(\mathbf{OCP})_\alpha$ with $\alpha = 0.02$. . . . . 46
2.6 Controls for the resolution of $(\mathbf{OCP})$. . . . . 46
2.7 Trajectory for the resolution of problem $(\mathbf{OCP})$. The attitude of the launcher is controlled from the initial state $x_0$ ($\diamond$) to the final state $x_f$ ($\diamond$). . . . . 47
2.8 Controls for the resolution of the problem $(\mathbf{OCP})$ with a state constraint. . . . . 48
2.9 Trajectory for the resolution of the problem $(\mathbf{OCP})$ with a constraint on the transverse angular velocities $q$ and $r$. . . . . 48

3.1 Continuation procedure to solve $(\mathcal{P})_{via,0}$. . . . . 62
3.2 Evolution of the angular velocity during the continuation. . . . . 64
3.3 Solution of $(\mathcal{P})_{via,0}$, steering the system from $x_0$ ($\diamond$) from $x_f$ ($\diamond$), satisfying a via-point constraint ($\diamond$). . . . . 65
3.4 Controls for the resolution of $(\mathcal{P})_{via,0}$. . . . . 65

4.1 Needle-like variation $u_{\pi_1}$ of a control $u$. . . . . 69
4.2 Changing the switching times induces a displacement at the final time. . . . . 70
4.3 Principle of adding needles. . . . . 82
4.4 Improving the robustness of a trajectory adding needles. We lose optimality with respect to the consumption in order to gain robustness. . . . . 86

4.5 Size of the maximal perturbation absorbed with respect to the robustness of a trajectory . . . 87
4.6 Reference, perturbed and corrected trajectories for $\varepsilon = 0.78$, $C_r = 2.22$. . . . 88
4.7 Tracking results for several values of $\varepsilon$. . . . 89
4.8 Shifting an opening time is equivalent to add a needle. . . . 91

5.1 Continuation procedure to solve ($\mathbf{OCPPDE_1}$) for $T = 60$. . . . 107
5.2 Intermediate steps of the continuation procedure for the test case 1. . . . 108
5.3 Continuation procedure to solve ($\mathbf{OCPPDE_1}$) for $T = 80$. . . . 109
5.4 Evolution of the constraint (5.2) during the continuation. . . . 110
5.5 Raising the maximal values $u_1^{max}$, $u_2^{max}$ for the controls. . . . 110
5.6 Evolution of $n_C$ for the optimal solution of ($\mathbf{OCPPDE_1}$). In black with a thick line, the initial condition $n_C(0, \cdot)$, with lighter shades of red, the evolution of $n_C(t, x)$ as time increases. At final time, the population of cancer cells is drawn with a thick red line. . . . 111
5.7 Adding a term accounting for the $L^1$ norm $\int \rho_C$ in the cost. . . . 112

A.1 Description of the software to optimize a complete ballistic phase. The previously mentionned data that has to be provided by the user is given in the file `data.txt`. 119
A.2 Trajectory for the optimization of a whole ballistic phase, starting from $\diamond$. Each $\diamond$ stands for the separation of a body. . . . 120
A.3 Angular velocity for the optimization of a whole ballistic phase, starting from $\diamond$. Each $\diamond$ stands for the separation of a body. Each $\diamond$ corresponds to the control of the angular velocity to 0. . . . 120
A.4 Controls for the optimization of a whole ballistic phase. . . . 121

C.1 In dimension 2, image by an inverible matrix of the unit sphere $\mathbb{S}^1$. . . . 130

# Introduction générale au problème de contrôle d'attitude

## Positionnement du problème

L'accès autonome à l'espace est un axe de développement national et européen majeur depuis la fin de la seconde guerre mondiale. Les enjeux géopolitiques sont plus importants que jamais. Un élément clé de cette politique est la disponibilité d'un lanceur, c'est à dire d'un véhicule ayant la capacité d'emporter des charges utiles (satellite commercial ou institutionnel, sonde d'exploration, cargo vers la station spatiale internationale...) vers une orbite depuis laquelle elles pourront réaliser leur mission.

La fonction principale d'un lanceur est alors d'injecter un satellite sur une orbite, avec un état cinématique requis. Nous entendons par là qu'il est nécessaire de pouvoir assurer la séparation du satellite dans une certaine orientation, avec une certaine vitesse angulaire. Ces données d'attitude [1] sont d'une importance cruciale dans la pratique ; citons quelques exemples de contraintes qui imposent d'être capable de contrôler l'attitude au moment de la séparation :

- L'orientation par rapport au soleil pour des besoins thermiques, ou énergétiques en présence de panneaux solaires.
- L'orientation par rapport à la terre, afin d'assurer une visibilité depuis les stations au sol recevant les données télémétriques.
- La mise en rotation des charges utiles lors de leur séparation. En effet, sous certaines conditions géométriques, un corps rigide tournant selon son axe principal présente des propriétés de stabilité.

### Géométrie générale d'un lanceur Ariane 5

Dans cette partie, on souhaite donner rapidement quelques éléments sur la géométrie d'un lanceur Ariane. Le dernier vol du lanceur Ariane 4 ayant eu lieu en 2003, c'est actuellement le programme d'Ariane 5 qui est exploité, et c'est sur celui ci que l'on se concentre. Ce programme a été voté en 1987, pour un premier vol en 1996. Il est actuellement prévu que les lancements se poursuivent jusqu'au début des années 2020. Ce lanceur ayant été conçu afin de rester compétitif au cours de cette longue période, plusieurs versions successives ont vu le jour. Mentionnons par exemple (*Source : CNES*) :

- Ariane 5 G,
- Ariane 5 G+,

1. Nous reviendrons dans la suite sur une définition de ce terme d'attitude. Pour l'instant il est suffisant de savoir qu'il désigne à la fois l'orientation du lanceur ou du satellite dans l'espace, ainsi que sa vitesse angulaire.

- Ariane 5 GS,
- Ariane 5 ES,
- Ariane 5 ECA.

Ces différentes versions ont permis l'introduction de modifications (allant du remplacement d'un moteur au remplacement d'un étage complet) permettant par exemple d'augmenter la performance du lanceur (i.e., d'augmenter sa capacité à envoyer des charges utiles de plus en plus lourdes en orbite), ou d'acquérir de la versatilité pour l'étage supérieur (possibilité de rallumage en orbite). Aujourd'hui, la performance mise en avant par le CNES et Arianespace est d'une dizaine de tonnes en orbite géostationnaire pour la version "ECA" (*Source : CNES*).

De part la grande variabilité au sein de cette famille de lanceur, nous nous contenterons de donner des éléments de géométrie qui nous paraissent représentatifs du programme de développement Ariane 5, et permettent de donner une idée générale de la chronologie d'un lancement Ariane. Sur la Figure 1, nous donnons une vue globale d'un lanceur Ariane 5. Dans les prochains paragraphes, nous donnerons plus de détails sur les différents composants du lanceur.

**EAP et EPC.** Ariane 5, dans sa partie basse, est composée de son Étage Principal Cryotechnique au centre, entouré de deux Étages d'Accélération à Poudre (EAP), comme représenté sur la Figure 2.

Les EAP fournissent 92% de la poussée au moment du décollage. Dotés d'une propulsion solide, une de leurs particularités est de ne pas pouvoir être éteints après leur mise à feu. Lorsque l'ordinateur de bord détecte une baisse significative de la poussée, environ deux minutes après le décollage, ils sont séparés du lanceur et retombent dans l'océan.

L'EPC quant à lui est allumé 7 secondes avant le décollage. Même s'il ne fournit que les 8% de poussées restants au moment du décollage, le moteur Vulcain qui l'équipe assure seul l'essentiel de la poussée du lanceur dès que les EAP sont séparés. Il fonctionne alors environ 7 minutes supplémentaires, avant de s'éteindre et l'EPC peut être séparé à son tour.

Ces données proviennent du manuel utilisateur d'Ariane 5 [Ari16], qui contient largement plus de détails sur la conception et la composition des EAP et de l'EPC.

**Composite supérieur.** Posée sur l'EPC, la partie supérieure d'Ariane 5 est représentée sur la Figure 3 dans différentes versions. Ce composite est formé de l'étage supérieur (avec notamment ses réservoirs et son moteur), la case à équipement du lanceur (contenant notamment toute l'avionique d'Ariane 5), l'adaptateur de charges utiles, la ou les charge(s) utile(s), l'éventuel système de lancement double (sur lequel nous nous attardons au paragraphe suivant) et la coiffe protégeant tous ces éléments. Notons que bien que faisant partie du composite supérieur, la coiffe est séparée avant l'EPC. En effet, sa vocation est de protéger les charges utiles des frottements avec l'atmosphère lors du décollage. Lorsque ces frottements deviennent suffisamment faibles, la coiffe est larguée afin d'alléger le lanceur. L'un des organes essentiels de l'étage supérieur pour ce travail de thèse est le système de contrôle d'attitude du lanceur. Nous reviendrons plus en détails dans la section suivante.

**Lancement double Ariane.** Dans ce paragraphe, on souhaite insister sur un élément particulier du composite supérieur, le SYstème de Lancement Double Ariane (SYLDA), qui a été utilisé pour la première fois en 2000. Son introduction a été d'une grande importance pratique, car il permet de réaliser de manière systématique des lancements doubles, en plaçant deux satellites en orbite. En effet, en plaçant deux charges utiles en orbite par vol d'Ariane 5, le coût de lancement d'un satellite est diminué. Cela impose alors de concevoir des phases balistiques plus complexes, avec plus de contraintes provenant, entre autres, des différents largages. Lorsqu'elle

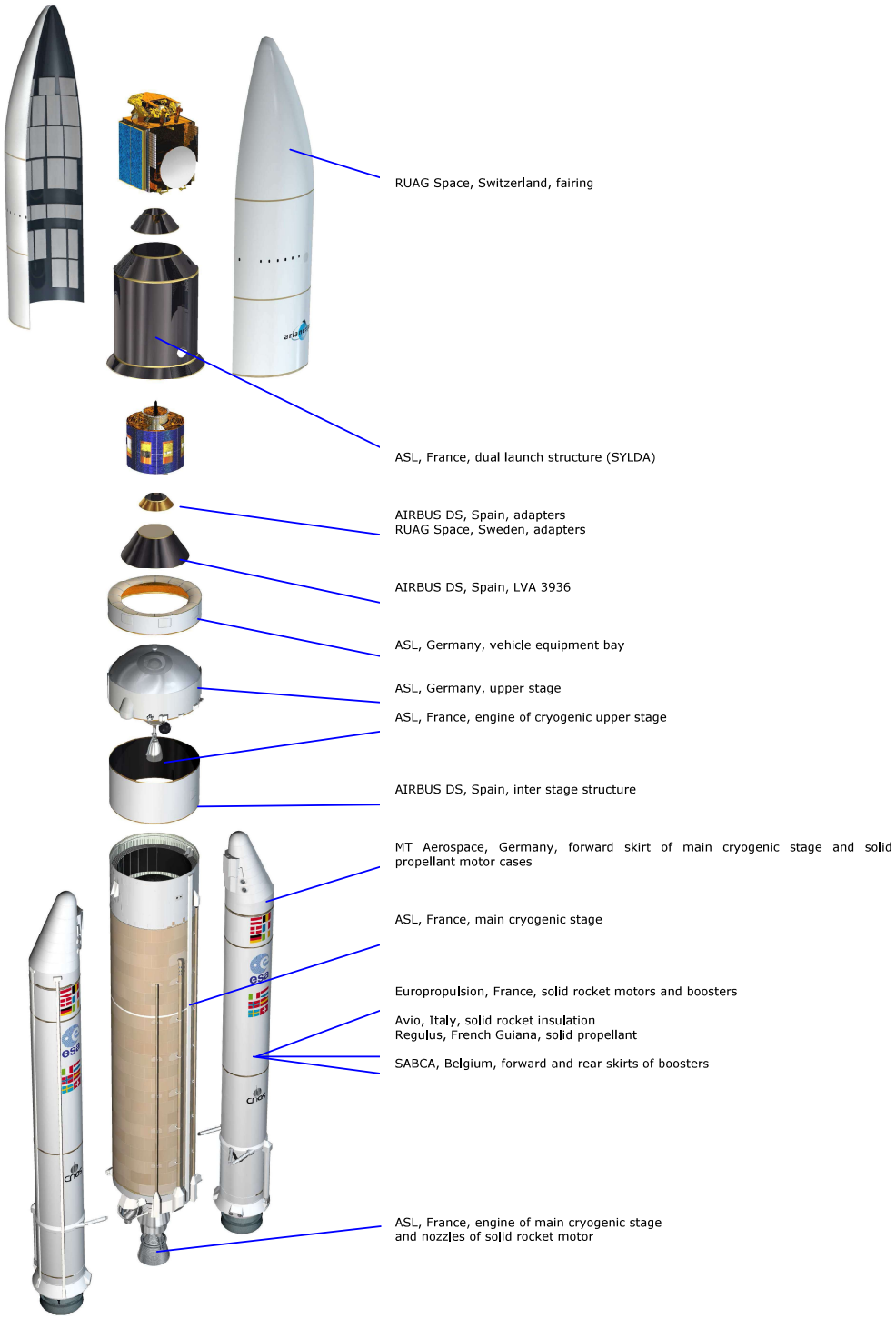



FIGURE 1 – Vue éclatée d'un lanceur Ariane 5 (*Source : Manuel utilisateur Ariane 5*).



FIGURE 2 – Représentation de la partie basse d'Ariane 5, avec les deux EAP et l'EPC. *Source : CNES.*

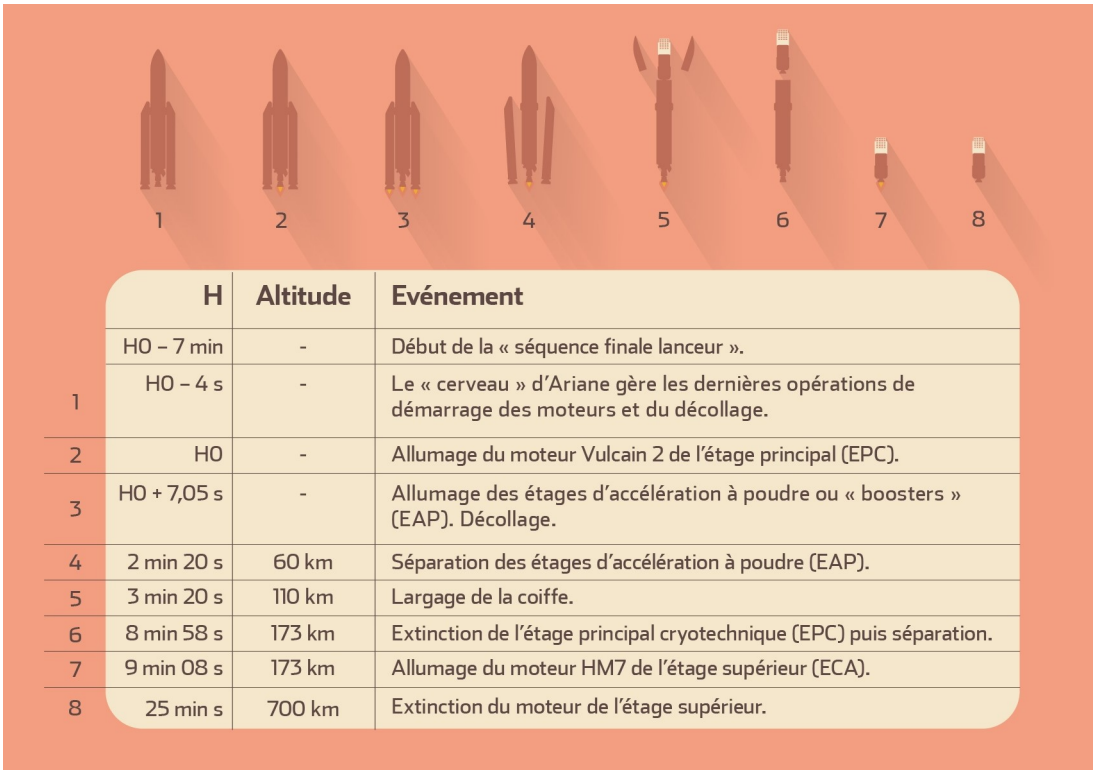



FIGURE 3 – Représentation de la partie haute d'Ariane 5. *Source : CNES.*

est utilisée, cette structure doit également être séparée, après le largage du premier satellite, et avant d'entamer les manœuvres menant à la séparation du second satellite. Ce système est bien visible sous la coiffe d'Ariane 5 ECA, à la Figure 3.

**Chronologie d'un lancement Ariane 5.** La Figure 4 récapitule les différentes étapes de la phase propulsée d'un lancement Ariane 5 (c'est-à-dire, jusqu'à l'extinction du moteur principal

de l'étage supérieur). Il s'agit de valeurs moyennes, données pour un lancement vers une orbite de transfert géostationnaire (GTO).




| | H | Altitude | Evénement |
|---|---|---|---|
| | H0 – 7 min | - | Début de la « séquence finale lanceur ». |
| 1 | H0 – 4 s | - | Le « cerveau » d'Ariane gère les dernières opérations de démarrage des moteurs et du décollage. |
| 2 | H0 | - | Allumage du moteur Vulcain 2 de l'étage principal (EPC). |
| 3 | H0 + 7,05 s | - | Allumage des étages d'accélération à poudre ou « boosters » (EAP). Décollage. |
| 4 | 2 min 20 s | 60 km | Séparation des étages d'accélération à poudre (EAP). |
| 5 | 3 min 20 s | 110 km | Largage de la coiffe. |
| 6 | 8 min 58 s | 173 km | Extinction de l'étage principal cryotechnique (EPC) puis séparation. |
| 7 | 9 min 08 s | 173 km | Allumage du moteur HM7 de l'étage supérieur (ECA). |
| 8 | 25 min s | 700 km | Extinction du moteur de l'étage supérieur. |

FIGURE 4 – Chronologie d'un vol Ariane 5 ECA. *Source : CNES.*

Cette phase de vol, dite phase propulsée, constitue en soit un sujet de recherche à part entière, riche en questions mathématiques diverses. Néanmoins, le centre d'intérêt de cette thèse est la phase de vol suivante, dite phase balistique, qui va permettre de séparer sur l'orbite souhaitée les satellites. Nous nous attardons plus en détails sur cette phase balistique dans la sous-section suivante.

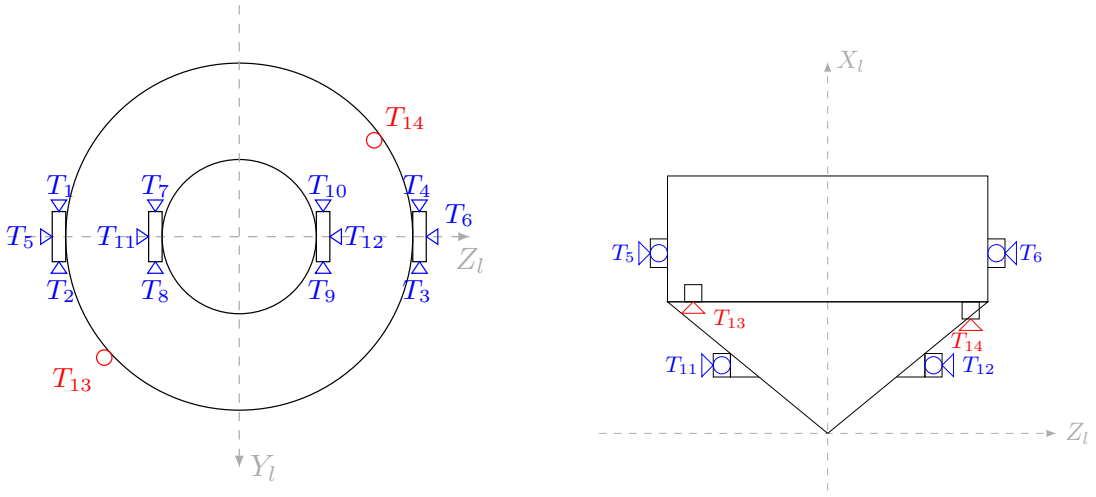
## Phase balistique

Rappelons pour commencer que la fonction d'un lanceur est de séparer une ou plusieurs charges utiles sur une orbite donnée, *dans un état d'attitude prescrit*. L'atteinte de l'orbite visée est assurée par les phases propulsées du vol. Il s'agit des phases où, successivement, les EAP, le moteur principal de l'EPC, puis celui de l'étage supérieur sont actifs. Il suit une phase dite balistique durant laquelle le contrôle de l'attitude du lanceur est assuré par le Système de Contrôle d'Attitude (SCA), en vue de la séparation des charges utiles. Par opposition aux phases de poussées, cette phase désigne la période pendant laquelle les moteurs principaux sont éteints. Le rôle du SCA est donc d'orienter l'étage supérieur et ses charges utiles afin d'atteindre une attitude donnée permettant de satisfaire les différentes contraintes liées au lanceur ou au(x) satellite(s).

Le SCA est l'ensemble des composants assurant la génération de la poussée nécessaire à la réalisation des objectifs de la phase balistique : le moteur principal du lanceur est éteint, et de

petites poussées sont réalisées par un ensemble de tuyères réparties sur l'engin.

Nous représentons sur la Figure 5 un schéma du SCA, comme considéré dans les travaux de ce travail de thèse. Il est constitué d'un ensemble de tuyères (14 sur le schéma) dont le nombre peut varier d'un lanceur à l'autre. L'alimentation du SCA diffère également entre les différentes versions du développement d'Ariane 5. Par exemple, pour Ariane 5 ECA, il est alimenté par du dihydrogène gazeux ou du dioxygène gazeux ; sur Ariane 5 ES, de l'hydrazine est également utilisée. Il est possible d'ouvrir ou de fermer chaque tuyère afin de produire une force de poussée et un couple, mais on n'en contrôle ni le débit, ni l'orientation. C'est donc le SCA qui nous permet d'exercer un *contrôle* sur le système. Précisons que sur un lanceur de type Ariane 5, deux tuyères sont généralement utilisées pour les contraintes d'éloignement entre les corps. Sur la Figure 5, il s'agit des tuyères 13 et 14, représentées en rouge. Ces tuyères n'étant pas strictement utilisées pour faire du contrôle d'attitude, nous les omettrons parfois dans la suite de la thèse.




(a) Vue du dessus (b) Vue de profil

FIGURE 5 – Schéma du Système de Contrôle d'Attitude

C'est également lors de la phase balistique que sont réalisées les manœuvres nécessaires aux dispositions de fin de vie et de passivation du composite supérieur. Ces exigences sont liées à la Loi sur les Opérations Spatiales, relatives notamment à la sécurité des personnes et au respect de l'environnement orbital.

**Objectifs de la thèse.** La recherche d'une séquence d'activation des tuyères permettant d'amener le lanceur dans les états d'attitude souhaités est conditionnée au respect d'un certain nombre de contraintes physiques provenant de la conception du lanceur. Par exemple, la quantité d'ergols disponibles pour effectuer les manœuvres est limitée. La durée d'ouverture des tuyères ou la durée totale de la phase balistique sont également sujettes à des bornes que l'on ne doit pas dépasser. Nous reviendrons, lors de la formulation du problème de contrôle optimal, sur le choix d'un coût permettant d'intégrer ces contraintes.

Bien que conçue de façon spécifique pour chaque analyse de mission, la phase balistique n'a pas jusqu'à présent fait l'objet à notre connaissance d'une méthode d'élaboration systématique et déposée, sa conception reposant essentiellement sur le savoir-faire des analystes de mission.

C'est donc la recherche de la trajectoire du lanceur au cours de la phase balistique que l'on étudie dans cette thèse. Nous y développons un cadre mathématique précis permettant de

formuler ce problème comme un problème de contrôle optimal. L'objectif de la thèse est de concevoir et d'implémenter une méthode mathématique permettant d'automatiser et d'optimiser l'élaboration et la planification d'une phase balistique.

# Modélisation du problème de contrôle d'attitude

## Évolution de l'orientation d'un lanceur.

On se donne un repère mobile $(\overrightarrow{X}_\ell(t), \overrightarrow{Y}_\ell(t), \overrightarrow{Z}_\ell(t))$, attaché au lanceur en son centre de gravité, ainsi qu'un repère inertiel $(\overrightarrow{X}_i, \overrightarrow{Y}_i, \overrightarrow{Z}_i)$. Repérer l'attitude du lanceur revient à repérer la position du repère mobile par rapport au repère inertiel, c'est-à-dire calculer l'expression de la matrice de passage entre le repère mobile et le repère inertiel :

$$R(t) = \begin{pmatrix} \langle \overrightarrow{X}_\ell(t), \overrightarrow{X}_i \rangle & \langle \overrightarrow{X}_\ell(t), \overrightarrow{Y}_i \rangle & \langle \overrightarrow{X}_\ell(t), \overrightarrow{Z}_i \rangle \\ \langle \overrightarrow{Y}_\ell(t), \overrightarrow{X}_i \rangle & \langle \overrightarrow{Y}_\ell(t), \overrightarrow{Y}_i \rangle & \langle \overrightarrow{Y}_\ell(t), \overrightarrow{Z}_i \rangle \\ \langle \overrightarrow{Z}_\ell(t), \overrightarrow{X}_i \rangle & \langle \overrightarrow{Z}_\ell(t), \overrightarrow{Y}_i \rangle & \langle \overrightarrow{Z}_\ell(t), \overrightarrow{Z}_i \rangle \end{pmatrix},$$

qui est un élément de $SO_3(\mathbb{R})$.

Le vecteur de vitesse angulaire du lanceur est défini, en repère lanceur, par $[\overrightarrow{\omega}]_l = (p, q, r)$. C'est à dire que

$$\overrightarrow{\omega} = p\overrightarrow{X_\ell} + q\overrightarrow{Y_\ell} + r\overrightarrow{Z_\ell}.$$

Suivant ses caractéristiques géométriques, un lanceur présente généralement un axe principal d'inertie : sur la Figure 5, il s'agit de l'axe $\overrightarrow{X}_l$. Dans la suite de ce travail de thèse, on appellera parfois vitesse de roulis la vitesse angulaire suivant cet axe (c'est-à-dire $p$), et on utilisera l'appellation vitesses angulaires transverses pour les composantes $q$ et $r$.

L'équation décrivant l'évolution de l'orientation du lanceur est alors

$$\dot{R}(t) = \begin{pmatrix} 0 & r & -q \\ -r & 0 & p \\ q & -p & 0 \end{pmatrix} R(t),$$

qui exprime la rotation du repère $(\overrightarrow{X}_\ell(t), \overrightarrow{Y}_\ell(t), \overrightarrow{Z}_\ell(t))$ à la vitesse angulaire $[\overrightarrow{\omega}]_l = (p, q, r)$.

## Évolution de la vitesse angulaire

**Paramétrisation de $SO_3(\mathbb{R})$.** Dans ce travail de thèse, on a choisit de repérer la position des axes du lanceur par trois rotations, dont les angles sont parfois désignés *angles de Cardan*, et qui constituent une variation des angles d'Euler. Le repère lanceur $(\overrightarrow{X}_l, \overrightarrow{Y}_l, \overrightarrow{Z}_l)$ est obtenu à partir du repère inertiel $(\overrightarrow{X}_i, \overrightarrow{Y}_i, \overrightarrow{Z}_i)$ par la série suivante de rotations, représentée à la Figure 6 :

- d'angle $\theta$ autour de l'axe $\overrightarrow{Z}_i$, qui donne le repère $(\overrightarrow{X}_1, \overrightarrow{Y}_1, \overrightarrow{Z}_1)$,
- d'angle $\psi$ autour de l'axe $\overrightarrow{Y}_1$, qui donne le repère $(\overrightarrow{X}_2, \overrightarrow{Y}_2, \overrightarrow{Z}_2)$,
- et enfin d'angle $\varphi$ autour de l'axe $\overrightarrow{X}_2$, qui donne le repère $(\overrightarrow{X}_l, \overrightarrow{Y}_l, \overrightarrow{Z}_l)$,

Rappelons que la vitesse angulaire est définie, en repère lanceur, par $\overrightarrow{\omega} = p\overrightarrow{X_\ell} + q\overrightarrow{Y_\ell} + r\overrightarrow{Z_\ell}$.

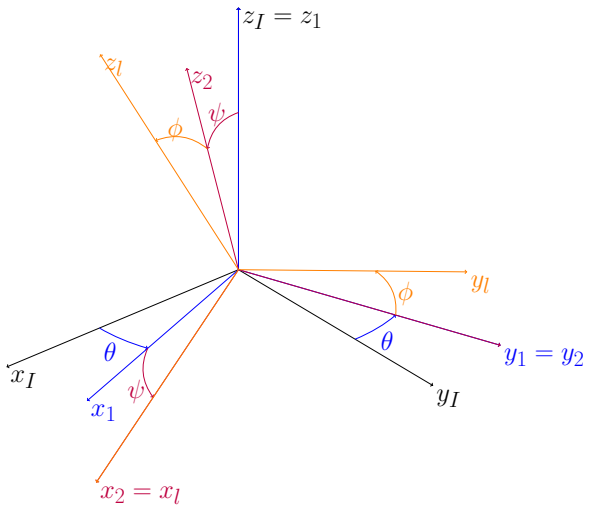



FIGURE 6 – Séquence des trois angles d'attitude.

Or, en utilisant la séquence de rotations précédemment introduite, on a également

$$\begin{aligned}
\overrightarrow{\omega} &= \dot{\theta}\overrightarrow{Z_i} + \dot{\psi}\overrightarrow{Y_1} + \dot{\varphi}\overrightarrow{X_2} \\
&= \dot{\theta}\overrightarrow{Z_1} + \dot{\psi}\overrightarrow{Y_1} + \dot{\varphi}\overrightarrow{X_2} \\
&= \dot{\theta}(-\sin\psi\overrightarrow{X_2} + \cos\psi\overrightarrow{Z_2}) + \dot{\psi}\overrightarrow{Y_2} + \dot{\varphi}\overrightarrow{X_2} \\
&= (\dot{\varphi} - \dot{\theta}\sin\psi)\overrightarrow{X_2} + \dot{\psi}\overrightarrow{Y_2} + \dot{\theta}\cos\psi\overrightarrow{Z_2} \\
&= (\dot{\varphi} - \dot{\theta}\sin\psi)\overrightarrow{X_\ell} + \dot{\psi}(\cos\varphi\overrightarrow{Y_\ell} - \sin\varphi\overrightarrow{Z_\ell}) + \dot{\theta}\cos\psi(\sin\varphi\overrightarrow{Y_\ell} + \cos\varphi\overrightarrow{Z_\ell}) \\
&= (\dot{\varphi} - \dot{\theta}\sin\psi)\overrightarrow{X_\ell} + (\dot{\theta}\cos\psi\sin\varphi + \dot{\psi}\cos\varphi)\overrightarrow{Y_\ell} + (\dot{\theta}\cos\psi\cos\varphi - \dot{\psi}\sin\varphi)\overrightarrow{Z_\ell}
\end{aligned}$$

Ainsi, en identifiant les termes, on obtient l'expression suivante pour la vitesse angulaire

$$\begin{pmatrix} p \\ q \\ r \end{pmatrix} = \begin{pmatrix} 1 & 0 & -\sin\varphi \\ 0 & \cos\varphi & \cos\psi\sin\varphi \\ 0 & -\sin\varphi & \cos\psi\cos\varphi \end{pmatrix} \cdot \begin{pmatrix} \dot{\varphi} \\ \dot{\psi} \\ \dot{\theta} \end{pmatrix}$$

Le calcul du déterminant de la matrice donne :

$$\det \begin{pmatrix} 1 & 0 & -\sin\varphi \\ 0 & \cos\varphi & \cos\psi\sin\varphi \\ 0 & -\sin\varphi & \cos\psi\cos\varphi \end{pmatrix} = \cos^2\varphi\cos\psi + \sin^2\varphi\cos\psi = \cos\psi.$$

Cette représentation par des angles de Cardan introduit donc une singularité quand $\cos\psi = 0$, i.e., $\psi \equiv \pi/2 \pmod{\pi}$. Si $\psi \not\equiv \pi/2 \pmod{\pi}$ la matrice est inversible et on obtient les équations d'évolution pour les angles $\theta$, $\psi$ et $\varphi$ :

$$\begin{pmatrix} \dot{\varphi} \\ \dot{\psi} \\ \dot{\theta} \end{pmatrix} = \begin{pmatrix} 1 & \sin\varphi\tan\psi & \cos\varphi\tan\psi \\ 0 & \cos\varphi & -\sin\varphi \\ 0 & \frac{\sin\varphi}{\cos\psi} & \frac{\cos\varphi}{\cos\psi} \end{pmatrix} \cdot \begin{pmatrix} p \\ q \\ r \end{pmatrix}. \tag{1}$$

**Remarque 0.1: Représentation de $SO_3(\mathbb{R})$**

Le choix de la série de rotations pour repérer la position des axes du lanceur n'est pas unique. La matrice $R(t)$ introduite au paragraphe précédent est un élément de $SO_3(\mathbb{R})$, une sous-variété de $\mathcal{M}_3(\mathbb{R})$ de dimension 3, et de manière plus générale, il existe plusieurs représentations possibles pour $SO_3(\mathbb{R})$. Mentionnons par exemple les représentations classiques :

- avec trois paramètres, les angles d'Euler, les paramètres de Rodrigues.
- avec quatre paramètres, les quaternions, la représentation angle/axe.

Notons que les représentations à trois paramètres ne sont pas globales : par exemple, quelle que soit la séquence choisie pour les angles d'Euler, une singularité apparaîtra. Quant aux représentations à quatre paramètres, elles ne présentent pas de propriété d'unicité : par exemple, si $q$ est un quaternion unitaire représentant une attitude $R \in SO_3(\mathbb{R})$, le quaternion $-q$ représente la même attitude.
L'article [Cea11] et le livre [BFT06] évoquent plus en détails le problème du choix d'une représentation pour $SO_3(\mathbb{R})$.

**Équations pour la vitesse angulaire.** On va maintenant établir les équations d'évolution pour la vitesse angulaire $\vec{\omega} = (p, q, r)$, en s'inspirant de la présentation de [D.10]. Soit $\vec{H}_G$ le moment cinétique par rapport au centre de gravité du lanceur, qui s'exprime

$$\vec{H}_G = I\vec{\omega},$$

où $I$ désigne la matrice d'inertie du lanceur. Le théorème du moment cinétique affirme que la dérivée du moment cinétique est égale à la somme des moments (par rapport au centre de gravité) des forces s'exerçant sur l'objet :

$$\frac{d}{dt}\vec{H}_G = \sum \vec{M}_G(\vec{f}).$$

En outre, le repère lanceur étant attaché de manière rigide au lanceur, on peut également exprimer la dérivée du moment cinétique

$$\frac{d}{dt}\vec{H}_G = \left.\frac{d}{dt}\right|_{rel} \vec{H}_G + \vec{\omega} \wedge \vec{H}_G.$$

Le premier terme du membre de droite désigne la dérivée dans le repère mobile du moment cinétique, c'est-à-dire que

$$\left.\frac{d}{dt}\right|_{rel} \vec{H}_G = I\dot{\vec{\omega}},$$

et on obtient finalement en regroupant les équations précédentes

$$I\dot{\vec{\omega}} + \vec{\omega} \wedge I\vec{\omega} = \sum \vec{M}_G(\vec{f}).$$

Une tuyère, placée au point $A_j$ par rapport au centre de gravité $G$, et produisant à l'instant $t$ une force de poussée $\vec{P}(t)$ induit un couple sur le lanceur $\vec{P}(t) \wedge \overrightarrow{A_jG}$. Comme on ne contrôle ni le débit d'ergols, ni l'orientation de la tuyère, la poussée peut s'écrire $\vec{P} = u(t)\vec{P}$, où $\vec{P}$ est un vecteur constant de $\mathbb{R}^3$ et la fonction $u(\cdot)$ est une fonction constante par morceaux, avec $u = 1$

si la tuyère est ouverte et $u = 0$ si la tuyère est fermée. Le couple peut alors se réécrire $u\overrightarrow{b_j}$, où $\overrightarrow{b_j} = \overrightarrow{P} \wedge \overrightarrow{A_jG}$ est un vecteur constant de $\mathbb{R}^3$. En notant $m$ le nombre de tuyères sur le lanceur, l'équation d'évolution pour la vitesse angulaire s'écrit donc :

$$I\dot{\overrightarrow{\omega}}(t) + \overrightarrow{\omega}(t) \wedge I\overrightarrow{\omega}(t) = \sum_{j=1}^{m} u(t)\overrightarrow{b_j}. \tag{2}$$

En regroupant les équations (1) et (2), on obtient les équations d'Euler complètes pour l'attitude d'un corps rigide :

$$\left\{ \begin{array}{rcl} \dot{\theta}(t) & = & \frac{\sin\varphi(t)}{\cos\psi(t)} q(t) + \frac{\cos\varphi(t)}{\cos\psi(t)} r(t) \\ \dot{\psi}(t) & = & \cos\varphi(t) \cdot q(t) - \sin\varphi(t) \cdot r(t) \\ \dot{\varphi}(t) & = & p(t) + \sin\varphi(t)\tan\psi(t) \cdot q(t) + \cos\varphi(t)\tan\psi(t) \cdot r(t) \\ I\dot{\overrightarrow{\omega}}(t) & = & I\overrightarrow{\omega}(t) \wedge \overrightarrow{\omega}(t) + \sum_{j=1}^{m} u(t)\overrightarrow{b_j}. \end{array} \right. \tag{3}$$

Dans la suite, on utilisera souvent la notation $x = (\theta, \psi, \varphi, p, q, r)$ pour désigner l'état du lanceur, $u = (u_i)_{1 \leqslant i \leqslant m}$ pour désigner le contrôle, et on notera la dynamique sous la forme condensée

$$\dot{x}(t) = f(x(t), u(t)),$$

ou encore, afin de faire apparaître le caractère affine par rapport aux contrôles,

$$\dot{x}(t) = f_0(x(t)) + \sum_{j=1}^{m} u_j(t) f_j(x(t)),$$

où $f_0$ correspond aux équations libres du mouvement, et pour $j \geqslant 1$, $f_j(x(t))$ est un champ de vecteurs constant, égal à $(0_{\mathbb{R}^3}, \overrightarrow{b_j})$.

**Cas d'une matrice d'inertie diagonale.** Si les axes du repère mobile sont alignés avec les axes principaux du lanceur, la matrice d'inertie est diagonale

$$I = \begin{pmatrix} I_x & 0 & 0 \\ 0 & I_y & 0 \\ 0 & 0 & I_z \end{pmatrix}.$$

Dans ce cas, les équations pour la vitesse angulaire dans (3) deviennent

$$\begin{aligned} I_x\dot{p} &= (I_y - I_z)qr + \sum_{j=1}^{m} u(t) b_j^1 \\ I_y\dot{q} &= (I_z - I_x)pr + \sum_{j=1}^{m} u(t) b_j^2 \\ I_z\dot{r} &= (I_x - I_y)pq + \sum_{j=1}^{m} u(t) b_j^3. \end{aligned}$$

En introduisant les coefficients numériques

$$a_1 = \frac{I_y - I_z}{I_x}, \quad a_2 = \frac{I_z - I_x}{I_y}, \quad a_3 = \frac{I_x - I_y}{I_z},$$

et avec un léger abus de notation, car on gardera l'écriture $\overrightarrow{b}_j$ pour désigner le couple normalisé $I^{-1}\overrightarrow{b}_j$ produit par une tuyère $j$, on peut écrire les équations sous la forme

$$\begin{aligned}
\dot{p} &= a_1 qr + \sum_{j=1}^{m} u(t) b_j^1 \\
\dot{q} &= a_2 pr + \sum_{j=1}^{m} u(t) b_j^2 \\
\dot{r} &= a_3 pq + \sum_{j=1}^{m} u(t) b_j^3.
\end{aligned}$$

Cette simplification est justifiée en pratique par la géométrie du lanceur, qui présente (presque) une symétrie le long de son axe principal. Dans la suite de ce travail de thèse, et sauf mention du contraire, c'est le cadre que l'on considère.

**Équations d'attitude d'un corps rigide - Matrice d'inertie diagonale**

$$\left\{\begin{array}{rcl}
\dot{\theta}(t) &=& \frac{\sin\varphi(t)}{\cos\psi(t)} q(t) + \frac{\cos\varphi(t)}{\cos\psi(t)} r(t) \\
\dot{\psi}(t) &=& \cos\varphi(t) \cdot q(t) - \sin\varphi(t) \cdot r(t) \\
\dot{\varphi}(t) &=& p(t) + \sin\varphi(t)\tan\psi(t) \cdot q(t) + \cos\varphi(t)\tan\psi(t) \cdot r(t) \\
\dot{p}(t) &=& a_1 q(t) r(t) + \sum_{j=1}^{m} u(t) b_j^1 \\
\dot{q}(t) &=& a_2 p(t) r(t) + \sum_{j=1}^{m} u(t) b_j^2 \\
\dot{r}(t) &=& a_3 p(t) q(t) + \sum_{j=1}^{m} u(t) b_j^3.
\end{array}\right. \tag{4}$$

## Contrôle optimal.

Outre la recherche de trajectoires respectant les équations de la dynamique (3.3.1) et permettant d'amener le lanceur dans l'état d'attitude souhaité, on cherche des trajectoires optimales pour un certain critère. Le choix de ce critère revêt une importance particulière dans la modélisation et la formulation d'un problème de contrôle optimal. En effet, nous verrons au Chapitre 2 que l'expression des conditions nécessaires du *principe du maximum de Pontryagin* peut conduire à des contrôles présentant des structures bien différentes.

Une des contraintes imposée par la conception du SCA est de ne permettre l'utilisation de contrôles ne prenant que les valeurs 0 ou 1. Dans la littérature, de tels contrôles sont généralement qualifiés de *bang-bang*. Le coût que l'on cherche à minimiser lors de la phase balistique est la consommation en ergols du lanceur, qui est proportionnelle à la durée d'ouverture totale des tuyères (et dont nous verrons au Chapître 2 qu'il mène bien à des contrôles bang-bang) :

$$\int_0^{t_f} \sum_{i=1}^{m} |u_j(t)|\, dt.$$

Le temps final $t_f$ étant destiné à être laissé libre, la minimisation de ce seul critère peut conduire à obtenir une suite de trajectoires en temps tendant vers $+\infty$. Afin de s'en prévenir, on ajoute une pondération

$$\int_0^{t_f} \sum_{i=1}^{m} |u_j(t)|\, dt + \lambda_0 t_f, \tag{5}$$

où $\lambda_0$ dépend de l'importance que l'on souhaite donner au temps final.

Le problème de contrôle optimal consiste alors, étant donné un point initial $x_0$ et un point final $x_f$, à trouver un contrôle $u(\cdot)$, optimal pour le coût (5), tel que la trajectoire associée $x(\cdot)$ vérifie $x(0) = x_0$ et $x(t_f) = x_f$ :

$$(\mathbf{OCP}) \begin{cases} \min \quad \int_0^{t_f} \sum_{i=1}^{m} |u_j(t)|\, dt + \lambda_0 t_f, \\ \dot{x}(t) = f(x(t), u(t)), \\ \forall i \in [\![1, m]\!], \quad 0 \leqslant u_i(t) \leqslant 1 \quad \text{p.p. on } [0; t_f], \\ x(0) = x_0, \\ x(t_f) = x_f. \end{cases}$$

On a beaucoup insisté dans l'introduction sur le fait que la conception du SCA impose d'avoir des contrôles dans $\{0, 1\}^m$. Or, dans l'écriture du problème (**OCP**), on donne la contrainte sur les contrôles

$$\forall i \in [\![1, m]\!], \quad 0 \leqslant u_i(t) \leqslant 1.$$

Cela se justifie par le fait qu'il est commode de choisir un ensemble convexe pour appliquer les résultats usuels du contrôle optimal. En outre, nous montrerons au Chapitre 2 que le choix d'un critère $L^1$ tel que (5) mène bien à des contrôles bang-bang.

Nous verrons aussi au Chapitre 3 que nous ajouterons à cette formulation des contraintes intermédiaires sur l'état, qui s'écrivent génériquement sous la forme :

$$c(x(t_1)) = 0.$$

## Structure du manuscrit et description des contributions

Ce travail de thèse combine plusieurs études théoriques sur le contrôle optimal de systèmes non linéaires en dimension finie et la mise en œuvre numérique des algorithmes de résolution. Cette partie numérique comprend entre autres le développement d'un logiciel à destination du CNES capable d'optimiser la trajectoire d'un lanceur pour différentes situations de phase balistique, à chaque fois par la méthode numérique la plus appropriée.

Les chapitres 1 et 2 serviront à présenter un état de l'art rapide sur des résultats déjà existants de la théorie du contrôle.

Dans le chapitre 1, nous commencerons par rappeler en détails la preuve du résultat de contrôlabilité pour les équations d'attitude d'un corps rigide. Cela signifie que pour toute donnée initiale et toute donnée finale, et sous certaines hypothèses sur la géométrie du lanceur, et notamment la disposition des tuyères, il existe un contrôle permettant de réaliser le transfert entre ces deux états. De manière générale, des résultats de contrôlabilité existent lorsqu'il n'y a pas de contraintes sur le contrôle. Citons par exemple les *critères de Kalman* pour des systèmes de contrôle linéaires autonomes ou non autonomes, ou la *méthode du retour* [Cor92] pour des systèmes non linéaires. Lorsqu'il y a des contraintes sur le contrôle (ce qui est le cas dans ce travail de thèse), la question de la contrôlabilité peut être plus délicate. Cependant, les techniques du contrôle géométrique permettent de répondre à cette question lorsque le système présente

une structure affine en les contrôles. La contrôlabilité d'un système de contrôle affine peut alors s'obtenir par la combinaison de deux éléments : la *stabilité au sens de Poisson* introduite par Poincaré dans [Poi90], et une *condition de rang* sur l'algèbre de Lie engendrée par les champs de vecteurs constituant la dynamique. Ce chapitre suivra en grande partie la démonstration présentée dans [BFT06].

Le chapitre 2 présentera les éléments théoriques usuels en théorie du contrôle optimal. On cherche alors à trouver des contrôles qui sont optimaux vis à vis d'un certain critère. Dans cette thèse, il s'agit d'une combinaison linéaire entre le temps mis pour effectuer la manœuvre et la consommation en ergols. Le résultat clé de la théorie du contrôle optimal est le Principe du Maximum de Pontryagin (PMP) [PBGM62] qui énonce un ensemble de conditions nécessaires pour qu'un contrôle soit optimal. Même si ces conditions ne sont pas suffisantes, on se limite souvent dans la pratique à la recherche de solutions les satisfaisant. Les méthodes dites *indirectes* exploitent le PMP pour réduire le problème à la recherche des zéros d'une certaine fonction. Les méthodes directes quant à elles reposent sur une discrétisation totale du problème de contrôle optimal pour se ramener à un problème d'optimisation en dimension finie. À la fin du chapitre 2 , nous montrerons comment nous avons eu recours à une méthode de *continuation* afin de résoudre le problème (**OCP**) par une méthode indirecte. L'utilisation d'une telle technique est désormais un procédé standard, voir par exemple les travaux [CHT12, CHT17, GH06, CDG12, AG90]. Dans le cas d'une phase balistique simple avec un seul largage de corps, nous avons développé pour le CNES un logiciel en C implémentant cette méthode de continuation, afin d'être capable de résoudre génériquement ce type de problème de contrôle. Nous illustrerons également le principe des méthodes directes en résolvant un problème de contrôle d'attitude avec des contraintes sur l'état.

Lors du traitement de phases balistiques complexes, notamment avec plusieurs largages de charges utiles, l'utilisation du logiciel précédemment mentionné n'est plus suffisant. En effet, les différents largages induisent des contraintes aux instants des séparations successives, qui ne concernent pas nécessairement les 6 composantes décrivant l'état du lanceur. C'est par exemple le cas lors d'une séparation d'un corps spiné selon son axe principal d'inertie : l'état $\varphi$ est généralement laissé libre.

Au chapitre 3, nous utiliserons le formalisme des systèmes de contrôle hybrides de [DK08, DK11] pour résoudre un problème de contrôle optimal avec des contraintes dites *intermédiaires*. Il s'agit de contraintes sur l'état à un certain instant au cours de la trajectoire. Nous y montrerons des principes du maximum pour cette classe de problèmes. À la différence du PMP usuel présenté au chapitre 2, le vecteur adjoint n'est plus absolument continu et présente des discontinuités aux instants des contraintes intermédiaires. Même si on peut trouver des principes du maximum similaires dans la littérature, par exemple dans [BH75], nous n'avons pas trouvé de travaux généraux sur leur mise en œuvre numérique appliquée à des exemples non académiques. Nous proposons dans ce chapitre une procédure numérique permettant de résoudre, avec la grande précision offerte par les méthodes indirectes, un problème de contrôle optimal avec une contrainte intermédiaire. La contrainte intermédiaire est d'abord introduite par pénalisation dans le coût. Une fois que la pénalisation est suffisamment contraignante, la résolution d'un ultime problème de tir permet de satisfaire de manière exacte une contrainte de type $c(x(t_1)) = 0$.

Nous verrons également dans ce chapitre que lorsque le nombre de contraintes intermédiaires devient trop important (c'est par exemple le cas lorsque le CNES traite une phase balistique complexe) cette procédure peut ne plus suffire, et nous montrerons dans l'appendice A comment les méthodes directes permettent de résoudre, relativement simplement mais au prix d'une perte de précision, un tel problème. Un logiciel en C (qui utilise un algorithme de point intérieur) a d'ailleurs été développé pour le CNES permettant de résoudre par une méthode directe et en toute généralité une phase balistique complète, avec un nombre quelconque de séparations, et

également un nombre quelconque de contraintes intermédiaires à chaque largage. L'appendice A est ainsi complémentaire du chapitre 3, ces deux parties du manuscrit s'adressant aux mêmes classes de problèmes, mais par des approches différentes.

Les techniques présentées aux chapitres 2 et 3 permettent de donner des stratégies de contrôle pour des systèmes idéalisés. On entend par là qu'il n'y a pas d'incertitudes dans la dynamique, ni de perturbations au cours du mouvement. Dans le cas du système de contrôle d'attitude étudié avec le CNES dans cette thèse, les conditions réelles de vol ne sont jamais nominales, et appliquer en boucle ouverte une stratégie de contrôle préalablement calculée ne permettrait pas de contrer une éventuelle dérive au cours de la mission. Au chapitre 4, nous proposons *un algorithme de contrôle robuste*, permettant de faire face à des perturbations. L'originalité de notre approche réside dans la *préservation de la structure bang-bang* des contrôles par cet algorithme. Nous identifions également un critère permettant de quantifier la robustesse d'un contrôle bang-bang. Alors que la littérature sur les systèmes de contrôle robustes est extrêmement riche[2], nous n'avons pas connaissance d'une théorie générale permettant de traiter des perturbations par des variations du contrôle qui préservent sa structure bang-bang. Découle également de notre approche une stratégie pouvant permettre de rendre un contrôle nominal plus robuste. De manière informelle, les temps de commutation des contrôles peuvent être vus comme des degrés de liberté dans la commande du système. Avec cette vision, plus il y a de degrés de liberté dans la commande, plus le pilote a de possibilités pour lutter contre les perturbations. Nous verrons d'ailleurs que notre stratégie pour "robustifier" les contrôles consiste à ajouter des temps de commutation additionnels.

Enfin, au chapitre 5 nous présenterons les résultats de travaux avec Camille Pouchol. L'originalité de ce chapitre est la combinaison d'une méthode de continuation, comme celles présentées aux chapitres 2 et 3, avec une méthode directe. On y étudie un système d'équations intégro-différentielles structurées en phénotype représentant l'évolution au cours du temps de populations de cellules saines et cancéreuses. Dans ce système, le contrôle est l'administration ou non de deux types de médicaments, cytotoxiques ou cytostatiques, et on cherche à minimiser le nombre de cellules cancéreuses. L'étude théorique de ce type de système est difficile et il n'existe pas actuellement (à notre connaissance) de résultats dans le cas le plus général. La difficulté vient notamment de la présence de plusieurs contraintes sur l'état, qui rendent d'ailleurs l'utilisation de méthodes indirectes délicate. Dans le cas sans diffusion, des résultats ont cependant été obtenus dans [PT17]. On propose dans ce chapitre une procédure permettant de résoudre numériquement le problème de contrôle optimal correspondant. Le système est d'abord grandement simplifié pour permettre d'appliquer un principe du maximum en dimension infinie, puis on se ramène au problème initial par une continuation. Même si le cadre de ce chapitre s'éloigne du problème de contrôle d'attitude, nous souhaitons insister sur le fait que la technique générale s'appliquerait à une classe beaucoup plus vaste de problèmes, par exemple en aérospatial, quand le système est trop compliqué pour permettre une initialisation du programme d'optimisation sous-jacent.

Nous conclurons cette thèse en donnant quelques perspectives et en mentionnant certains problèmes ouverts.

2. Mentionnons par exemple les approches $\mathcal{H}_2$ et $\mathcal{H}_\infty$, ou la théorie linéaire quadratique permettant de "suivre" des trajectoires. Il existe également des papiers où l'algorithme de contrôle robuste préserve bien la structure bang-bang des contrôles, mais pour des systèmes bien particuliers pour lesquels la démarche ne se généralise pas. Nous ferons un état de l'art plus détaillé sur le sujet au début du Chapitre 4.

**Contributions principales de la thèse**

Résumons ici les contributions principales qui ont été apportées dans ce travail de thèse :

- Au chapitre 3, l'étude d'une procédure numérique permettant de résoudre par une méthode indirecte un problème de contrôle optimal avec des contraintes intermédiaires.
- Au chapitre 4, la conception d'un algorithme de contrôle robuste permettant de traiter des perturbations tout en préservant la structure bang-bang des contrôles. Nous y proposons également un critère pour quantifier la robustesse des trajectoires, ainsi qu'une heuristique pour "robustifier" un contrôle de référence.
- Au chapitre 5, la combinaison de méthodes directes et d'une continuation pour résoudre un problème de contrôle optimal pour une équation aux dérivées partielles. Pour initialiser la continuation, nous y montrons un résultat sur la structure des contrôles optimaux en appliquant un PMP en dimension infinie.
- En parallèle à ces travaux "théoriques", la conception et l'écriture d'un logiciel en C pour le CNES, n'utilisant que des librairies "open source". Il permet :
  - pour une phase balistique simple avec un seul largage, de calculer la trajectoire optimale du point de vue de la consommation. Il implémente la méthode de continuation présentée à la fin du chapitre 2.
  - pour une phase balistique complexe, avec un nombre quelconque de largages et de contraintes intermédiaires, de trouver par une méthode directe la solution optimale du point de vue de la consommation. C'est l'objet de l'appendice A.

Chapter 1

# Controllability of the attitude for a rigid spacecraft

## Contents


**Positionnement du problème** . . . . . **1**
Géométrie générale d'un lanceur Ariane 5 . . . . . 1
Phase balistique . . . . . 5
**Modélisation du problème de contrôle d'attitude** . . . . . **7**
Évolution de l'orientation d'un lanceur. . . . . 7
Évolution de la vitesse angulaire . . . . . 7
Contrôle optimal. . . . . 11
**Structure du manuscrit et description des contributions** . . . . . **12**

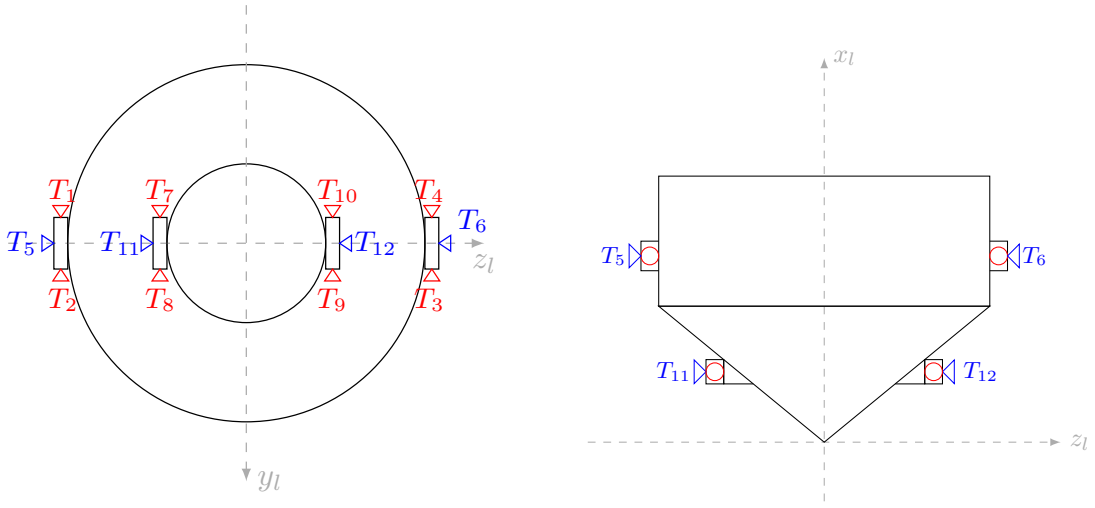

In this first chapter, we shall start by showing that the problem of interest in this thesis is well posed, that is, the attitude equations for a rigid body are controllable. It means that for every initial condition $x_0 = (\theta_0, \psi_0, \varphi_0, p_0, q_0, r_0)$ and every final condition $x_f = (\theta_f, \psi_f, \varphi_f, p_f, q_f, r_f)$, there exist a final time $t_f$ and a control $u(\cdot)$ defined on $[0, t_f]$ such that the associated trajectory, solution to the Cauchy problem

$$\begin{cases} \dot{x}(t) = f(x(t), u(t)), \\ x(0) = x_0, \end{cases}$$

is well-defined on $[0, t_f]$ and satisfies $x(t_f) = x_f$.

To do so, we follow the presentation made in [BFT06], where controllability is shown for an attitude control system equiped with opposite gaz jets. Mathematically, it means that the control can take the values $\{-1, 0, 1\}$.

Recall that in the setting of this thesis, the controls in the attitude equations (3.3.1) can only take the values $\{0, 1\}$. However, the results presented in this chapter will still be sufficient to conclude to the controllability of this system, as some thrusters are placed in order to produce opposite torques. On Figure 1.1, we give again a representation of the "SCA", drawing in red the pairs of thrusters that yield a control taking its values in the set $\{-1, 0, 1\}$.




(a) View from above (b) Side view

Figure 1.1 – Scheme of the SCA. Some pairs of thrusters create opposite torques.

Note also that we are going to show the controllability of the system (3.3.1) while considering the attitude of the launcher as an element $R(t) \in SO_3(\mathbb{R})$, and the dynamics as a differential system on the submanifold $SO_3(\mathbb{R}) \times \mathbb{R}^3$ (of dimension 6) of $\mathbb{R}^{12}$. In that case, the differential equations describing the evolution of the state $y(t) = (R(t), \overrightarrow{\omega}(t)) \in \mathbb{R}^{12}$ is

$$\left\{ \begin{array}{rcl} \dot{R}(t) & = & S(\overrightarrow{\omega}(t))R(t), \\ I_x\dot{p}(t) & = & (I_y - I_z)qr + \sum_{j=1}^m u_j(t)b_j^1, \\ I_y\dot{q}(t) & = & (I_z - I_x)p(t)r(t) + \sum_{j=1}^m u_j(t)b_j^2, \\ I_z\dot{r}(t) & = & (I_x - I_y)p(t)q(t) + \sum_{j=1}^m u_j(t)b_j^3, \end{array} \right. \tag{1.1}$$

where

$$S(\overrightarrow{\omega}) = \begin{pmatrix} 0 & r & -q \\ -r & 0 & p \\ q & -p & 0 \end{pmatrix}.$$

(In this chapter, we will denote the state of the system with the letter $y$, in order to emphasize the fact that it belongs to the space $SO_3(\mathbb{R}) \times \mathbb{R}^3$, and is therefore different from the state $x = (\theta, \psi, \varphi, p, q, r) \in \mathbb{R}^6$.)

Note that this system is a control-affine system that can be written under the form

$$\dot{y}(t) = f_0(y(t)) + \sum_{j=1}^m u_j(t) f_j(y(t)). \tag{1.2}$$

The vector field $f_0$ gives the dynamics for the free (uncontrolled) system $\dot{y}(t) = f_0(y(t))$, and for each $j \in [\![1, m]\!]$, the vector field $f_j$ is constant, equal to $(0_{\mathbb{R}^9}, \vec{b}_j)$.

The proof for the controllability of this system is based on the combination of two elements:

- The Poisson stability of the vector field $f_0$, corresponding to the uncontrolled dynamics. This stablity means that for almost every initial condition, the free system will come back arbitrarily close to the initial condition, in a time arbitrarily long.

- A rank condition on the Lie algebra spaned by the vector fields $(f_0, f_1, \ldots, f_m)$.

## 1.1 Poisson stability of a vector field

Let us start by defining the Poisson stability for a vector field, for which we make the assumption that the associated trajectories, solutions to the differential equation $\dot{y}(t) = X(y(t))$ are well-defined on $\mathbb{R}$.

**Definition 1.1 (Poisson stability for a vector field).** – *Let $X(\cdot)$ be a vector field. We say that $X$ is Poisson stable if for almost every initial condition $y_0$, every neighborhood $V$ of $y_0$ and every time $T > 0$, there exist times $t_1, t_2 \geqslant T$ such that $y(t_1, y_0) \in V$ and $y(-t_2, y_0) \in V$.*

In the previous definition, we denoted $y(t, y_0)$ the solution at time $t$ to the Cauchy problem:

$$\left\{ \begin{array}{rcl} \dot{y}(t) & = & X(y(t)), \\ y(0) & = & y_0. \end{array} \right.$$

This notion was introduced by H. Poincaré, following a work by S. D. Poisson, in his paper *Sur le problème des trois corps et les équations de la dynamique* [Poi90], where he undertook a study of the trajectories of the planets in the solar system.

The ingredients to have a Poisson stable vector field are the following:

- A finite measure $\mu$ on the phase space $Y$.
- A flow $\varphi$ that preserves the measure $\mu$, that is, for every $A \in Y$, $\mu(\varphi(A)) = \mu(A)$.

For the free part of the attitude equations

$$\left\{ \begin{array}{rcl} \dot{R}(t) & = & S(\overrightarrow{\omega}(t))R(t), \\ I_x \dot{p}(t) & = & (I_y - I_z)q(t)r(t), \\ I_y \dot{q}(t) & = & (I_z - I_x)p(t)r(t), \\ I_z \dot{r}(t) & = & (I_x - I_y)p(t)q(t), \end{array} \right. \tag{1.3}$$

that we write under the condensed form

$$\dot{y}(t) = f_0(y(t)),$$

The kinetic energy $I = (I_x p^2 + I_y q^2 + I_z r^2)/2$ remains constant over time. Indeed,

$$\begin{aligned} \dot{I}(t) &= I_x p(t)\dot{p}(t) + I_y q(t)\dot{q}(t) + I_z r(t)\dot{r}(t) \\ &= (I_y - I_z)p(t)q(t)r(t) + (I_z - I_x)p(t)q(t)r(t) + (I_x - I_y)p(t)q(t)r(t) \\ &= 0, \end{aligned}$$

Besides, the term $R(t)$ remains bounded as well. Let us define

$$J(t) := \mathrm{Tr}(R(t)R(t)^T) = \|R(t)\|_F,$$

where $\|\cdot\|_F$ denotes the Frobenius norm of the matrix $R(t)$. It holds

$$\begin{aligned}\dot{J}(t) &= \mathrm{Tr}(\dot{R}(t)R(t)^T + R(t)\dot{R}(t)^T)\\ &= \mathrm{Tr}(S(\overrightarrow{\omega}(t))R(t)R(t)^T + R(t)R(t)^T S(\overrightarrow{\omega}(t))^T)\\ &= \mathrm{Tr}\left(((S(\overrightarrow{\omega}(t)) + S(\overrightarrow{\omega}(t))^T)R(t)R(t)^T\right)\\ &= 0,\end{aligned}$$

as the matrix $S(\omega)$ is skew-symmetric, and the Frobenius norm of the matrix $R(t)$ is constant.

Therefore, the trajectories of the differential equations (1.3) remain bounded over time.

We now show that the flow associated to the dynamics (1.3) preserves the Lebesgue measure.

**Definition 1.2 (Flow of a vector field).** – *Let $X$ be a vector field. Under regularity assumptions on $X$, for all initial condition $y_0 \in \mathbb{R}^n$ there exist a unique solution $y(t, y_0)$ to the Cauchy problem*

$$\left\{\begin{array}{rcl}\dot{y}(t) &=& X(y(t)),\\ y(0) &=& y_0,\end{array}\right.$$

*that we denote* $\exp(tX)(y_0) := y(t, y_0)$.

In order to show that the flow preserves the Lebesgue measure, we are going to use a more general result, stating that the flow associated to a differential system $\dot{y}(t) = X(y(t))$ preserves this measure as soon as the divergence of the vector field $X$ is zero. In the literature, this result is known as *Liouville's Theorem*. We state this theorem now and give the proof in Appendix B.

**Proposition 1.1 (Liouville's Theorem).** – *Let* $\exp(tX)$ *be the flow of a non-linear differential equation* $\dot{y}(t) = X(y(t))$ *such that the divergence of the vector field $X$ is zero:*

$$\nabla \cdot X(y) = \mathrm{Tr}(dX(y)) = 0.$$

*Then the flow preserves the Lebesgue measure.*

It is then easy to check that the vector field $f_0$ corresponding to the uncontrolled dynamics (1.3) has zero divergence. Indeed, in the expression of the differential $df_0(y)$, all the diagonal coefficients are zero. It follows that the flow associated to the free part of the attitude equations preserves the Lebesgue measure.

We are now set to show the Poisson stability for the vector field $f_0$, which is sometimes stated as *Poincaré's recurrence Theorem*.

**Theorem 1.1 (Poincaré's recurrence theorem).** – *The vector field $f_0$ in the system (1.2) is Poisson stable.*

*Proof.* Let $\exp(tf_0)$ be the flow for the uncontrolled equations (1.3). We already know that this flow preserves the Lebesgue measure, as it has zero divergence. Let $A$ be a connected and bounded open set in $\mathbb{R}^{12}$. As the trajectories are bounded, there exists a compact set $Y$ such that every trajectory starting from $A$ (in positive or in negative time) remains in $Y$.

As $Y$ is compact, his Lebesgue measure is finite:

$$|Y| < +\infty.$$

For $p \in \mathbb{N}$, let $U_p := \cup_{k=p}^{+\infty} \exp(-kf_0)(A)$. Then, as $U_p$ is a subset of $Y$, his Lebesgue measure is also finite:

$$|U_p| < +\infty.$$

Besides, we have that $U_p \subset U_0 = A \cup (\cup_{k=1}^{+\infty} \exp(-kf_0)(A))$. But il also stands true that $U_p = \exp(-pf_0)(U_0)$ and the preservation of the Lebegue measure by the flow $\exp(-pf_0)$ yields

$$|U_p| = |U_0|\,.$$

From this, we deduce that

$$|U_p \backslash U_0| = 0,$$

that is,

$$|\{y \in U_0, y \notin U_p\}| = 0.$$

As $A$ is a subset of $U_0$, it follows that

$$\begin{aligned} |\{y \in A, y \notin U_p\}| &= 0, \\ |\{y \in A, \forall k \geqslant p, y \notin \exp(-kf_0)(A)\}| &= 0, \\ |\{y \in A, \forall k \geqslant p, \exp(kf_0)(y) \notin A\}| &= 0. \end{aligned}$$

Taking the countable reunion of those sets for $p \in \mathbb{N}$, the measure remains zero. Thus, we have shown that for almost every point $q \in A$ and for all $p \in \mathbb{N}$, there exists an integer $k_1 \geqslant p$ such that $\exp(k_1 f_0)(y) \in A$. With the same reasonning, for almost every $y \in A$ we can construct an integer $k_2 \geqslant p$ such that $\exp(-k_2 f_0)(y) \in A$: it is exactely the Poisson stability of the vector field $f_0$. □

## 1.2 Lie algebra spanned by vector fields and controllability

In this section, we are going to detail geometric conditions on the vector fields $f_0$, $f_1, \ldots, f_m$, that are necessary and sufficient to conclude to the controllability of a control-affine system under the form (1.2), as in [BFT06]:

$$\begin{aligned} \dot{y}(t) &= f_0(y(t)) + \sum_{j=1}^{m} u_j(t) f_j(y(t)) \\ &= f(y(t), u(t)), \end{aligned}$$

on a connected submanifold $\mathcal{M}$.

Let us start by introducing some definitions, that will be useful to study the controllability of the system $\dot{y}(t) = f(y(t), u(t))$. We define the set of vector fields $D$ by

$$D := \{f(\cdot, u) \mid \forall i \in [\![1, m]\!], u_i \in \{-1, 0, 1\}\}\,. \tag{1.4}$$

We also define

$$S(D) = \{\exp t_1 X_1 \circ \cdots \circ \exp t_k X_k \mid k \in \mathbb{N}, t_i \geqslant 0, X_i \in D\}\,. \tag{1.5}$$

With this definition, the set of reachable points starting from $y_0$ is $S(D)(y_0)$. The system will be controllable if for all $y_0 \in \mathcal{M}$, $S(D)(y_0) = \mathcal{M}$.

Intuitively, in order to have results for global controllability, it is necessary to be able to move in every direction of the tangent space of the submanifold $\mathcal{M}$. In the case of the attitude control problem, it means being able to move in every direction of the tangent space of $SO_3(\mathbb{R}) \times \mathbb{R}^3$.

Of course, it is possible to move in the directions given by

$$f_0 + \sum_{j=1}^{m} u_j f_j,$$

where $u_j \in \{-1, 0, 1\}$. However, we will now see that other directions are also available, combining properly the previous vector fields.

### 1.2.1 Lie bracket and Lie algebra

Given two vector fields $X$ and $Y$, the main notion to describe the directions available when moving along the vector fields $X$ and $Y$ is the Lie bracket of $X$ and $Y$. We now give a definition and a property of this object.

**Proposition 1.2 (Lie bracket).** – *The Lie bracket for the vector fields $X$ and $Y$ is the vector field, denoted by $[X, Y]$, such that*

$$e^{-tX} \circ e^{-tY} \circ e^{tX} \circ e^{tY}(y_0) =_{t\to 0} y_0 + t^2[X, Y](y_0) + o(t^2).$$

*Besides, we have*

$$[X, Y](y_0) = dY(y_0) \cdot X(y_0) - dX(y_0) \cdot Y(y_0).$$

If follows from this definition that $[X, Y] = 0$ if the vector fields $X$ and $Y$ commute locally. In the definition, it appears clearly that one needs to be able to move in the direction $X$ and $-X$ (and $Y$ and $-Y$) in order to be able to move in the direction of the Lie bracket $[X, Y]$, as shown on Figure 1.2.

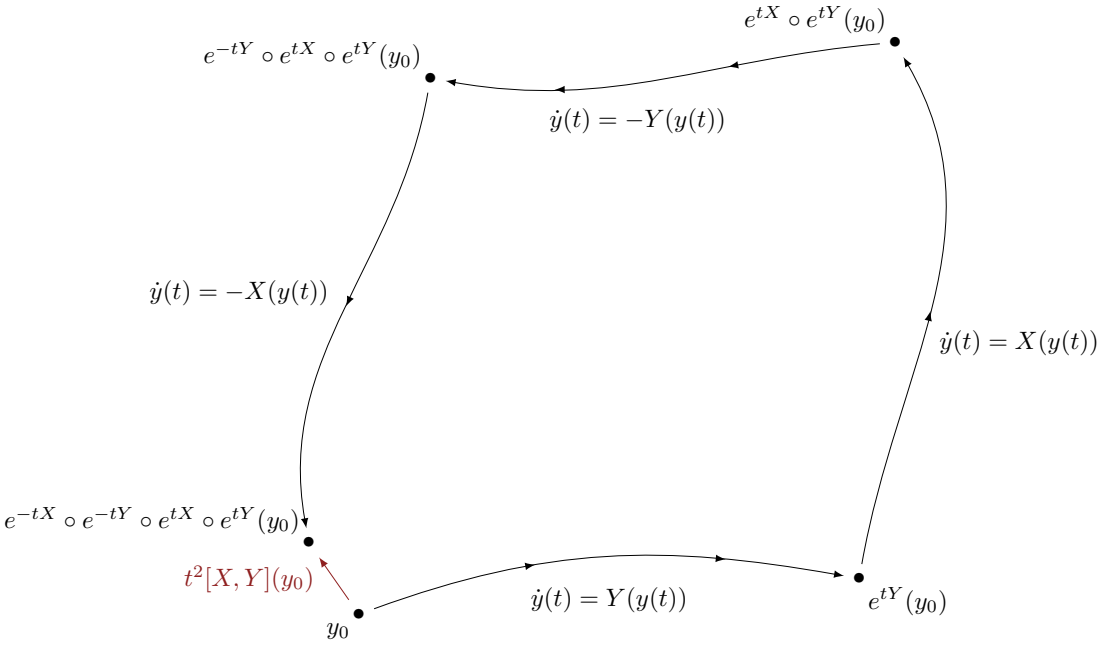



Figure 1.2 – Crochet de Lie $[X, Y]$ de deux champs de vecteurs $X$ et $Y$.

Given the differential system $\dot{y}(t) = f(y(t), u(t))$, we recall the definition 1.4 of the set of vector fields $D$,

$$D = \{f(\cdot, u) \mid \forall i \in [\![1, m]\!], u_i \in \{-1, 0, 1\}\},$$

which corresponds to the set of vector fields one gets by applying constant controls. We can then define the *Lie algebra* spanned by $D$.

**Definition 1.3 (Lie algebra).** – *The Lie algebra spanned by $D$, denoted by* $\mathrm{Lie}(D)$*, is the set of vector fields such that*

- *For all $X \in D$, $X \in \mathrm{Lie}(D)$.*
- *For all $X, Y \in \mathrm{Lie}(D)$, $[X, Y] \in \mathrm{Lie}(D)$.*

We are now able to give a first controllability result, as soon as the Lie algebra spanned by $D$ is of maximal dimension, that is, for all $y_0 \in \mathcal{M}$, $\mathrm{Lie}(D)(y_0) = T_{y_0}\mathcal{M}$, and $D$ is symmetric, that is for all $X \in D$, $-X \in D$.

**Theorem 1.2 (Symmetric case).** – *Assume that the Lie algebra spanned by $D$ is of maximal dimension, and that $D$ is symmetric. Then the system is controllable.*

*Proof.* The proof of this result is easy when $\mathcal{M} = \mathbb{R}^3$ and $D$ contains two vector fields $X$ and $Y$, and can give a good insight to the proof in the general case. Assume that the vector fields $X$, $Y$ and $[X, Y]$ are linearly independent at every point $q \in \mathbb{R}^3$. Let $\lambda \in \mathbb{R}$, $y_0 \in \mathbb{R}^3$ and $\varphi$ be the application:

$$\varphi : (t_1, t_2, t_3) \mapsto \exp(\lambda X) \circ \exp(t_3 Y) \circ \exp(-\lambda X) \circ \exp(t_2 Y) \circ \exp(t_1 X)(y_0).$$

Then, using the Baker-Campbell-Haussdorff formula, we get that

$$\varphi(t_1, t_2, t_3) = \exp(t_1 X + (t_2 - t_3)Y + \lambda t_3 [X, Y] + \ldots)(y_0)$$

Thus,

$$\left.\frac{\partial \varphi}{\partial t_1}\right|_{t_1=0} = X(y_0), \quad \left.\frac{\partial \varphi}{\partial t_2}\right|_{t_2=0} = Y(y_0), \quad \left.\frac{\partial \varphi}{\partial t_3}\right|_{t_3=0} = \lambda [X, Y](y_0) - Y(y_0) + o(\lambda),$$

and when $\lambda$ is small enough, those three vectors are linearly independant and the differential $d\varphi(t_1, t_2, t_3)$ is invertible. From this, we deduce that $S(D)(y_0)$ is a neighborhood of $y_0$. Besides, as we will soon see it the following, the set $S(D)(y_0)$ is closed, and as $\mathcal{M} = \mathbb{R}^3$ is connected, we get that $S(D)(y_0) = \mathbb{R}^3$.

This proof can be generalized if $\mathcal{M}$ is any connected submanifold. □

Let us end this section by giving a local result on the set of reachable points from $y_0$, which does not assume a symmetry hypothesis on the set $D$. The proof of this result can be found in [BFT06].

**Proposition 1.3.** – *Assume that the Lie algebra spanned by $D$ is of maximal rank. Then, for every $y_0 \in \mathcal{M}$ and all neighborhood $V$ of $y_0$, there exists an non-empty open set $U^+$ contained in $V \cap S(D)(y_0)$.*

Note that in the previous proposition, $S(D)(y_0)$ stands for the reachable points from $y_0$, in positive time. We could give a similar result replacing this set by the set of reachable points in negative time $S^-(D)(y_0)$. Then for every neighborhood $V$ of $y_0$, there exists a non empty open set $U^-$ contained in $V \cap S^-(D)(y_0)$.

### 1.2.2 Results in the non-symmetric case

The controllability result in Proposition 1.2 assumes a strong symmetry hypothesis on the system, namely that if a vector $X$ belongs in $D$, then $-X$ also belongs to $D$. This hypothesis is not satisfied in the case of a control-affine system with drift

$$\dot{y}(t) = f_0(y(t)) + \sum_{j=1}^{m} u_j(t) f_j(y(t)),$$

because of the drift term $f_0$ corresponding to the uncontrolled movement. Indeed, the vector fields $f_0+f_1$ or $f_0-f_1$ are in the set $D$, but the vector field $-f_0-f_1$ is not in $D$. The controllability of the system can however be obtained thanks to the Poisson stability of the vector field $f_0$.

#### Enlargement of the set $D$

Let us start by giving an important corollary of the Proposition 1.3: the system is controllable if and only if $\overline{S(D)(y_0)} = \mathcal{M}$ for every $y_0 \in \mathcal{M}$. Indeed, saying that the system is controllable means that $S(D)(y_0) = \mathcal{M}$, and we then also have that $\overline{S(D)(y_0)} = \mathcal{M}$. Reciprocally, assume that $\overline{S(D)(y_0)} = \mathcal{M}$. Let $y \in \mathcal{M}$. Following Proposition 1.3, for every neighborhood $V$ of $q$, there exists a non-empty open set $U^-$ such that $U^- \subset V \cap S^-(D)(y)$. As $\overline{S(D)(y_0)} = \mathcal{M}$, we get that the intersection $S(D)(y_0) \cap U^-$ is not empty: there exists a point $y_1$ such that $y_1 \in S(D)(y_0) \cap U^-$. Thus $y_1$ is reachable in positive time from $y_0$, and in negative time from $y$, i.e., $y$ is reachable is positive time from $y_1$. We deduce that $y$ is reachable in positive time from $y_0$, that is, $S(D)(y_0) = \mathcal{M}$ and the system is controllable.

With that in mind, we denote $\overline{D}$ the bigger (in the sense of the inclusion) set of vector fields such that

$$\overline{S(D)(y_0)} = \overline{S(\overline{D})(y_0)}.$$

It consists in the reunion of all the sets of vector fields $D'$ such that $\overline{S(D)(y_0)} = \overline{S(D')(y_0)}$. The following result shows how the Poisson stability compensates for the lack of symmetry in the family $D$.

**Proposition 1.4.** – *(i) If $X \in D$ and $X$ is Poisson stable, then $-X \in \overline{D}$.*
*(ii) If $X, Y \in D$, then $X + Y \in \overline{D}$.*

*Proof.* Let us start by proving $(i)$. Let $X$ be a vector field that is Poisson stable. We wish to show that

$$\overline{S(D \cup \{-X\})(y_0)} = \overline{S(D)(y_0)}.$$

The sense $\supset$ is clear, and we have to show that $\overline{S(D \cup \{-X\})(y_0)} \subset \overline{S(D)(y_0)}$. It is enough to show that $S(D \cup \{-X\})(y_0) \subset \overline{S(D)(y_0)}$. Let $y \in S(D \cup \{-X\})(y_0)$. First, we assume that $y$ is obtained from $y_0$ as

$$y = \exp(-tX)y_0,$$

with $t \geqslant 0$. Let $V$ be a neighborhood of $y$. As $X$ is Poisson stable, there exists a time $T > t$ such that $\exp(TX)y \in V$. It follows that

$$\exp(TX) \circ \exp(-tX)y_0 \in V$$
$$\exp((T-t)X)y_0 \in V.$$

It means that $y \in \overline{S(D)(y_0)}$. The general case, when there exist an integer $k \in \mathbb{N}$, non-negative

times $t_i \geqslant 0$ and vector fields $X_i \in D \cup \{-X\}$ such that

$$y = \exp(t_1 X_1) \circ \cdots \exp(t_k X_k) y_0$$

can be obtained by applying the same reasoning to each piece of the trajectory where the vector field $X_i$ is equal to $-X$.

Finally, to show $(ii)$, we use the Baker-Campbell-Haussdorff that leads to

$$\exp(\frac{t}{n}X) \circ \exp(\frac{t}{n}Y) = \exp(\frac{t}{n}(X+Y) + O(\frac{1}{n^2})).$$

Composing this relation $n$ times, it follows that

$$\prod_{1 \leqslant i \leqslant n} \exp(\frac{t}{n}X) \circ \exp(\frac{t}{n}Y) = \exp(t(X+Y) + O(\frac{1}{n})).$$

Letting $n \to +\infty$, we get that $X + Y \in \overline{D}$. □

### Controllability of a control-affine system

**Theorem 1.3.** – *On the submanifold $\mathcal{M}$, let us consider the control-affine system*

$$\dot{y}(t) = f_0(y(t)) + \sum_{j=1}^{m} u_j(t) f_j(y(t)).$$

*Assume that the vector field $f_0$ is Poisson stable, ant that the Lie algebra* $\mathrm{Lie}(f_0, f_1, \ldots, f_m)$ *is of maximal rank. Then the system is controllable*

*Proof.* The vector field $f_0$ is Poisson stable and belongs to the set $D$, therefore, according to Proposition 1.4, $-f_0 \in \overline{D}$. Besides, for all $j \in [\![1, m]\!]$, $f_0 \pm f_j \in D$ thus, also according to Proposition 1.4,

$$f_0 \pm f_j + (-f_0) \in \overline{D},$$

i.e., $\pm f_j \in \overline{D}$. It follows that $\overline{D}$ contains the set of vector fields $\{\pm f_0, \pm f_1, \ldots, \pm f_m\}$, which is symmetric and satisfies the rank condition : $\mathrm{Lie}(\pm f_0, \pm f_1, \ldots, \pm f_m)$ is of maximal rank. According to Theorem 1.2, we get that the control-affine system is controllable. □

**Remark 1.1: Analytic case**

In the case when the system is analytic, this condition is also necessary accroding to Sussmann's theorem [SJ72]. We will use this fact in the next section.

## 1.3 Controllability of the attitude of a rigid body

In this section, we are going to apply the result of Theorem 1.3 to the attitude control system equiped with opposite gaz jets (1.1)

$$\left\{\begin{array}{rcl} \dot{R}(t) & = & S(\overrightarrow{\omega}(t))R(t), \\ I_x\dot{p}(t) & = & (I_y - I_z)q(t)r(t) + \sum_{j=1}^m u_j(t)b_j^1, \\ I_y\dot{q}(t) & = & (I_z - I_x)p(t)r(t) + \sum_{j=1}^m u_j(t)b_j^2, \\ I_z\dot{r}(t) & = & (I_x - I_y)p(t)q(t) + \sum_{j=1}^m u_j(t)b_j^3. \end{array}\right.$$

We are going to show that, under some geometric conditions on the placement of the thrusters, the system can be controlled with only one thruster producing a torque $\vec{b}$. Therefore, we start by considering the case $m = 1$:

$$\dot{y}(t) = f_0(y(t)) + u(t)f_1(y(t)),$$

with $y = (R, \overrightarrow{\omega})$,

$$f_0 = (S(\overrightarrow{\omega})R, \frac{I_y - I_z}{I_x}qr, \frac{I_z - I_x}{I_y}pr, \frac{I_x - I_y}{I_z}pq),$$

and $f_1$ is the constant vector field $(0_{\mathbb{R}^9}, \vec{b})$, where we still denote $\vec{b}$ the normalized torque $I^{-1}\vec{b}$.

We have shown in Theorem 1.1 that the vector field $f_0$ is Poisson stable. Thus, following Theorem 1.3, the system is controllable if and only if the Lie algebra $\mathrm{Lie}(f_0, f_1)$ is of dimension 6 at every point $y \in SO_3(\mathbb{R}) \times \mathbb{R}^3$. For this condition to hold, it is necessary that the Lie algebra spanned by the vector fields $((I_y - I_z)/I_x qr, (I_z - I_x)/I_y pr, (I_x - I_y)/I_z pq)$ and $\vec{b}$ is of dimension 3 at every point of $\mathbb{R}^3$.

Let us set

$$a_1 = \frac{I_y - I_z}{I_x}\ ,\ a_2 = \frac{I_z - I_x}{I_y}\ ,\ a_3 = \frac{I_x - I_y}{I_z},$$

and let us define the vector field $Q = (a_1qr, a_2pr, a_3pq)$. Note that when chosing the order of the axis, without loss of generality, one may always chose to have $I_x \geqslant I_y \geqslant I_z$. In that case, it follows that $a_1, a_3 \geqslant 0$ and $a_2 \leqslant 0$. In what follows, we are actually going to assume that $I_x > I_y > I_z$, which yields $a_1, a_3 > 0$ and $a_2 < 0$. At the end of the chapter, we will make a remark on what may happen when some of the inertia coefficients are identical. Geometrically, it corresponds to the case of a body with symmetry properties.

### 1.3.1 Dimension of $\mathrm{Lie}(Q, \vec{b})$

Let us point out first that $Q(0) = 0_{\mathbb{R}^3}$ and $dQ(0) = 0_{\mathcal{M}_3(\mathbb{R})}$, and that each component of $Q$ is a polynomial of degree 2. Thus, the Lie algebra $\mathrm{Lie}(Q, \vec{b})$ is of dimension 3 if and only if the Lie algebra spanned by the constant vector fields is itself of dimension 3. With a formal calculation software, like Maple or Mathematica, we can compute the constant vector fields:

$$g_1 := \vec{b} = (b^1, b^2, b^3), \tag{1.6}$$

$$g_2 := [[Q, g_1], g_1] = \left(2a_1b^2b^3, 2a_2b^1b^3, 2a_3b^1b^2\right), \tag{1.7}$$

$$\begin{aligned} g_3 &:= [[Q, g_1], g_2] \\ &= \left(2a_1b^1(a_2(b^3)^2 + a_3(b^2)^2), 2a_2b^2(a_1(b^3)^2 + a_3(b^1)^2), 2a_3b^3(a_1(b^2)^2 + a_2(b^1)^2)\right). \end{aligned} \tag{1.8}$$

and the next Lie brackets do not span new directions in $\mathbb{R}^3$:

$$\begin{aligned}[[Q,g_1],g_3] &= 2\begin{pmatrix} a_1b^3b^2\left(a_1a_2(b^3)^2+2a_2a_3(b^1)^2+a_1a_3(b^2)^2\right)\\ a_2b^1b^3\left(a_1a_2(b^3)^2+a_2a_3(b^1)^2+2a_1a_3(b^2)^2\right)\\ a_3b^1b^2\left(2a_1a_2(b^3)^2+a_2a_3(b^1)^2+a_1a_3(b^2)^2\right)\end{pmatrix},\\ &= \lambda g_3 + 2a_1a_2a_3b^1b^2b^3\vec{b},\\ &= \lambda g_3 + 2a_1a_2a_3b^1b^2b^3g_1,\end{aligned}$$

with $\lambda = a_1a_2(b^3)^2 + a_2a_3(b^1)^2 + a_1a_3(b^2)^2$.

$$[[Q,g_2],g_2] = 8a_1a_2a_3b^1b^2b^3g_1.$$

$$\begin{aligned}[[Q,g_2],g_3] &= 4a_1a_2a_3\begin{pmatrix} b^1\left((b^2)^2(a_1(b^3)^2+a_3(b^1)^2)+(b^3)^2(a_1(b^2)^2+a_2(b^1)^2)\right)\\ b^2\left((b^1)^2(a_2(b^3)^2+a_3(b^2)^2)+(b^3)^2(a_1(b^2)^2+a_2(b^1)^2)\right)\\ b^3\left((b^1)^2(a_2(b^3)^2+a_3(b^2)^2)+(b^2)^2(a_1(b^3)^2+a_3(b^1)^2)\right)\end{pmatrix},\\ &= \mu_1g_1+\mu_2g_2,\end{aligned}$$

with $\mu_1 = 4a_1a_2a_3(a_1(b^2b^3)^2 + a_2(b^1b^3)^2 + a_3(b^1b^2)^2)$ and $\mu_2 = 4a_1a_2a_3b^1b^2b^3$.

$$\begin{aligned}[[Q,g_3],g_3] &= 8a_1a_2a_3\begin{pmatrix} b^2b^3\left(a_1(b^3)^2+a_3(b^1)^2\right)\left(a_1(b^2)^2+a_2(b^1)^2\right)\\ b^1b^3\left(a_1(b^2)^2+a_2(b^1)^2\right)\left(a_2(b^3)^2+a_3(b^2)^2\right)\\ b^1b^2\left(a_1(b^3)^2+a_3(b^1)^2\right)\left(a_2(b^3)^2+a_3(b^2)^2\right)\end{pmatrix},\\ &= \mu_3g_1+\mu_4g_2+\mu_5g_3,\end{aligned}$$

with $\mu_3 = 8a_1a_2a_3b^1b^2b^3(a_1a_2(b^3)^2+a_2a_3(b^1)^2+a_1a_3(b^2)^2)$, $\mu_4 = 4a_1a_2a_3(a_1(b^2b^3)^2+a_2(b^1b^3)^2+a_3(b^1b^2)^2)$ and $\mu_5 = -4a_1a_2a_3b^1b^2b^3$.

Besides, thanks to Jacobi identity,

$$\begin{aligned}[[Q,g_2],g_1] &= -[[g_2,g_1],Q]-[[g_1,Q],g_2]\\ &= [[Q,g_1],g_2]\\ &= g_3,\\ [[Q,g_3],g_2] &= -[[g_3,g_2],Q]-[[g_2,Q],g_3]\\ &= [[Q,g_2],g_3].\end{aligned}$$

Those computations show that the Lie algebra $\mathrm{Lie}(Q,\vec{b})$ has the same dimension than the space spanned by the vectors $g_1$, $g_2$ and $g_3$. The determinant of those three vectors is

$$\det(g_1,g_2,g_3) = 4\left(a_3(b^1)^2-a_1(b^3)^2\right)\left(a_2^2(b^1)^2(b^3)^3-a_2a_3(b^2)^2(b_1)^2-a_1a_2(b^2)^2(b^3)^2+a_1a_3(b^2)^4\right).$$

Therefore, the determinant is equal to zero if and only if $a_3(b^1)^2 = a_1(b^3)^2$, or

$$a_2^2(b^1)^2(b^3)^3-a_2a_3(b^2)^2(b_1)^2-a_1a_2(b^2)^2(b^3)^2+a_1a_3(b^2)^4 = 0.$$

We have made the assumption that $a_1$ and $a_3$ are positive and $a_2$ is negative. Thus, this last quantity is equal to zero when

$$a_2^2(b^1)^2(b^3)^3 = a_2a_3(b^2)^2(b_1)^2 = a_1a_2(b^2)^2(b^3)^2 = a_1a_3(b^2)^4 = 0,$$

i.e.,

$$(b^1)^2(b^3)^3 = (b^2)^2(b_1)^2 = (b^2)^2(b^3)^2 = (b^2)^4 = 0,$$

that is when $b^2 = 0$, and one of the real numbers $b^1$ and $b^3$ is zero.

We have therefore shown that the vectors $g_1$, $g_2$ and $g_3$ are linearly independant unless $a_3(b^1)^2 = a_1(b^3)^2$, or $b^2 = 0$ and one of the real numbers $b^1$ and $b^3$ is zero. We have found the condition for the Lie algebra $\mathrm{Lie}(Q, \vec{b})$ to be of dimension 3 at every point of $\mathbb{R}^3$, that we give in the following lemma.

**Lemma 1.1.** – *The Lie algebra* $\mathrm{Lie}(Q, \vec{b})$ *is of dimension 3 at every point of* $\mathbb{R}^3$ *unless* $\sqrt{a_1}b^3 = \pm\sqrt{a_3}b^1$, *or* $b^2 = 0$ *and one of the real numbers* $b^1$ *et* $b^3$ *is zero.*

### 1.3.2 Dimension de $\mathrm{Lie}(f_0, f_1)$

Let us now study the dimension of the Lie algebra $\mathrm{Lie}(f_0, f_1)$, under the geometric conditions of Lemma 1.1, that is when the vectors $g_1$, $g_2$ and $g_3$ previously introduced are linearly independant.

A first computation yields the expression of the following constant vector fields:

$$f_1 = (0_{\mathbb{R}^9}, g_1)\ ,\ [[f_0, f_1], f_1] = (0_{\mathbb{R}^9}, g_2)\ ,\ [[f_0, f_1], [[f_0, f_1], f_1]] = (0_{\mathbb{R}^9}, g_3).$$

Thus, as soons as the vectors $g_1$, $g_2$ and $g_3$ are linearly independant, the constant vector fields $f_1$, $[[f_0, f_1], f_1]$ and $[[f_0, f_1], [[f_0, f_1], f_1]]$ are also linearly independant, and the space they span matches $\mathrm{Vect}((0_{\mathbb{R}^9}, E_1), (0_{\mathbb{R}^9}, E_2), (0_{\mathbb{R}^9}, E_3))$ where $(E_1, E_2, E_3)$ stands for the canonical basis of $\mathbb{R}^3$.

Let us then compute the following vector fields, denoting by $(r_{ij})_{1 \leqslant i,j \leqslant 3}$ the components of the orientation matrix $R$.

$$[f_0, (0_{\mathbb{R}^9}, E_1)] = -\begin{pmatrix} 0 \\ 0 \\ 0 \\ r_{31} \\ r_{32} \\ r_{33} \\ -r_{21} \\ -r_{22} \\ -r_{23} \\ 0 \\ a_2 r \\ a_3 q \end{pmatrix}\ ,\ [f_0, (0_{\mathbb{R}^9}, E_2)] = -\begin{pmatrix} -r_{31} \\ -r_{32} \\ -r_{33} \\ 0 \\ 0 \\ 0 \\ r_{11} \\ r_{12} \\ r_{13} \\ a_1 r \\ 0 \\ a_3 p \end{pmatrix}\ ,\ [f_0, (0_{\mathbb{R}^9}, E_3)] = -\begin{pmatrix} r_{21} \\ r_{22} \\ r_{23} \\ -r_{11} \\ -r_{12} \\ -r_{13} \\ 0 \\ 0 \\ 0 \\ a_1 q \\ a_2 p \\ 0 \end{pmatrix}.$$

Those vector fields are linearly independant at every point of $SO_3(\mathbb{R}) \times \mathbb{R}^3$ if and only if the

vector fields

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ r_{31} \\ r_{32} \\ r_{33} \\ -r_{21} \\ -r_{22} \\ -r_{23} \end{pmatrix}, \quad \begin{pmatrix} -r_{31} \\ -r_{32} \\ -r_{33} \\ 0 \\ 0 \\ 0 \\ r_{11} \\ r_{12} \\ r_{13} \end{pmatrix}, \quad \begin{pmatrix} r_{21} \\ r_{22} \\ r_{23} \\ -r_{11} \\ -r_{12} \\ -r_{13} \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

are linearly independant at every point of $SO_3(\mathbb{R})$, which is indeed the case.

We deduce from the following computations that under the conditions of Lemma 1.1, the vector fields $f_1$, $[[f_0, f_1], f_1]$, $[[f_0, f_1], [[f_0, f_1], f_1]]$, $[f_0, (0_{\mathbb{R}^9}, E_1)]$, $[f_0, (0_{\mathbb{R}^9}, E_2)]$ and $[f_0, (0_{\mathbb{R}^9}, E_3)]$ are linearly independant at each point of $SO_3(\mathbb{R})$. It follows that the Lie algebra $\mathrm{Lie}(f_0, f_1)$ is of dimension 6.

### 1.3.3 Controllability condition

The vector field $f_0$ is Poisson stable, and using Theorem 1.3 we can conclude to the controllability of the attitude equations by means of opposite gaz jets.

**Theorem 1.4 (Case $m = 1$).** – *The attitude equations are controllable by means of one pair of opposite gaz jets except when $\sqrt{a_1}b^3 = \pm\sqrt{a_3}b^1$, or $b^2 = 0$ and one of the real numbers $b^1$ and $b^3$ is zero.*

Let us now analyse what the condition of non-controllability in Theorem 1.4 means. If two of the real numbers $b^i$ (including $b^2$) are zero, for instance $b^1$ and $b^2$, the equations for the angular velocity become

$$\left\{ \begin{array}{rcl} \dot{p} & = & a_1 qr, \\ \dot{q} & = & a_2 pr, \\ \dot{r} & = & a_3 pq + b^3 u, \end{array} \right.$$

and the line $\{p = q = 0\}$ is invariant under the action of any control $u$.

If $\sqrt{a_1}b^3 = \pm\sqrt{a_3}b^1$, for instance $\sqrt{a_1}b^3 = \sqrt{a_3}b^1$, then the vector $\vec{b}$ belongs to the plane of equation $\{\sqrt{a_1}r = \sqrt{a_3}p\}$. It is then easy to check that if $\overrightarrow{\omega} = (p, q, r)$ belongs to this plane, then the vector $Q(\overrightarrow{\omega})$ also belongs to this plane, that is, this plane is invariant. In that case, whatever the control $u$ may be, the plane of equation $\left\{\sqrt{a_1}r = \sqrt{a_3}p\right\}$ is invariant.

We now have everything to state the controllability condition for the attitude equations in the general case, when $m \geqslant 1$.

**Theorem 1.5 (Case $m \geqslant 1$).** – *The attitude equations are controllable by means of opposite gaz jets producing the torques $\{\overrightarrow{b}_1, \ldots, \overrightarrow{b}_m\}$ unless the space spanned by the vectors $(\overrightarrow{b}_i)_{1 \leqslant i \leqslant m}$ matches one of the invariant line of equation $\{p = q = 0\}$, $\{p = r = 0\}$ or $\{q = r = 0\}$, or matches one of invariant the plane of equation $\left\{\sqrt{a_1}r = \pm\sqrt{a_3}p\right\}$.*

**Remark 1.2:**

Implicitely, when making the assumption $I_1 > I_2 > I_3$, we used several times in the proof of Theorem 1.4 the coupling between the equations for the angular velocity. Indeed, it implies that $a_1, a_3 > 0$ and $a_2 < 0$, and it enabled us to give a simple condition for the vectors $g_1$, $g_2$ and $g_3$ to be linearly independant. The fact that $a_1$, $a_2$ and $a_3$ are non-zero enables to use the coupling in

$$\begin{cases} \dot{p} &= a_1 qr, \\ \dot{y} &= a_2 pr, \\ \dot{r} &= a_3 pq. \end{cases}$$

If the three coefficients are zero (i.e., $I_x = I_y = I_z$), we easily get from the equations (1.6), (1.7) and (1.8) that only the vector $g_1$ is non zero. At least three torques (linearly independant) are thus required to control the attitude equations.
If one of the coefficient $a_i$ is zero, for instance $a_1 = 0$, $a_2 < 0$ and $a_3 > 0$, then

$$\det(g_1, g_2, g_3) = (b^1)^4 a_2 a_3 \left(a_2 (b^3)^2 - a_3 (b^2)^2\right).$$

As $a_3 > 0$ and $a_2 < 0$, this determinant is zero if and only if $b^1 = 0$ or $b^3 = b^2 = 0$. Thus, as soon as two of the real numbers $b^i$ (including $b^1$) are non zero, the attitude equations remain controllable by means of only one torque.

## 1.4 Conclusion of the chapter

In this chapter, we recalled the result of Theorem 1.5 which gives necessary and sufficient conditions to be able to control the attitude equations by means of pairs of opposite thrusters. Under some geometric hypothesis on the placement of the thrusters, the attitude can be controlled with only one pair of opposite gaz jets.

For the system presented in this thesis, the design of the SCA schematized on Figure 1.5 allows to consider some pairs of thrusters as opposite jets. The controllability of the equations presented in this thesis follows.

Nevertheless, the proof of the controllability is not constructive, and no effective control strategy has yet been exhibited. The proof uses the Poisson stability of some vector fields, which can yield very long transfer times. In the following chapter, in the setting of optimal control theory, we will study ways to numerically compute control strategies.

Chapter 2

# Optimal control in finite dimension

## Contents


**1.1 Poisson stability of a vector field . . . . . 19**

**1.2 Lie algebra spanned by vector fields and controllability . . . . . 21**

1.2.1 Lie bracket and Lie algebra . . . . . 22

1.2.2 Results in the non-symmetric case . . . . . 24

**1.3 Controllability of the attitude of a rigid body . . . . . 26**

1.3.1 Dimension of Lie$(Q, \vec{b})$ . . . . . 26

1.3.2 Dimension de Lie$(f_0, f_1)$ . . . . . 28

1.3.3 Controllability condition . . . . . 29

**1.4 Conclusion of the chapter . . . . . 30**


The purpose of this chapter is to give a brief insight of optimal control theory in finite dimension. The term "finite dimension" refers to the fact that the state vector $x(\cdot)$ belongs to the finite dimensional space $\mathbb{R}^n$. However, it is important to keep in mind that solving an optimal control problem in finite dimension like (**OCP**) requires being able to solve an infinite-dimensional optimization problem.

In chapter 1, we have shown that the attitude equations studied in this thesis are controllable. We are now interested in computing effectively control strategies, asking also for them to be optimal with respect to a given criterion.

We will first recall a classical result in optimal control theory, namely Pontryagin Maximum Principle, and we will then focus on numerical methods. At the end of the chapter, we will show how those methods can be implemented to solve numerically an attitude control problem corresponding to the separation of one satellite.

## 2.1 General setting

In this chapter, we wil consider the following general control problem: given $\mathcal{M}_0$ and $\mathcal{M}_1$ two submanifolds of $\mathbb{R}^n$ we aim at controlling the nonlinear system

$$\dot{x}(t) = f(t, x(t), u(t)) \quad \text{on } [0; t_f], \tag{2.1}$$

while minimizing the cost

$$C(t_f, u) = \int_0^{t_f} f^0(t, x(t), u(t))\, dt + g(t_f, x(t_f)), \tag{2.2}$$

and such that

$$x(0) \in \mathcal{M}_0, \quad x(t_f) \in \mathcal{M}_1.$$

In this description, $f$ is an application $\mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, $f^0 : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ and $g : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$. It is usual to assume the applications $f$, $f_0$ and $g$ to be of class $C^1$, even if this assupmtion can be weakened. The control $u(\cdot)$ belongs to the set $L^\infty([0; t_f], \Omega)$ where $\Omega$ is a convex subset of $\mathbb{R}^m$ and the final time $t_f$ can be left free or not. In the following we will use the notation $t_f$ when it is free, and $T$ when it is fixed.

We will say that a control $u(\cdot)$ defined on $[0; t_f]$ is *admissible* when the associated trajectory of the control system (2.1) is well-defined on $[0; t_f]$.

**Remark 2.1: Existence of an optimal solution**

It is far from obvious that there exists a solution to the previous optimal control problem. In the literature, the general existence results depend in particular on the compacity of the set $\Omega$ in which the control takes its values. In the case of the attitude control problem, the set $\Omega$ is compact. With that in mind, let us denote $\mathcal{U}$ the set of admissible controls $u \in L^\infty([0; t_f], \Omega)$ that steer the system from $\mathcal{M}_0$ to $\mathcal{M}_1$ in time $t(u)$. We assume that regularity assumptions on the applications $f$, $f_0$ and $g$ are fulfilled, and that the following assumptions hold:

(i) There exists $C_1 \geqslant 0$ such that for all $u \in \mathcal{U}$, $t(u) \leqslant C_1$,

(ii) There exists $C_2 \geqslant 0$ such that for all $u \in \mathcal{U}$, $\|x_u(\cdot)\|_\infty \leqslant C_2$,

(iii) For all $(t, x) \in \mathbb{R} \times \mathbb{R}^n$, the set

$$V(t, x) = \left\{ \begin{pmatrix} f(t, x, u) \\ f^0(t, x, u) + \gamma \end{pmatrix} \mid u \in \Omega, \gamma \geqslant 0 \right\}$$

is convex.

Then there exists a solution $u^\star$ defined on an interval $[0, t(u^\star)]$ to the optimal control problem. The proof of this result, as well as some extensions can be found in [Tré05a, LM67a, BC03a]. Hypotheses $(i) - (ii)$ ensure enough compacity to be able to extract converging (for the weak-star topology) subsequences in $L^\infty([0; t_f], \Omega)$. Hypothesis $(iii)$ allows to use the fact that closed (for the strong topology) convex sets are also weakly closed. Let us mention that this result can be generalized to the case of a control problem with state constraints.

## 2.2 Pontryagin Maximum Principle

In this section, we give a statement of Pontryagin Maximum Principle. It expresses necessary conditions for a pair $(x(\cdot), u(\cdot))$ to be optimal.

**Analogy with optimization in finite dimension.** It is crucial to keep in mind that the conditions in Pontryagin Maximum Principle are a set of necessary conditions: they consist in conditions of order 1. In a similar way, Lagrange Theorem, when applied to an optimization problem in finite dimension gives a set of necessary conditions. To be more specific, let us simplify the general optimal control stated in the previous section 2.1 before going further in the analogy. We assume that the submanifolds are singletons

$$\mathcal{M}_0 = \{x_0\} \ , \ \mathcal{M}_1 = \{x_1\},$$

and that the final time $T$ is fixed.

Given two applications in finite dimension $J : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}^k$, Lagrange Theorem states that if a point $x^\star$ is a solution to the finite-dimensional optimization problem

$$\min_{s.t.\ g(x)=0} J(x), \tag{2.3}$$

then there exists a pair $(\lambda^0, \lambda) \in \mathbb{R} \times \mathbb{R}^k$ of Lagrange multipliers such that

$$\lambda^0 \nabla J(x^\star) + \langle \lambda, g(x^\star) \rangle = 0.$$

We shall now introduce the definition of the end-point mapping, that we will use again further in the thesis (notably in chapter 4). It allows us to rewrite the optimal control under a form close to (2.3).

**Definition 2.1 (End-point mapping).** — *Let $u(\cdot) \in L^\infty([0; t_f], \Omega)$ be an admissible control. The end-point mapping is defined as the response of the system to the control $u$:*

$$E_{x_0,T}(u) = x_u(T),$$

*where $x_u(\cdot)$ is the solution to the ordinary differential equation (2.1) with initial condition $x_u(0) = x_0$.*

Therefore, solving the previous optimal control problem amounts to solving the optimization problem in infinite dimension

$$\min_{s.t.\ E_{x_0,T}(u)=x_1} C(T, u). \tag{2.4}$$

In a similar way to Lagrange Theorem in finite dimension, if a control $u^\star$ is optimal, then there exists a pair $(\psi^0, \psi) \in \mathbb{R} \times \mathbb{R}^m$ such that

$$\psi^0 \frac{\partial C}{\partial u}(T, u^\star) + \psi \cdot dE_{x_0,T}(u^\star) = 0.$$

Pontryagin Maximum Principle, that we we will present in details in the following paragraph, can be seen as a development of this last equality.

**Statement of the PMP.** Let us now give a precise statement of the PMP. A proof of this result can be found in [LM67b, PBGM62, BFT06, ST10a]. Before going forward, let us point

out that a key ingredient in the proof is the use of needle-like variations of the control, an idea that we will use again in Chapter 4.

**Definition 2.2 (Hamiltonian).** – *The Hamiltonian of the system is defined in the following way*

$$H = \left\{ \begin{array}{ccc} \mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^n\times\mathbb{R} & \longrightarrow & \mathbb{R} \\ (t,x,u,p,p^0) & \mapsto & \langle p, f(t,x,u)\rangle + p^0 f^0(t,x,u) \end{array} \right.$$

**Theorem 2.1 (Pontryagin maximum principle).** – *Let $(x(\cdot),u(\cdot))$ be a trajectory of the system (2.1), optimal with respect to the cost (2.2). Then there exists a non-trivial pair $(p(\cdot),p^0)$ such that :*

- $p^0 \leqslant 0$.
- *$p(\cdot)$ is absolutely continous on $[0;t_f]$.*
- *For almost every $t\in[0;t_f]$,*

$$\dot{x}(t) = \frac{\partial H}{\partial p}(t,x(t),u(t),p(t),p^0), \tag{2.5}$$

$$\dot{p}(t) = -\frac{\partial H}{\partial x}(t,x(t),u(t),p(t),p^0). \tag{2.6}$$

- *For almost every $t\in[0;t_f]$, the control $u(t)$ maximizes the Hamiltonian $H$ :*

$$H(t,x(t),u(t),p(t),p^0) = \max_{v\in\Omega} H(t,x(t),v,p(t),p^0). \tag{2.7}$$

- *The adjoint vector $p(\cdot)$ satisfies the transversality conditions.*

$$p(0) \perp T_{x(0)}\mathcal{M}_0, \tag{2.8}$$

$$p(t_f) - p^0\frac{\partial g}{\partial x}(t_f,x(t_f)) \perp T_{x(t_f)}\mathcal{M}_1, \tag{2.9}$$

  *where $T_x\mathcal{M}$ is the notation for the tangent space of the submanifold $\mathcal{M}$ at point $x$.*

- *If the final time $t_f$ is free, there is an additional transversality condition*

$$\max_{v\in\Omega} H(t_f,x(t_f),v,p(t_f),p^0) = -p^0\frac{\partial g}{\partial t}(t_f,x(t_f)). \tag{2.10}$$

**Definition 2.3 (Extremal).** – *We call extremal a tuple $(x(\cdot),u(\cdot),p(\cdot),p^0)$ that is solution to the equations (2.5), (2.6) and (2.7). If $p^0<0$, the extremal is said to be normal, and abnormal if $p^0=0$.*

Note that the couple $(p(\cdot),p^0)$ in the statement of Theorem 2.1 is defined up to a positive multiplicative constant. Therefore, if $p^0<0$, i.e., if the extremal is normal, one can always set $p^0=-1$.

**Remark 2.2: Sufficient conditions for optimality**

We wish to emphasize once more that the PMP gives a set of necessary conditions. If one wants to check the optimality status (at least locally) of a given extremal $(x(\cdot), u(\cdot), p(\cdot), p^0)$, the following condition, known as the strong Legendre condition, is sufficient : there exists $\alpha > 0$ such that for each $v \in \mathbb{R}^m$,

$$\frac{\partial^2 H}{\partial u^2}(t, x(t), u(t), p(t), p^0) \cdot (v, v) \leqslant -\alpha \|v\|^2 .$$

From a practical point of view, it can be satisfactory to find a trajectory satisfying the necessary conditions of the PMP. This is what we will do in the sequel.

There exists in the literature a wide variety of variations for the PMP, for more general control systems than the one in the statement of Theorem 2.1. Let us mention for instance the following generalizations:

- In [AS04], the author shows a maximum principle for a control system where both the state and the control belong to submanifolds, that is when the dynamics can be written under the form
$$\dot{x}(t) = f(x(t), u(t)),$$
where $f : \mathcal{M} \times \mathcal{N} \to T\mathcal{M}$, with $\mathcal{M}$ (resp. $\mathcal{N}$) a submanifold of $\mathbb{R}^n$ (resp. $\mathbb{R}^m$). In this setting, the adjoint vector $p(t)$ is an element of the dual space of $T_{x(t)}\mathcal{M}$, in order to give a meaning to the quantity $\langle p(t), f(x(t), u(t)) \rangle$. Note that in the case of the attitude control problem for a rigid body, the result may be of importance when considering the dynamics as a differential equation on $SO_3(\mathbb{R}) \times \mathbb{R}^3$, as we did in Chapter 1.
- The papers [DK08, DK11, GP05a] state maximum principles for *hybrid* control systems, where the dynamics can change over time. Inequality or equality constraints at times when the dynamics changes can also be dealt with in this setting.
- In [Cla90], Clarke considers state-constrained control problems, where the constraint on the state is written under the form
$$c(x) \leqslant 0 \quad \text{almost everywhere on } [0, t_f].$$
A major difficulty arises when solving such problems: the adjoint vector $p(\cdot)$ becomes a measure and is not anymore absolutely continuous. There are jumps in the evolution of the adjoint vector every time the state meets the fronteer of the allowed domain, $\{x \in \mathbb{R}^n \,|\, c(x) = 0\}$. Numerically, one way to tackle this issue may be to penalize the state constraint in the cost.
- Finally, let us mention the existence of maximum principles for control systems in infinite dimension. In [LLY95], a statement is given for an evolution equation
$$\dot{y}(t) = Ay(t) + f(t, y(t), u(t)),$$
where for each $t$, $y(t) \in X$ and $u(t) \in U$, with $X$ a Banach space and $U$ a separable metric space. In chapter 5, it is the setting we will consider to derive the structure of the controls for an evolution problem for populations of cells structured in phenotype.

In chapter 3, we will show two maximum principles for a control problem with via-point

constraints, as in [BH75]. It will consist in ponctual constraints under the form

$$c(x(t_1)) \leqslant 0,$$

with $t_1 \in [0; t_f]$. We will see that such constraints can be dealt with using the formalism of [DK08, DK11] for hybrid control systems.

## 2.3 Numerical methods in optimal control

Amongst the existing numerical methods to solve an optimal control problem, it is usual to make the distinction between *direct methods* and *indirect methods*. Indirect methods exploit the necessary conditions stated by the PMP to reduce the problem of finding an optimal trajectory to finding the zeros of some function in finite dimension. This is then often done by Newton-like methods. Direct methods, for their part, consist in discretizing totally the state and the control to end up with an optimization problem in finite dimension. Such a problem can then be solved by means of the usual optimization techniques.

We will now give more details on this two families of numerical methods, starting first with the direct methods and moving on then to the indirect ones. The survey paper [Tré12] gives a state of the art for numerical methods in optimal control, putting the emphasis on aerospace applications.

### 2.3.1 Direct methods

The main idea behind a direct method is to undertake a complete discretization of the optimization space: let $0 = t_0 < t_1 < \cdots < t_N = t_f$ be a subdivision of the time interval $[0; t_f]$ (for the sake of simplicity, we will consider here that the discretization is uniform. We denote $h := t_1 - t_0 = t_f/N$ the step of this subdivision. The dynamics is also discretized using some numerical scheme. In our case, we have considered an explicit/implicit scheme: for each $i \in [\![0, N-1]\!]$,

$$x(t_{i+i}) \approx x(t_i) + \frac{h}{2}\left(f(t_i, x(t_i), u(t_i)) + f(t_{i+1}, x(t_{i+1}), u(t_{i+1}))\right).$$

Let us denote $x_i \in \mathbb{R}^n$ (resp. $u_i \in \mathbb{R}^m$) an approximation of $x(t_i)$ (resp. $u(t_i)$), and let $x^h \in \mathbb{R}^{n\times(N+1)}$ be the vector $(x_0, x_1, \ldots, x_N)$, and $u^h \in \mathbb{R}^{m\times(N+1)}$ be the vector $(u_0, u_1, \ldots, u_N)$. We also consider some discretization of the integral cost (2.2), for instance thanks to the rectangle method:

$$C(t_f, u) \approx C^h(t_f, u^h) := h \cdot \sum_{i=0}^{N-1} f^0(t_i, x_i, u_i).$$

Numerically, solving the optimal control problem amounts to solving the *finite-dimensional* optimization problem:

$$\text{minimize} \quad C^h(t_f, u^h)$$

under the constraints

$$\begin{array}{ll} x_{i+i} = x_i + \frac{h}{2}\left(f(t_i, x_i, u_i) + f(t_{i+1}, x_{i+1}, u_{i+1})\right), & \forall i \in [\![0, N-1]\!], \\ u_i \in \Omega, & \forall i \in [\![0, N]\!], \\ x_0 \in \mathcal{M}_0, & \\ x_N \in \mathcal{M}_1. & \end{array}$$

**Numerical aspects.** Eventually, when performing a direct method, one is left with solving some optimization problem in finite dimension under the form

$$\begin{array}{cc} \min & f(X). \\ g(X) = 0 & \\ h(X) \leqslant 0 & \end{array}$$

(Assuming that the constraints on the control $u \in \Omega$ and on $x_0$ and $x_N$ can be written under the form of an equality or an inequality constraint.)

The literature is full of various numerical methods to tackle such a problem. To perform the numerical simulations in this thesis, we chose the open-source solver IPOPT [WB06a], based on the implementation of some interior-point algorithm.

Let us mention that the solver IPOPT can be used jointly with the modelling language AMPL [FGK93]. The interface provided by AMPL allows for a very easy implementation of the optimization problem, with an intuitive syntax. For instance, AMPL uses automatic differentiation to compute the derivatives of the constraints and of the cost function. In constrast, if one wishes to solve efficiently an optimal control problem using only the solver IPOPT (for instance through the C, C++ or Fortran interfaces), it is required to implement as well the methods computing the derivatives.

**Remark 2.3:**

When compared to the indirect methods we are going to present hereafter, direct methods offer the possibility to tackle, at a low computational cost, constraints on the state variable. This can be of importance in practice, for instance if the CNES wishes to forbid some angular domain during the whole ballistic phase, or control the transverse angular velocities, as we will show at the end of the chapter.

### 2.3.2 Indirect methods

In contrast with direct methods where a full discretization of the optimal control problem is undertaken first, indirect methods exploit the duality and the necessary conditions stated in the PMP.

We denote $z := (x, p)$ the pair formed by the state variable $x$ and the adjoint vector $p$. Under usual regularity assumptions, the maximization condition (2.7) allows to express the control $u$ as a function of $z$: $u = u(x, p)$, and the dynamics (2.5)-(2.6) can then be written under the closed form $\dot{z}(t) = F(t, z(t))$. We will denote $z(t, z_0)$ the solution at time $t$ to the Cauchy problem

$$\left\{ \begin{array}{rcl} \dot{z}(t) & = & F(t, z(t)), \\ z(0) & = & z_0. \end{array} \right.$$

We also denote $R(z_0, z(t_f))$ the transversality conditions stated in the PMP, as well as the initial and final constraints on the state. Finding an extremal satisfying the set of equations of the PMP amounts to finding an initialization $z_0$ and a final time $t_f$ (if it is not fixed) such that $R(z_0, z(z_0, t_f)) = 0$.

We denote $G(z_0, t_f) := R(z_0, z(z_0, t_f))$. It follows that finding an extremal satisfying the set of equations of the PMP boilds down to finding a zero of the function $G$. In the literature, the function $G$ is often named a *shooting function*, and looking for a zero of this function is a *shooting problem*.

**Remark 2.4: Resolution of a shooting problem**

Solving a shooting problem, i.e., finding a zero of the shooting function is usually done by means of a Newton-like method. Those methods are famous for having very fast convergence rates, while having a potentially small convergence domain. It means that the initialization of the shooting problem can be very intricate. Therefore, being able to design a good enough initialization ensuring the convergence of the Newton method is a challenge when solving an optimal control problem with an indirect method. In Section 2.4.1 we will detail a numerical procedure to adress this issue, deforming continuously the optimal control problem at hand.

### 2.3.3 Comparison between the methods

We are now concluding this section by putting together some of the elements of the previous paragraphs, in order to give some elements of comparison between direct and indirect methods. The survey paper [Tré12] also compares those two families of methods.

Based on a full discretization of the optimization problem, direct methods are often described as robust methods, in the sense that they do not require much knowledge *a priori* on the structure of the solution, (even if obviously, carefully chosing the initialization of the optimization algorithm can increase the speed of convergence). Besides, they allow to take into account all type of constraints, including state constraints. However, the discretization of the optimization problem can be a cause for the apparition of local minima, and the user of the optimization software can not have the guarantee to obtain a global solution. Such a problem may arise for instance when the discretization is too fine. Moreover, when compared with indirect methods, the numerical accuracy obtained with direct methods may be not as good. Aerospace is a field often put forth as a domain of application requiring high-level of numerical precision.

As for indirect methods, they rely on writing a maximum principle and solving some shooting problem. This resolution is often done by means of Newton-like methods, and therefore indirect methods inherit from the strengths and weaknesses of Newton methods: the rate of convergence is quadratic, and the method is both fast to converge and very precise. Note also that the integration of the differential system in the shooting function can be done using a numerical integrator, which can be very precise. However, the domain of convergence of the method can be very small, making its initialization difficult. It is therefore often required to have *a priori* knowledge on the structure of the solution (for instance the number of switchings of the control, or the number of fronteer arcs, as in [BFLT03]). Besides, using an indirect method to solve an optimal control problem with state constraints imply to use a PMP including such constraints, which can be very intricate.

Roughly speaking, it is often said that direct methods discretize the problem first before applying a dual method, whereas indirect methods first exploit the duality in the PMP before discretizing the problem.

Let us mention the existence of a large family of methods, namely hybrid methods, based on the combination of direct and indirect methods. For instance, when an optimal control problem is solved with the solver IPOPT, the output contains the value of the Lagrange multipliers for the underlying finite-dimensional optimization problem. Up to the sign, those multipliers are an approximation of the adjoint vector $p(\cdot)$ in the statement of the PMP. Therefore, they can be used to build the initialization of the adjoint vetor $p(0)$. With such a procedure, one can hope to benefit from the strength of both direct methods - little knowledge *a priori* on the structure of the solution - and indirect methods with a fast and precise convergence.

## 2.4 Application to the attitude control problem for a rigid body

In this section, we aim at showing how the two families of numerical methods presented in the previous Section 2.3 can be implemented to solve the attitude control problem with minimization of the consumption (**OCP**) we stated in the Introduction:

$$(\mathbf{OCP}) \begin{cases} \min \quad \int_0^{t_f} \sum_{i=1}^m |u_j(t)|\, dt + \lambda_0 t_f, \\ \dot{x}(t) = f_0(x(t)) + \sum_{j=1}^{14} u_j(t) f_j(x(t)), \\ \forall i \in [\![1, m]\!], \quad 0 \leqslant u_i(t) \leqslant 1 \quad \text{p.p. on } [0; t_f], \\ x(0) = x_0, \\ x(t_f) = x_f. \end{cases}$$

Note that the term accounting for the final time in the cost can be written under the form

$$\lambda_0 t_f = \lambda_0 \int_0^{t_f} 1\, dt$$

if one wishes to have no function $g$ in the general cost (2.2). This is what we do in the following.

### 2.4.1 With an indirect method

As explained in Remark 2.4, one of the difficulty when solving a shooting problem is the initialization of the underlying Newton-like method to find a zero of the shooting function. It is well-known that the problem of minimizing the consumption is part of the problems for which this initialization is indeed difficult, as in [CGN03, HMG04, GH06, CDG12].

#### Continuation procedure

A common technique to overcome this difficulty is the use of a *continuation procedure* (sometimes also named homotopy procedure). The idea is to introduce a parametrization in the expression of the optimal control problem, in order to deform the initial problem, deemed to be hard to solve, into an easier problem for which an initialization can easily be provided, either because one has *a priori* knowledge on the structure of the solution, or because the convergence domain of the Newton method is wide enough.

This deformation can be introduced as a change in the expression of the cost, in order to benefit from convexity properties. This is what is done for instance in [CGN03, HMG04, GH06, CDG12]. In our case, for $\alpha \in [0, 1]$, we introduce the optimal control problem $(\mathbf{OCP})_\alpha$ :

$$(\mathbf{OCP})_\alpha \begin{cases} \min \quad \alpha \int_0^{t_f} \sum_{j=1}^m u_j(t)^2\, dt + (1-\alpha) \int_0^{t_f} \sum_{j=1}^m |u_j(t)|\, dt + \lambda_0 t_f, \\ \dot{x}(t) = f_0(x(t)) + \sum_{j=1}^m u_j(t) f_j(x(t)), \\ \forall i \in [\![1, m]\!], \quad 0 \leqslant u_i(t) \leqslant 1 \quad \text{p.p. on } [0; t_f], \\ x(0) = x_0, \\ x(t_f) = x_f. \end{cases}$$

Thus, when $\alpha = 0$, one recognizes the initial problem (**OCP**), and when $\alpha = 1$, it consists in the problem of minimizing the energy:

$$\text{minimize} \quad \int_0^{t_f} \sum_{j=1}^m u_j(t)^2\, dt + \lambda_0 t_f.$$

We recall that in order to be exploitable by the "SCA" of Ariane 5, the control has to be bang-bang, i.e., has to take its values in $\{0,1\}$. However, it is not obvious at first that the above control problems yield bang-bang controls. We will see in the following that it is not the case when $\alpha \neq 0$. However the computation of those regular controls will allow us to eventually solve (**OCP**). At the end, the procedure will result in bang-bang controls.

### Application of the PMP to $(\mathbf{OCP})_\alpha$

We shall start by detailing the application of the PMP to the optimal control problem $(\mathbf{OCP})_\alpha$. The Hamiltonian of the system writes

$$H(x,u,p,p^0) = \langle p, f(x,u)\rangle + p^0\left(\alpha\sum_{j=1}^m u_j(t)^2 + (1-\alpha)\sum_{j=1}^m |u_j(t)| + \lambda_0\right),$$

We denote $p = (p_\theta, p_\psi, p_\varphi, p_p, p_q, p_r)$ the adjoint vector. If a trajectory $(x(\cdot), u(\cdot))$ is optimal, then there exists a non trivial pair $(p(\cdot), p^0)$ with $p^0 \leqslant 0$ such that the dynamics of $p(\cdot)$ are given by

$$\left\{\begin{array}{rcl}
\dot p_\theta &=& 0,\\
\dot p_\psi &=& -p_\theta\left(\frac{\sin\varphi\sin\psi}{\cos^2\psi}q + \frac{\cos\varphi\sin\psi}{\cos^2\psi}r\right) - p_\varphi\left(\sin\varphi q + \cos\varphi r\right)\left(1+\tan^2\psi\right),\\
\dot p_\varphi &=& -p_\theta\left(\frac{\cos\varphi}{\cos\psi}q - \frac{\sin\varphi}{\cos\psi}r\right) - p_\psi(-\sin\varphi q - \cos\varphi r) - p_\varphi(\cos\varphi\tan\psi q - \sin\varphi\tan\psi r),\\
\dot p_p &=& -a_2 p_q r - a_3 p_r q - p_\varphi,\\
\dot p_q &=& -a_1 p_p r - a_3 p_r p - p_\theta\frac{\sin\varphi}{\cos\psi} - p_\psi\cos\varphi - p_\varphi\sin\varphi\tan\psi,\\
\dot p_r &=& -a_1 p_p q - a_2 p_q p - p_\theta\frac{\cos\varphi}{\cos\psi} - p_\psi\sin\varphi - p_\varphi\cos\varphi\tan\psi.
\end{array}\right.$$

Besides, because of the maximization condition (2.7), for all $j \in [\![1,m]\!]$, each component $u_j$ of the control maximizes almost everywhere the quantity

$$u_j(t)\langle p(t), f_j\rangle + p^0\left(\alpha u_j(t)^2 + (1-\alpha)u_j(t)\right).$$

In what follows, we restrict ourselves to finding normal extremals, i.e., extremals with $p^0 < 0$. It can then be assumed that $p^0 = -1$.

**Case $\alpha \neq 0$.** In that case, the control maximizes almost everywhere the quadratic function

$$u_j(t)\langle p(t), f_j\rangle - \alpha u_j(t)^2 - (1-\alpha)u_j(t),$$

over the interval $[0,1]$. This function reaches its maximum at the unique point where its derivative vanishes, or on the boundary of the interval. We get that, for almost every time $t$,

$$u_j(t) = \max\left(0, \min\left(1, \frac{\langle p(t), f_j\rangle - (1-\alpha)}{2\alpha}\right)\right).$$

Therefore the control has the same regularity as the adjoint vector $p(\cdot)$. In particular, as soon as $p$ is continuous, the control is a continuous function of the time. However, such a control does not fulfill the requirement, imposed by the design of the "SCA", to be bang-bang.

**Case $\alpha = 0$.** In that case, each component $u_j$ of the control can be obtained by minimizing the affine function

$$u_j(t)\langle p(t), f_j\rangle - u_j(t).$$

It follows that

$$u_j(t) = \text{sign}\left(\langle p(t), f_j\rangle - 1\right),$$

where sign$(\cdot)$ is the sign function defined by:

$$\text{sign}(x) = \begin{cases} 1 & \text{if} \quad x > 0, \\ 0 & \text{if} \quad x < 0. \end{cases}$$

The function $t \mapsto \langle p(t), f_j\rangle - 1$ can sometimes be found under the name *switching function* in the literature, as its sign will decide if the thruster $j$ is to be closed or opened. Let us point out that $u_j$ is undetermined when the switching function vanishes. If it happens on a countable subset of the time interval, it has no effect as the maximization condition (2.7) of the Hamiltonian stands almost everywhere. However, when there exists a time interval $[t_1, t_2]$ on which the switching function vanishes, the control $u_j$ can not be computed directly [1]. Such a control is then often said to be singular.

### Algorithmic procedure

First, recall that solving an optimal control problem with an indirect method as explained in Subsection 2.3.2 boils down to finding the zeros of some shooting function 2.3.2. For the optimal control problem $(\mathbf{OCP})_\alpha$, the transversality conditions (2.10) on the final time in the PMP and the constraints on the state at final time write

$$\begin{aligned} \max_{v\in\Omega} H_\alpha(t_f, x(t_f), v, p(t_f), p^0) = 0, \\ x(t_f) - x_f = 0. \end{aligned}$$

As the initial and terminal submanifolds are singletons $\mathcal{M}_0 = \{x_0\}$ and $\mathcal{M}_1 = \{x_f\}$, the transversality equations on the adjoint vector (2.8) and (2.9) are trivial.

In the problem of interest, the initial state of the launcher is fixed, with $x(0) = x_0$. Therefore, the switching function only depends on the initialization of the adjoint vector $p(0)$ and on the final time $t_f$, and we denote it $G_\alpha(p_0, t_f)$. We will also denote $Z$ the variable of the function $G_\alpha$, and $\tilde{Z}_\alpha$ a zero of the function $G_\alpha$. We emphasize that the dimension of the variable $Z$ is 7 (6 components for $p_0$ and one for $t_f$), and the function $G_\alpha$ also has 7 components (one equation for the transversality condition on the final time, and 6 equations for the constraint on the final state $x(t_f) - x_f = 0$). Therefore, the shooting problem is well-posed.

**Simple continuation procedure.** Our final goal is therefore to find a zero of the function $G_0$, which corresponds to the shooting function for the problem (**OCP**), with minimization of the consumption. To do so, we look for a sequence of parameters $(\alpha_k)_{k\in[\![0,N]\!]}$ such that $\alpha_0 = 1$ and $\alpha_N = 0$, and such that for each $k \in [\![0, N]\!]$, we know a zero $\tilde{Z}_{\alpha_k}$ of the shooting function $G_{\alpha_k}$. The interest of the procedure lies in the fact that for each $k \in [\![0, N-1]\!]$, we can use the solution $\tilde{Z}_{\alpha_k}$ of the problem $(\mathbf{OCP})_{\alpha_k}$ to initialize the search for a zero of the shooting function $G_{\alpha_{k+1}}$. An implicit assumption is made that we are able to compute the solution at the first step

1. The usual technique to derive the expression of the control is to differentiate the relation $\langle p(t), f_j\rangle - 1 = 0$ a number of times sufficient for the control $u_j$ to appear explicitely. Generically, the resulting controls are not bang-bang, and can not be used to solve the control problem at hand in this thesis

of the continuation procedure, namely a zero $\tilde{Z}_1$ of $G_1$, the shooting function for the problem $(\mathbf{OCP})_1$. Hereafter at Algorithm 1, we give in pseudo-code the algorithmic principle of a simple continuation, and the Figure 3.1 schematizes this procedure.

**Algorithm 1** General principle of the continuation procedure

$\overline{Z} = \tilde{Z}_1$ ▷ Initialization for $\alpha = 1$
step $\in [0, 1]$ ▷ Reference step
$\text{step}_m \in [0, \text{step}]$ ▷ Minimal step
**while** $\alpha > 0$ et $\text{step} > \text{step}_m$ **do**
  $\text{step} \leftarrow \min(\text{step}, \alpha)$
  $\overline{\alpha} = \alpha - \text{step}$ ▷ $\alpha$ decreases
  Look for $\tilde{Z}$, zero of the function $G_{\tilde{\alpha}}(Z)$, with $\overline{Z}$ serving as an initialization.
  **if** success **then**
    $\alpha \leftarrow \overline{\alpha}$
    $\overline{Z} \leftarrow \tilde{Z}$ ▷ We move on
  **else**
    $\text{step} \leftarrow \frac{\text{step}}{2}$ ▷ We decrease the step and start again
  **end if**
**end while**

This algorithm could be improved in many ways. For instance, in case of a success in the resolution, it can be decided to increase the step in order to improve the speed of convergence of the algorithm. We refer to the book [AG90] for more details on the numerical implementation of a continuation procedure. Besides, they are many existing softwares available online, as the open source HamPath [CCG12].

In the next paragraph, we will present an improvement of the Algorithm 1, introduced to decrease its runtime.

**Continuation procedure with linear prediction.** Behind this method is the idea that we can do better than just using $\tilde{Z}_\alpha$ as an approximation of $\tilde{Z}_{\alpha-\Delta\alpha}$ when initializing the shooting problem. Assume that we have already made two successive resolutions, yielding $\tilde{Z}_{\alpha+\Delta\alpha_1}$ and $\tilde{Z}_\alpha$, for two values $\alpha+\Delta\alpha_1$ and $\alpha$. Assuming some regularity on the path of zeros, an approximation of $\tilde{Z}_{\alpha-\Delta\alpha_2}$ for a new value $\alpha - \Delta\alpha_2$ is given by

$$\tilde{Z}_{\alpha-\Delta\alpha_2} \approx \tilde{Z}_\alpha - \frac{\Delta\alpha_2}{\Delta\alpha_1}\left(\tilde{Z}_{\alpha+\Delta\alpha_1} - \tilde{Z}_\alpha\right),$$

as displayed on Figure 2.2.

This is the procedure we used throughout this thesis each time a continuation is performed (in the context of indirect methods), as we could experimentally witness an improvement in the runtime of the algorithm.

### Numerical results

We give here some numerical results of the implementation of the continuation procedure with linear prediction applied to the attitude control problem. We used the following numerical

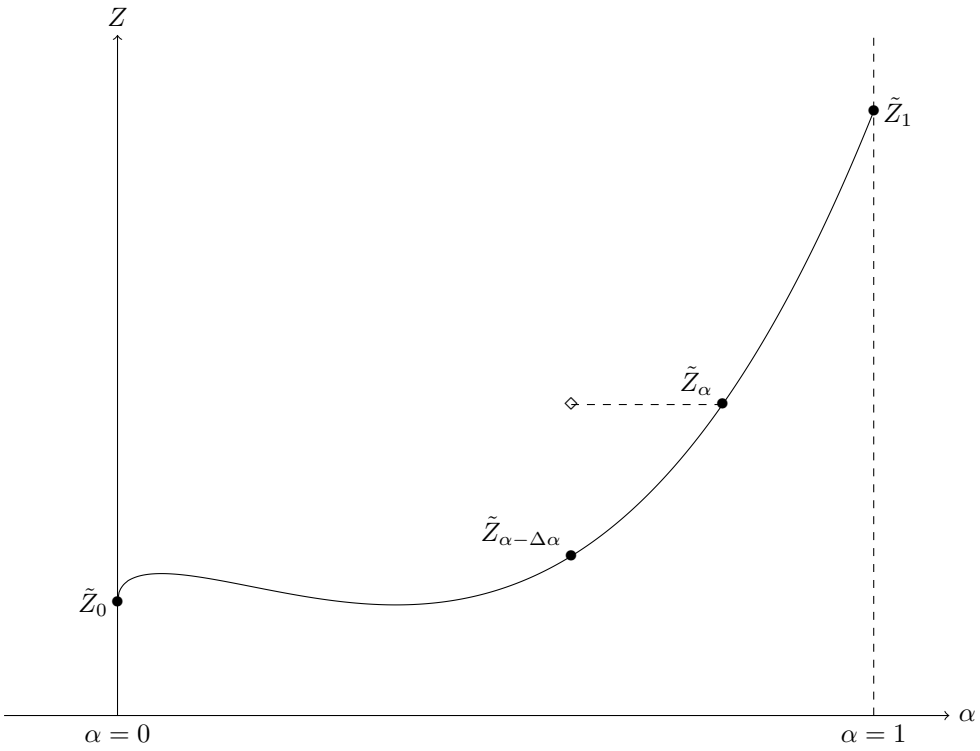



Figure 2.1 – General principle of the continuation procedure. The resolution of $(\mathbf{OCP})_\alpha$ is used to initialize the resolution for $\alpha - \Delta\alpha$.

values for the initial and final conditions

$$(\theta_0, \psi_0, \varphi_0, p_0, q_0, r_0) = (0.04, 0.06, 7.7, -0.027, 0, 0),$$
$$(\theta_f, \psi_f, \varphi_f, p_f, q_f, r_f) = (0.63, 0.82, 7.0, -0.008, 0, 0),$$

and we chose the following expression for the cost

$$J_\alpha(u) = \alpha \int_0^{t_f} \sum_{j=1}^{m} u_j(t)^2 \, dt + (1-\alpha) \int_0^{t_f} \sum_{j=1}^{m} |u_j(t)| \, dt + \frac{t_f}{2},$$

i.e., we set the parameter $\lambda_0 = 1/2$. Note that the angles are expressed in radians, and the angular velocity in radians per second.

The integration of the differential system to compute the shooting function $G_\alpha$ is done using the numerical integrator DOP853. It consists in an explicit Runge-Kutta method with adaptative step comparing the methods RK8, RK5 and RK3. The description of the algorithm can be found in [HNW08].

On the Figures 2.3, 2.4, 2.5 et 2.6, we display the evolution of the controls during the continuation on the parameter $\alpha$. We represent in black the controls for the problem of minimizing the energy $(\mathbf{OCP})_1$, and in red the controls for the minimization of the consumption $(\mathbf{OCP})$. We also chose to display the controls at two intermediate stages of the continuation, for the values $\alpha = 0.32$ and $\alpha = 0.02$. It appears clearly how the controls are deformed progressively from a

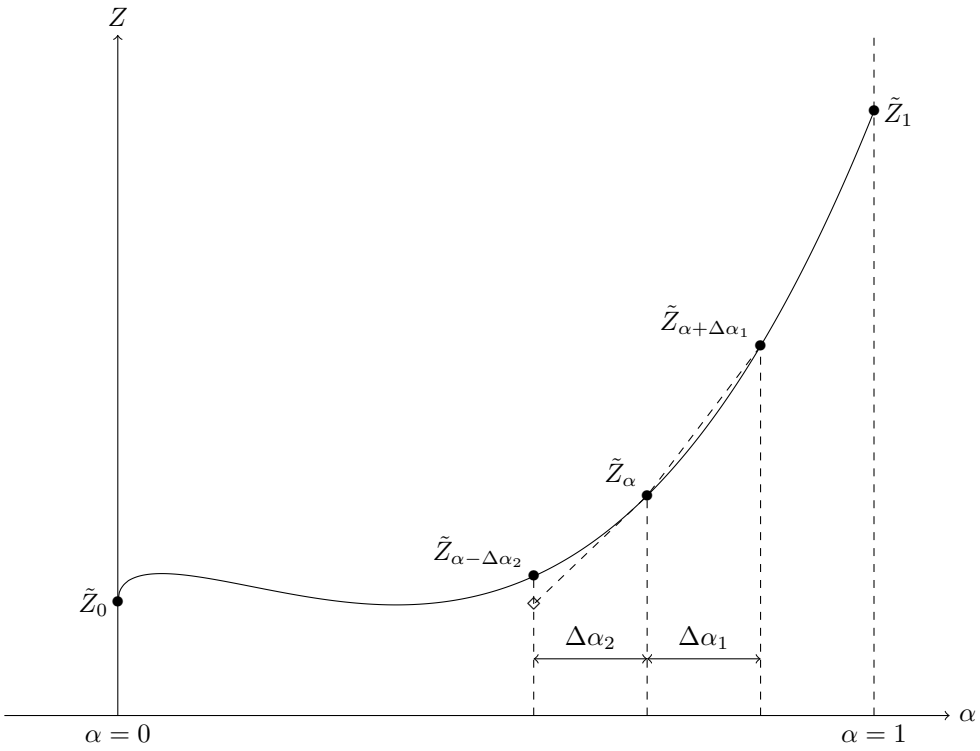



Figure 2.2 – Principle of the continuation procdeure with linear prediction. The solutions of the problem $(\mathbf{OCP})_\alpha$ and $(\mathbf{OCP})_{\alpha+\Delta\alpha_1}$ are used to initialize the shooting problem for $\alpha - \Delta\alpha_2$, doing an affine extrapolation in order to get an approximation for $Z_{\alpha-\Delta\alpha_2}$.

continuous command law (for $(\mathbf{OCP})_1$), to a bang-bang command law (for $(\mathbf{OCP})$).

On Figure 2.7, we also represent the trajectory for the minimization of the consumption in $(\mathbf{OCP})$. Physically, it corresponds to steering the launcher from a spinned state along its main inertia axis (the roll velocity $p$ is non zero, and the transverse angular velocities $q$ and $r$ are zero) to another spinned state in a different orientation ($(\theta_0, \psi_0, \varphi_0) \neq (\theta_f, \psi_f, \varphi_f)$). The angles (resp. the angular velocities) are expressed in radians (resp. radians per second).

### 2.4.2 With a direct method

Now, we are going to illustrate how the attitude control problem with minimization of the consumption can be tackled with a direct method. In order to insist on the fact that such a method does not require a priori knowledge on the structure of the solution, yet allowing to easily consider state constraints, we consider the optimal control problem $(\mathbf{OCP})$ [2] to which we add an additionnal constraint: during the maneuver, the transverse angular velocities $q$ and $r$

2. Solving directly $(\mathbf{OCP})$ does not pose a problem. When we performed the numerical simulations, we obtained the same controls and the same trajectory as those computed with the indirect method in the previous Subsection.

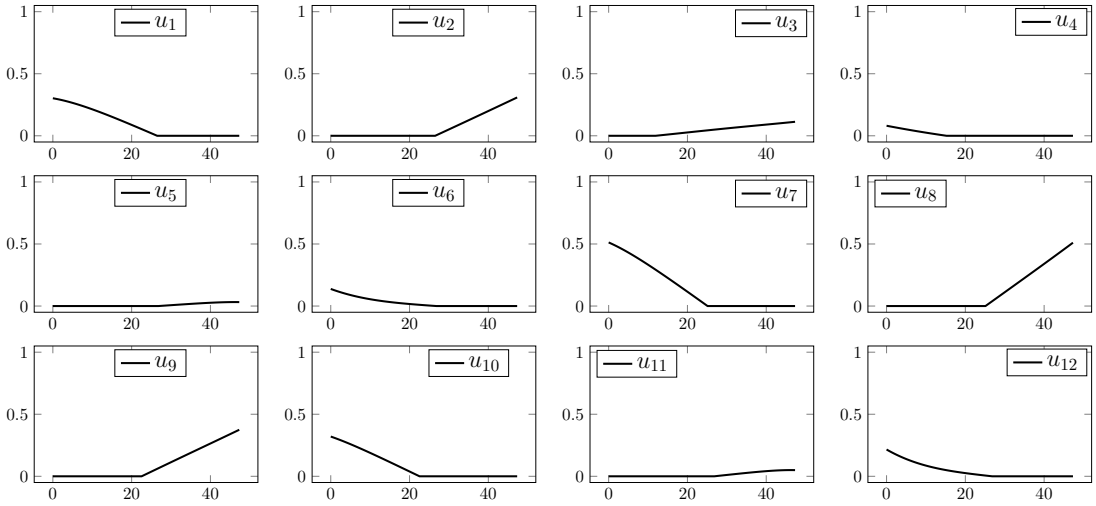



Figure 2.3 – Controls for the resolution of $(\mathbf{OCP})_1$.

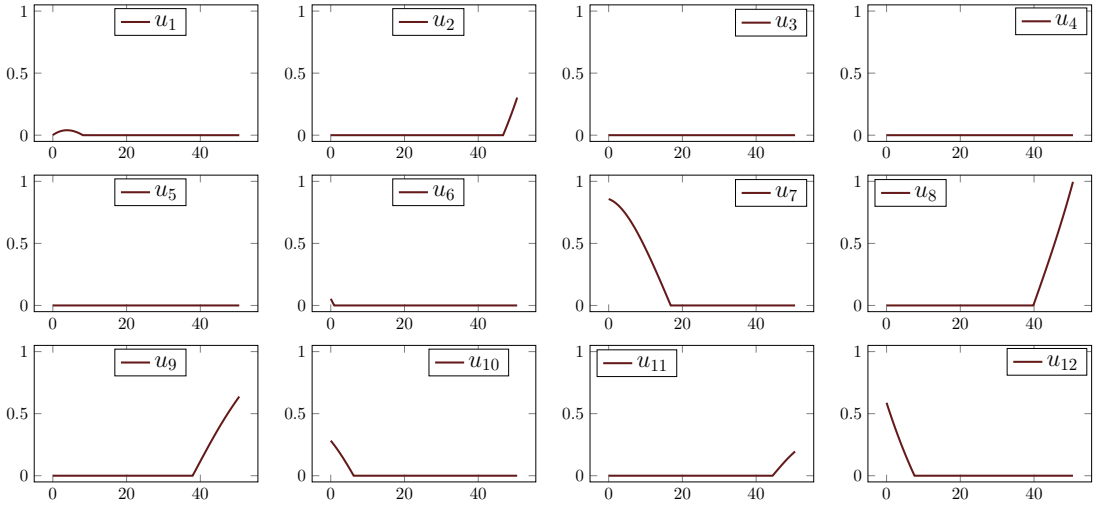



Figure 2.4 – Controls for the resolution of $(\mathbf{OCP})_\alpha$ with $\alpha = 0.36$.

are requested to stay below a certain level, namely

$$|q| \leqslant \omega_t^{\max}, \quad |r| \leqslant \omega_t^{\max}. \tag{2.11}$$

We emphasize again that when using the interior-point solver IPOPT [WB06a] (with the modelling language AMPL [FGK93] or not), taking into account state constraints such as (2.11) does not make the implementation harder.

We consider the same initial and final conditions as in the previous section, where the launcher is controlled from a spinned state along its principal inertia axis to another spinned state, in a different orientation. We chose the following numerical value for the constraint on the state

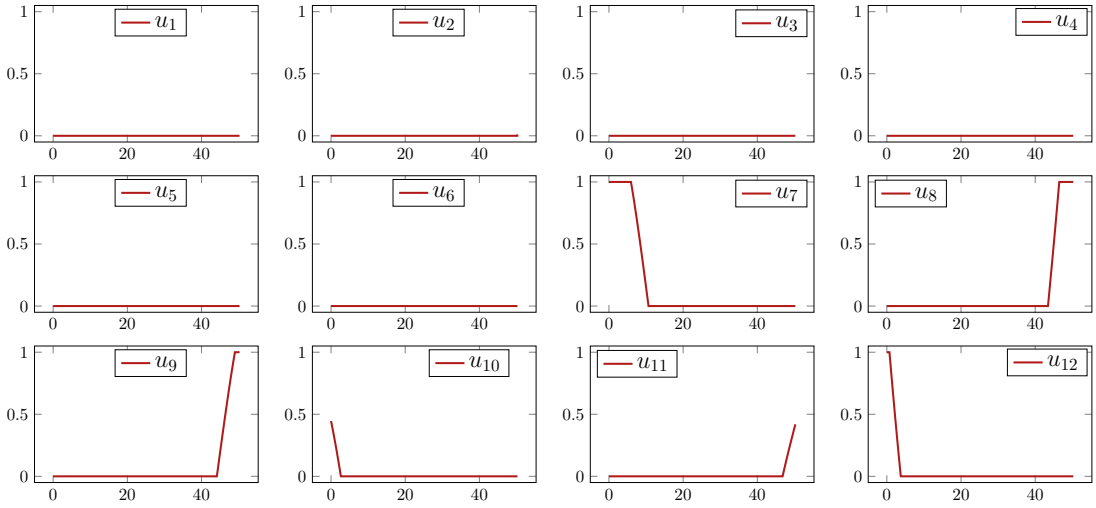



Figure 2.5 – Controls for the resolution of $(\mathbf{OCP})_\alpha$ with $\alpha = 0.02$.

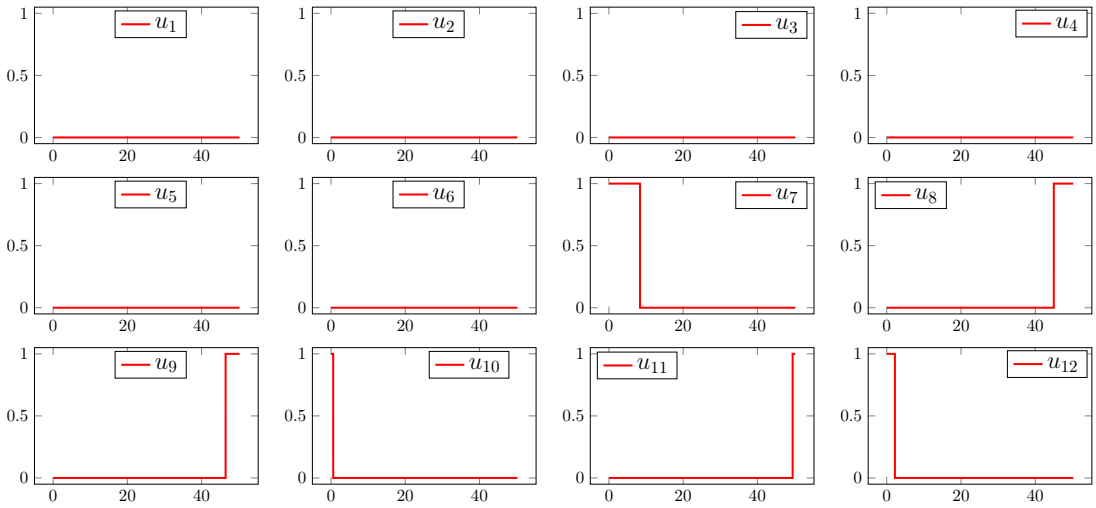



Figure 2.6 – Controls for the resolution of $(\mathbf{OCP})$.

(2.11)

$$\omega_t^{\max} = 0.007 \text{ rad.s}^{-1}.$$

On Figure 2.8 are displayed the controls and on Figure 2.9 the trajectory of the launcher. A remarkable fact appears clearly on those two figures: when the state of the system saturates the constraint (2.11), the controls are not anymore bang-bang. In order to give an insight on this fact, we quickly give some theoretical elements on control systems with a state constraint.

For the sake of simplicity, we restrict ourselves to the more simple system

$$\dot{y}(t) = f_0(y(t)) + uf_1(y(t)),$$

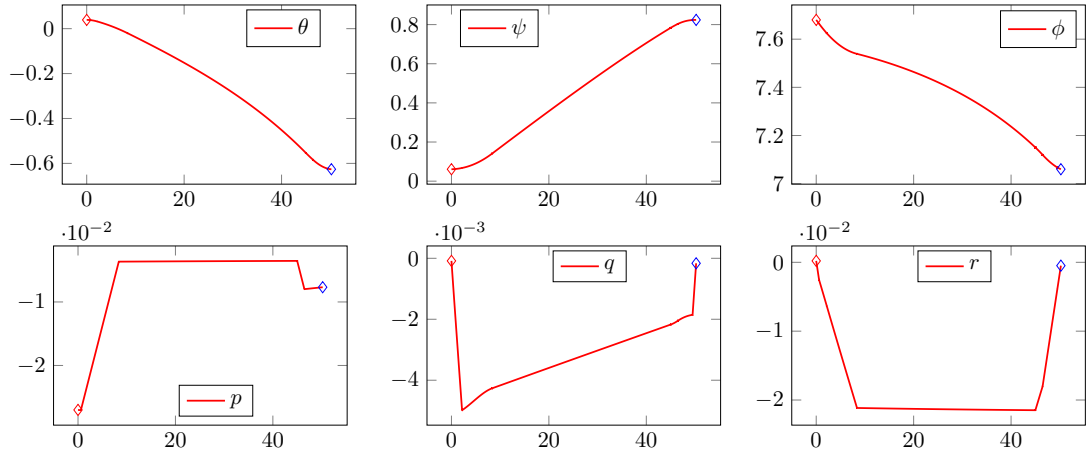



Figure 2.7 – Trajectory for the resolution of problem (**OCP**). The attitude of the launcher is controlled from the initial state $x_0$ (◇) to the final state $x_f$ (◇).

with only one control $u$, and where the state $y$ belongs to $\mathbb{R}^n$ (we denote $y$ this state and not $x$ as in the rest of the chapter in order to insist on the fact that it is *not* the attitude control system). Besides, we add a state constraint under the form

$$c(y) \leqslant 0,$$

with $c : \mathbb{R}^n \to \mathbb{R}$. Assume that the constraint is active between the times $t_1$ and $t_2$. One gets an expression for the control by differentiating the relation $c(y(t)) \equiv 0$ on $[t_1, t_2]$. Differentiating this relation once, we get

$$\begin{aligned} \nabla c(y(t)) \cdot \dot{y}(t) &= 0 \\ \nabla c(y(t)) \cdot (f_0(y(t)) + u(t) f_1(y(t))) &= 0 \end{aligned}$$

Thus, if $\nabla c(y(t)) \cdot f_1(y(t) \neq 0$, the control can be expressed under the feedback form:

$$u(t) = -\frac{\nabla c(y(t)) \cdot f_0(y(t))}{\nabla c(y(t)) \cdot f_1(y(t))}.$$

Note that the terms $\nabla c(y(t)) \cdot f_i(y(t)$ can be written under the form $(f_i c)(y(t))$ if we consider that the vector field $f_i$ acts as a derivation on $c$:

$$(f_i c)(y(t)) := \nabla c(y(t)) \cdot f_i(y(t)).$$

Let $M$ be the number of times one needs to differentiate the relation $t \mapsto c(y(t)) \equiv 0$ in order to have $f_1^M c \neq 0$ for the first time. By an easy iteration, it follows that

$$(f_0 f_1^{M-1} c)(y(t)) + u(t)(f_1^M c)(y(t)) = 0,$$

and the control can again be expressed under a feedback form

$$u(t) = -\frac{(f_0 f_1^{M-1} c)(y(t))}{(f_1^M c)(y(t))},$$

which is then not bang-bang.

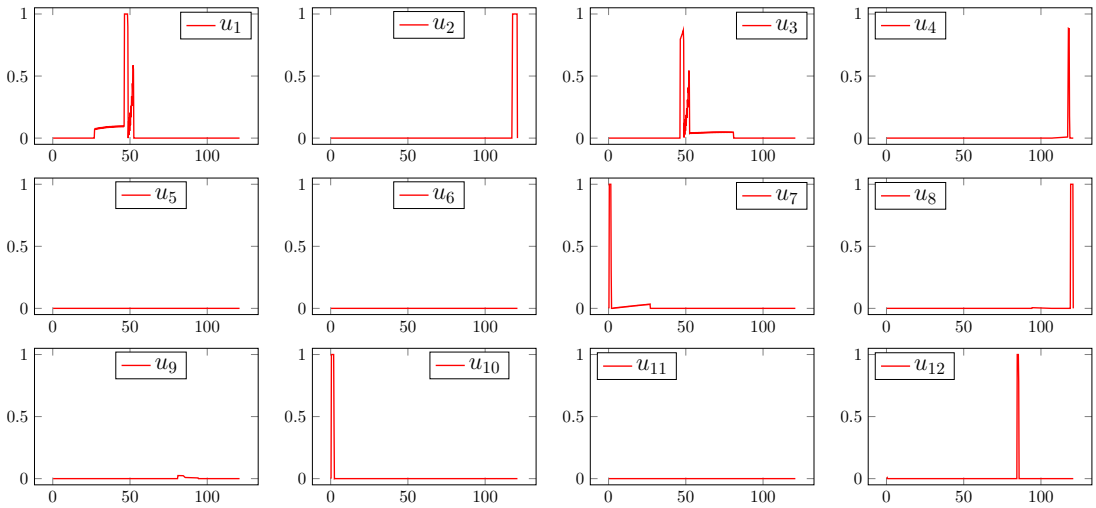



Figure 2.8 – Controls for the resolution of the problem (**OCP**) with a state constraint.

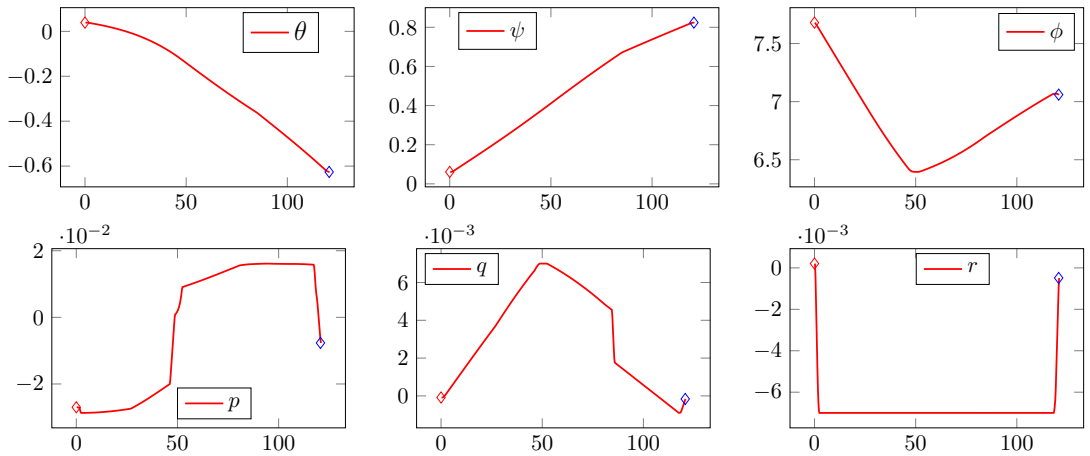



Figure 2.9 – Trajectory for the resolution of the problem (**OCP**) with a constraint on the transverse angular velocities $q$ and $r$.

## 2.5 Conclusion of the chapter

In this chapter, we recalled the statement of Pontryagin maximum principle for a nonlinear control system with constraints on the control but no state constraints, as well as some of the usual methods in optimal control theory. In the following chapter, we will intensively use both direct and indirect methods to solve numerically optimal control problems, chosing the most adequate method depending on the problem at hand.

One of the key idea presented here is the use of a continuation procedure to solve a problem deemed to be hard to solve. When implementing such a procedure, the initial control problem is embedded in a family of problems depending on one (or several parameters). The aim is to deform it in order to end up with an "easier" problem. In the next chapters of this thesis, we will thoroughly use continuation procedures: each time we compute an optimal trajectory with respect to the $L^1$ cost, it actually comes from a continuation $L^2 \to L^1$ as the one we presented in this chapter. Besides, in chapters 3 and 5, we will use such a procedure in various settings, each time to solve numerically the optimal control problem under consideration.

The numerical examples displayed at the end of the chapter show how to compute an optimal trajectory when only one satellite is to be separated. However, the indirect method we presented here is deficient to tackle a more complex ballistic phase when several satellites are boarded on the launcher. In the context of optimal control problems with intermediate constraints, this is what we are going to study in the next chapter.

Chapter 3

# Optimal control with intermediate constraints

## Contents


**2.1 General setting . . . . . 32**
**2.2 Pontryagin Maximum Principle . . . . . 33**
**2.3 Numerical methods in optimal control . . . . . 36**
2.3.1 Direct methods . . . . . 36
2.3.2 Indirect methods . . . . . 37
2.3.3 Comparison between the methods . . . . . 38
**2.4 Application to the attitude control problem for a rigid body . . . 39**
2.4.1 With an indirect method . . . . . 39
2.4.2 With a direct method . . . . . 44
**2.5 Conclusion of the chapter . . . . . 49**


In the previous chapter, we explained how a continuation procedure can be implemented to solve an optimal control problem arising during a simple ballistic phase, when only one satellite is put into orbit. During a complex ballistic phase, several bodies are successively put into orbit. At the times of the separations, there may be in the description of the mission constraints on the state of the system that do not concern the 6 components of the state. It can be also requested during a ballistic phase to cancel at a given time the angular velocity of the launcher while not constraining the angles $\theta$, $\psi$ and $\varphi$.

In this chapter, we will give a general numerical algorithm to solve an optimal control problem with intermediate constraints by means of an indirect method. We write it in a general setting as we believe it could be applied to a wide range of problems. However, we will also point out that our procedure sometimes fails to converge when the number of intermediate constraints becomes too important.

That is why, in Appendix A, we will also present the results given by an optimization software designed for the CNES. The software tackles the question of finding the optimal solution for a complex ballistic phase, with any number of separations, and in a general fashion, any number of intermediate constraints. For this reason, this chapter and Appendix A are very complementary, as they address the same problems with different approaches.

## 3.1 Introduction of the chapter

**Optimal control problems with intermediate constraints.** Let $n$, $m$ and $p$ be positive integers. In this chapter, we consider the general nonlinear control system

$$\dot{x}(t) = f(t, x(t), u(t)), \tag{3.1}$$

where the state $x(t) \in \mathbb{R}^n$ and the control is subject to the constraint $u(t) \in \Omega = [0,1]^m$. Our goal is to find a control $u(\cdot)$ and a final time $t_f$ that steer the control system (3.1) from an initial point $x_0$ to a final point $x_f$ (both fixed), while minimizing an integral cost

$$J(u) = \int_0^{t_f} f^0(t, u(t), x(t))\, dt, \tag{3.2}$$

and enforcing an *intermediate constraint* (or *interior-point constraint*) at some (fixed) intermediate time $t_1 \in (0, t_f)$ that we write under the generic form :

$$g(t_1, x(t_1)) = 0, \tag{3.3}$$

where $g : \mathbb{R}^n \to \mathbb{R}^p$ is a smooth function.

Note that this formulation (and all the results presented thereafter) can easily be extended to a problem with several intermediate constraints. However, for the sake of simplicity, we will only present first the case of one intermediate constraint.

The literature on control systems with intermediate constraints is abundant. Let us mention [BH75], a classical reference on optimization problems with interior-point constraints, and more generally with state constraints along the path. It has been shown in [DK08, DK11] that our problem can be seen as a particular instance of a *hybrid control problem* (see also [BBM98, GP05b, SC07, Sus99] for more details on hybrid control systems). The authors show how to reduce the optimal control problem (3.1)-(3.2)-(3.3) with intermediate constraints (as well as other general classes of hybrid optimal control problem) to a "classical" optimal control problem to which one can apply the usual PMP of [PBGM62]. We will use their results to derive the Propositions of Section 3.2. Recall that the PMP consits in a set of necessary conditions for a control and a trajectory to be optimal. Recall also that in the context of indirect methods, those conditions can be used to reduce the resolution of an optimal control problem to finding the zeros of some shooting function. We will elaborate on this issue in more details in the core of this chapter.

Let us denote $(\mathcal{P})_0$ the optimal control problem of steering the control system (3.1) from $x_0$ to $x_f$, while minimizing the cost (3.2), *without the intermediate constraint* (3.3). Throughout this chapter, we will assume that $(\mathcal{P})_0$ has at least one optimal solution, that we will denote $(\overline{x}(\cdot), \overline{u}(\cdot))$.

A first idea to solve the initial problem with the intermediate constraint (3.3) is to introduce the constraint by *continuation*, or *homotopy*, solving a sequence of problems that depend on a parameter $s \in [0,1]$, each problem containing a constraint:

$$g(t_1, x(t_1)) = s \cdot g(t_1, \overline{x}(t_1)). \tag{3.4}$$

For $s = 1$, one can notice that $(\overline{x}(\cdot), \overline{u}(\cdot))$ is a solution of the problem, and for $s = 0$, one finds our initial problem. In the following, we will denote $(\mathcal{P})_{via,s}$ the optimal control problem of steering the system (3.1) from $x_0$ to $x_f$ while minimizing the cost (3.2) and satisfying the constraint (3.4) with a continuation parameter $s$. Therefore, the goal of this chapter is to propose a robust and

efficient procedure to solve $(\mathcal{P})_{via,0}$, based on a mathematically sound theory.

**Numerical difficulty.** Even though performing a continuation on the parameter $s$ can sometimes be enough to solve $(\mathcal{P})_{via,0}$, we experimentally noticed that in some cases, the procedure fails to converge, even for values close to $s = 1$. Unfortunately, so far we have not been able to identify clearly the reason for this failure. We give hereafter two possible reasons, that need to be further investigated:

- because of a local loss of controllability around some value $s_1 > 0$. It could happen that the problem $(\mathcal{P})_{via,s_1}$ admits a solution, but that the problem $(\mathcal{P})_{via,s}$ does not for values $s < s_1$ close enough to $s_1$. In that case, there is a barrier somewhere during the continuation.
- because of the presence of *singular trajectories* along the path, that forbids convergence of the underlying shooting method

In [Tré12] conditions ensuring local and global convergence of numerical continuation methods in optimal control are given.

**Penalizing the intermediate constraint.** To avoid this numerical difficulty, we consider another optimal control problem, consisting of steering the control system (3.1) from $x_0$ to $x_1$ in some time $t_f$ while minimizing the cost functional:

$$J_\varepsilon(u) = \int_0^{t_f} f^0(t, u(t), x(t))\, dt + \frac{1}{\varepsilon} \left\| g(t_1, x(t_1)) \right\|^2. \tag{3.5}$$

Here, the intermediate constraint has been dropped and replaced by some penalization term included in the cost functional. It is therefore much less restraining than imposing a constraint of the form $g(t_1, x(t_1)) = s \cdot g(t_1, \overline{x}(t_1))$. Note that the penalization term is not completely standard since it is at the intermediate time $t_1$. Let us denote $(\mathcal{P})_{pen,\varepsilon}$ the optimal control problem of steering the system (3.1) from $x_0$ to $x_f$ while minimizing the cost (3.5), that depends on the parameter $\varepsilon$. When $\varepsilon \gg 1$, the cost $J_\varepsilon(u)$ can be approximated (at least formally)

$$J_\varepsilon(u) \approx \int_0^{t_f} f^0(t, u(t), x(t))\, dt$$

and one recovers $(\mathcal{P})_0$. When $\varepsilon \ll 1$, the solution of $(\mathcal{P})_{pen,\varepsilon}$ is expected to be close to a solution of $(\mathcal{P})_{via,0}$. Note that with this formulation, one can not ensure exactly that $g(t_1, x(t_1)) = 0$. Besides, if $\varepsilon$ becomes too small, one could face the numerical pitfall of dividing by $\varepsilon$. However, we will see that when $\varepsilon$ is small enough, the solution of $(\mathcal{P})_{pen,\varepsilon}$ provides a good enough initialization to solve the initial problem $(\mathcal{P})_{via,0}$.

Before going further, let us recall here the expressions of the two optimal control problems of interest in this chapter.

$$(\mathcal{P})_{via,s} \begin{cases} \min \quad \int_0^{t_f} f^0(t, u(t), x(t))\, dt, \\ \dot{x}(t) = f(t, x(t), u(t)), \\ \forall i \in [\![1, m]\!], \quad 0 \leqslant u_i(t) \leqslant 1 \quad \text{p.p. on } [0; t_f], \\ x(0) = x_0, \\ x(t_f) = x_f, \\ g(t_1, x(t_1)) = s \cdot g(t_1, \overline{x}(t_1)). \end{cases} \tag{3.6}$$

$$(\mathcal{P})_{pen,\varepsilon} \begin{cases} \min \quad \int_0^{t_f} f^0(t,u(t),x(t))\,dt + \frac{1}{\varepsilon}\left\|g(t_1,x(t_1))\right\|^2, \\ \dot{x}(t) = f(t,x(t),u(t)), \\ \forall i \in [\![1,m]\!], \quad 0 \leqslant u_i(t) \leqslant 1 \quad \text{p.p. on } [0;t_f], \\ x(0) = x_0, \\ x(t_f) = x_f. \end{cases} \tag{3.7}$$

**Outline of the chapter.** Initially, our goal was to solve problems coming from aerospace and involving intermediate constraints, as we will do in Section 3.3. However, we believe it is worth writing the theoretical results in the general setting presented in this introduction, as our method might well also be used in a various range of domains. The paper is organized as follows. In Section 3.2, we state and prove two Pontryagin maximum principles for the general problems $(\mathcal{P})_{via,s}$ and $(\mathcal{P})_{pen,\varepsilon}$. In addition to the classical statement of the PMP, the adjoint vector is not anymore continuous, and presents jumps at the intermediate points. In Section 3.3, we apply the theoretical results of Section 3.2 to the attitude control problem of a three dimensional rigid body, a problem of importance in aerospace. Section 3.4 contains numerical examples to illustrate our procedure.

## 3.2 Optimal control formulation

As presented in Section 3.1, we suggest in this chapter two optimal control formulations to account for the intermediate constraint of our problem. $(\mathcal{P})_{via,s}$ consists in steering the system (3.1) from $x_0$ to $x_f$ while minimizing the cost (3.2) and satisfying a constraint

$$g(t_1,x(t_1)) = sg(t_1,\overline{x}(t_1)),$$

whereas in $(\mathcal{P})_{pen,\varepsilon}$, we penalize it in the cost

$$J_\varepsilon(u) = \int_0^{t_f} f^0(t,u(t),x(t))\,dt + \frac{1}{\varepsilon}\cdot\left\|g(t_1,x(t_1))\right\|^2.$$

Let us emphasize once again that solving $(\mathcal{P})_{pen,\varepsilon}$ up to small values of $\varepsilon$ enables us to circumvent the numerical difficulties that come up when solving $(\mathcal{P})_{via,s}$.

In this section, we present two Pontryagin maximum principles for our two problems but first, we need to recall a statement of an hybrid maximum principle as in [DK11].

### 3.2.1 Hybrid maximum principle.

First, we state the main result of [DK11], that we are going to use to prove both propositions. Let $t_0 < t_1 < \cdots < t_\nu$. Given a trajectory $x : [t_0,t_\nu] \to \mathbb{R}^n$, we define the vector

$$v = ((t_0,x(t_0));(t_1,x(t_1));\ldots;(t_\nu,x(t_\nu))) \in \mathbb{R}^{(\nu+1)(n+1)}.$$

Let us consider the hybrid optimal control problem ($\Omega$ is a subset of $\mathbb{R}^m$):

$$\begin{cases} J = \varphi_0(v) \to \min, & \\ \dot{x}(t) = f(t,x(t),u(t)) & u \in L^\infty([0;t_f],\Omega), \\ \eta_j(v) = 0 & j = 1,\ldots,p, \\ \varphi_i(v) \leqslant 0 & i = 1,\ldots,q, \end{cases}$$

Note that if $\nu = 1$, there are no intermediate constraints, and the problem can be solved using the classical Pontryagin maximum principle. The problem contains some equality and/or inequality constraints, including for instance constraints on the initial and final states like

$$x(t_0) - x_0 = 0\ ,\ x(t_f) - x_f = 0.$$

**Theorem 3.1 (Hybrid maximum principle).** – *Assume that $(\tilde{x}(\cdot), \tilde{u}(\cdot), \tilde{v})$ is an optimal solution of the previous hybrid optimal control problem. Then, there exists a tuple $(\alpha, \beta, \lambda_x(\cdot), \lambda_t(\cdot))$ where $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_q) \in \mathbb{R}^{q+1}$, $\beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ such that, if we define the applications*

$$H(t, x, u, \lambda_x, \lambda_t) = \langle \lambda_x, f(t, x, u) \rangle + \lambda_t,$$

$$l(v) = \sum_{i=0}^{q} \alpha_i \varphi_i(v) + \sum_{j=1}^{p} \beta_j \eta_j(v),$$

*then the following conditions hold:*

- $(\alpha, \beta) \neq 0$ ;
- *For all* $i \in [\![0, q]\!]$, $\alpha_i \geqslant 0$ ;
- *For all* $i \in [\![1, q]\!]$, $\alpha_i \varphi_i(\tilde{v}) = 0$ ;
- *Almost everywhere on* $[t_0, t_\nu]$,

$$\dot{\lambda}_x(t) = -\frac{\partial H}{\partial x}(t, \tilde{x}(t), \tilde{u}(t), \lambda_x(t), \lambda_t(t)),$$

$$\dot{\lambda}_t(t) = -\frac{\partial H}{\partial t}(t, \tilde{x}(t), \tilde{u}(t), \lambda_x(t), \lambda_t(t));$$

- *The transversality conditions at initial and final time stand:*

$$\lambda_x(t_0) = \tfrac{\partial l}{\partial x(t_0)}(\tilde{v}) \quad \lambda_x(t_\nu) = -\tfrac{\partial l}{\partial x(t_\nu)}(\tilde{v}),$$

$$\lambda_t(t_0) = \tfrac{\partial l}{\partial t_0}(\tilde{v}) \quad \lambda_t(t_\nu) = -\tfrac{\partial l}{\partial t_\nu}(\tilde{v});$$

- *At every intermediate point, one has the following discontinuity condition : for all* $k \in [\![1, \nu - 1]\!]$,

$$\lambda_x(t_k^+) - \lambda_x(t_k^-) = \frac{\partial l}{\partial x(t_k)}(\tilde{v}),$$

$$\lambda_t(t_k^+) - \lambda_x(t_k^-) = \frac{\partial l}{\partial t_k}(\tilde{v});$$

- *Almost everywhere on* $[t_0, t_\nu]$, $H(t, \tilde{x}(t), \tilde{u}(t), \lambda_x(t), \lambda_t(t)) = 0$;
- *The following maximisation condition holds:*

$$H(t, \tilde{x}(t), \tilde{u}(t), \lambda_x(t), \lambda_t(t)) = \max_{w \in \Omega} H(t, \tilde{x}(t), w, \lambda_x(t), \lambda_t(t)).$$

In [DK11], the proof of this result is given considering each part of the time interval $[t_k, t_{k+1}]$ for $k \in [\![1, \nu - 1]\!]$, and doing a transformation allowing to apply the usual PMP.

### 3.2.2 PMP for $(\mathcal{P})_{via,s}$ and $(\mathcal{P})_{pen,\varepsilon}$

In view of the following, let us define here the Hamiltonian:

$$H(t,x,u,p,p^0) = \langle p, f(t,x,u)\rangle + p^0 f^0(t,x,u).$$

We also recall that the initial point $x_0$ and the final point $x_f$ are fixed.

**Proposition 3.1 (PMP for $(\mathcal{P})_{via,s}$).** – *Let $(x(\cdot),u(\cdot))$ be a solution of $(\mathcal{P})_{via,s}$. Then there exists a non-trivial tuple $(p(\cdot),p^0,\beta)$, with $\beta\in\mathbb{R}^p$, such that:*

- $\dot{x}(t) = \frac{\partial H}{\partial p}(t,x(t),u(t),p(t),p^0)$;
- $\dot{p}(t) = -\frac{\partial H}{\partial x}(t,x(t),u(t),p(t),p^0)$;
- $H(t,x(t),u(t),p(t),p^0) = \max_{v\in\mathcal{U}} H(t,x(t),v,p(t),p^0)$ *a.e. on* $[0,t_f]$;
- *At time $t_1$, the adjoint vector presents a discontinuity:*

$$p(t_1^+) - p(t_1^-) = \frac{\partial g}{\partial x}(t_1,x(t_1))^T\beta;$$

- $\max_{v\in\Omega} H(t_f,x(t_f),v,p(t_f),p^0) = 0$.

**Proposition 3.2 (PMP for $(\mathcal{P})_{pen,\varepsilon}$).** – *Let $(x(\cdot),u(\cdot))$ be a solution of $(\mathcal{P})_{pen,\varepsilon}$. Then there exists a non-trivial tuple $(p(\cdot),p^0)$ such that:*

- $(p(\cdot),p^0)\neq(0,0)$;
- $\dot{x}(t) = \frac{\partial H}{\partial p}(t,x(t),u(t),p(t),p^0)$;
- $\dot{p}(t) = -\frac{\partial H}{\partial x}(t,x(t),u(t),p(t),p^0)$;
- $H(t,x(t),u(t),p(t),p^0) = \max_{v\in\mathcal{U}} H(t,x(t),v,p(t),p^0)$ *a.e. on* $[0,t_f]$;
- *At time $t_1$, the adjoint vector presents a discontinuity:*

$$p(t_1^+) - p(t_1^-) = -\frac{p^0}{\varepsilon}\cdot\frac{\partial g}{\partial x}(t_1,x(t_1))^T g(t_1,x(t_1));$$

- $\max_{v\in\Omega} H(t_f,x(t_f),v,p(t_f),p^0) = 0$.

*Proof of Proposition 3.2.* First, we start by rewriting $(\mathcal{P})_{pen,\varepsilon}$ in order to apply the hybrid maximum principle 3.1. Let us introduce the augmented system

$$\begin{cases} \dot{x}(t) = f(t,x(t),u(t)), & x(0)=x_0 \\ \dot{y}(t) = f^0(t,x(t),u(t)), & y(0)=0 \end{cases}$$

Let $v := ((x(0),y(0));(t_1,x(t_1),y(t_1));(t_f,x(t_f),y(t_f)))$. We also introduce the notation:

$$h(t_1,x(t_1)) := \frac{1}{\varepsilon}\left\|g(t_1,x(t_1))\right\|^2.$$

The cost can then be written under the form

$$J(v) = h(t_1,x(t_1)) + y(t_f).$$

Recall that the intermediate time $t_1$ is fixed, say to some $\tilde{t}_1$. We introduce the equality constraints:

$$\begin{aligned}
\eta_1(v) &:= x(0) - x_0 = 0,\\
\eta_2(v) &:= y(0) = 0,\\
\eta_3(v) &:= x(t_f) - x_f = 0,\\
\eta_4(v) &:= t_0 = 0,\\
\eta_5(v) &:= t_1 - \tilde{t}_1.
\end{aligned}$$

Assume that $(x(\cdot), y(\cdot), u(\cdot))$ is a solution of the augmented optimal control problem. Then, according to Theorem 3.1, there exists $(\alpha^0, \beta, \lambda_x(\cdot), \lambda_y(\cdot), \lambda_t(\cdot))$ with $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5) \in \mathbb{R}^5$, such that if we define the functions $H(x, u, \lambda_x, \lambda_y, \lambda_t) = \langle \lambda_x, f(x, u)\rangle + \lambda_y f^0(x, u) + \lambda_t$ and $l(v) = \alpha^0 J(v) + \beta_1\eta_1(v) + \beta_2\eta_2(v) + \beta_3\eta_3(v) + \beta_4\eta_4(v) + \beta_5\eta_5(v)$, we have

$$(\alpha^0, \beta) \neq 0, \tag{3.8}$$

$$\alpha^0 \geqslant 0. \tag{3.9}$$

The dynamics of the adjoint vector is given by

$$\dot{\lambda}_x = -\tfrac{\partial H}{\partial x}, \quad \dot{\lambda}_y = -\tfrac{\partial H}{\partial y} = 0, \quad \dot{\lambda}_t = -\tfrac{\partial H}{\partial t} = 0; \tag{3.10}$$

and we have the transversality condition at initial time

$$\lambda_x(t_0) = \tfrac{\partial l}{\partial x(t_0)}(v) = \beta_1, \quad \lambda_y(t_0) = \tfrac{\partial l}{\partial y(t_0)}(v) = \beta_2, \quad \lambda_t(t_0) = \tfrac{\partial l}{\partial t_0}(v) = \beta_4; \tag{3.11}$$

and at final time

$$\lambda_x(t_f) = -\frac{\partial l}{\partial x(t_f)}(v) = \beta_3, \tag{3.12}$$

$$\lambda_y(t_f) = -\frac{\partial l}{\partial y(t_f)}(v) = -\alpha^0\frac{\partial J}{\partial y(t_f)}(v) = -\alpha^0, \tag{3.13}$$

$$\lambda_t(t_f) = -\frac{\partial l}{\partial t_f}(v) = 0. \tag{3.14}$$

Finally, the discontinuity condition writes

$$\lambda_x(t_1^+) - \lambda_x(t_1^-) = \alpha^0\frac{\partial J}{\partial x(t_1)}(v) = \alpha_0\frac{\partial h}{\partial x(t_1)}(x(t_1)) = \frac{\alpha_0}{\varepsilon}\cdot\frac{\partial g}{\partial x}(t_1, x(t_1))^T g(t_1, x(t_1)) \tag{3.15}$$

$$\lambda_y(t_1^+) - \lambda_y(t_1^-) = \frac{\partial l}{\partial y(t_1)}(v) = 0; \tag{3.16}$$

$$\lambda_t(t_1^+) - \lambda_t(t_1^-) = \frac{\partial l}{\partial t_1}(v) = \beta_5; \tag{3.17}$$

(from which we get that $\beta_5 = -\beta_4$). Combining Equations (3.10), (3.13), (3.14), (3.16) and (3.17), we get that the function $\lambda_y$ is constant on $[0, t_f]$, $\lambda_y \equiv -\alpha^0$, and $\lambda_t$ is piecewise constant on $[0, t_f]$, satisfying

$$\begin{cases} \lambda_t \equiv \beta_4 & \text{on } [0, t_1] \\ \lambda_t \equiv 0 & \text{on } [t_1, t_f] \end{cases}$$

Let us set $p^0 := -\alpha^0$. We get, exploiting the discontinuity condition (3.15), the jump on the adjoint vector :

$$\lambda_x(t_1^+) - \lambda_x(t_1^-) = -p^0 \nabla h(x(t_1)).$$

We obtain Proposition 3.2 by setting $p(t) = \lambda_x(t)$. □

*Proof of Proposition 3.1.* The sketch of the proof is similar to the previous one. We use the same trick of considering the augmented system

$$\begin{cases} \dot{x}(t) = f(t, x(t), u(t)), & x(0) = x_0, \\ \dot{y}(t) = f^0(t, x(t), u(t)), & x(t_f) = x_f. \end{cases}$$

Let $v = ((t_0, x(t_0), y(t_0)); (t_1, x(t_1), y(t_1)); (t_f, x(t_f), y(t_f)))$. $(\mathcal{P})_{via,s}$ consists in minimizing the cost

$$J(v) = y(t_f)$$

under the following constraints:

$$\begin{aligned} \eta_0(v) &= g(x(t_1)) - sg(\overline{x}(t_1)), \\ \eta_1(v) &= x(0) - x_0 = 0, \\ \eta_2(v) &= y(0) = 0, \\ \eta_3(v) &= x(t_f) - x_f = 0, \\ \eta_4(v) &:= t_0 = 0, \\ \eta_5(v) &:= t_1 - \tilde{t}_1. \end{aligned}$$

Let $(x(\cdot), y(\cdot), u(\cdot))$ be a solution of this optimization problem. Then, applying Theorem 3.1, there exists a tuple

$$(\alpha^0, \beta, \lambda_x(\cdot), \lambda_y(\cdot), \lambda_t(\cdot)),$$

with $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5) \in \mathbb{R}^p \times \mathbb{R}^5$ such that, if we define the function $H$ as in the previous proof and $l$ by:

$$l(v) = \alpha^0 J(v) + \langle \beta_0, \eta_0(v) \rangle + \beta_1 \eta_1(v) + \beta_2 \eta_2(v) + \beta_3 \eta_3(v) + \beta_4 \eta_4(v) + \beta_5 \eta_5(v),$$

we have

$$(\alpha^0, \beta) \neq 0, \quad \alpha^0 \geqslant 0.$$

The dynamics of the adjoint vector is given by

$$\dot{\lambda}_x = -\tfrac{\partial H}{\partial x}, \quad \dot{\lambda}_y = -\tfrac{\partial H}{\partial y} = 0, \quad \dot{\lambda}_t = -\tfrac{\partial H}{\partial t} = 0$$

The transversality conditions at initial and final time are the same as in the previous proof, and the jump on the adjoint vector is given by

$$\begin{aligned} \lambda_x(t_1^+) - \lambda_x(t_1^-) &= \frac{\partial l}{\partial x(t_1)}(v) = dg(x(t_1))^T \cdot \beta_0 \\ \lambda_y(t_1^+) - \lambda_y(t_1^-) &= \frac{\partial l}{\partial y(t_1)}(v) = 0 \end{aligned}$$

and we also have that $\lambda_t(t_1^+) - \lambda_t(t_1^-) = \beta_5$. One can then conclude the proof as before, the function $\lambda_y$ being constant, and $\lambda_t$ being piecewise constant. Letting $p^0 := -\alpha^0$ and $\beta := \beta_0 \in \mathbb{R}^p$, we get the formulation of Proposition 3.1. □

**Remark 3.1: Via-point constraint.**

An important case in practice is when the intermediate constraint consists in prescribing some components of the state $x$ at time $t_1$. Let $P : \mathbb{R}^n \to \mathbb{R}^p$ (with $p \leqslant n$) be a projection such that $P(x_1, \dots, x_n) = (x_{\sigma(1)}, \dots, x_{\sigma(p)})$, where $\sigma$ is a permutation of $\{1, \dots, n\}$. In that particular case, the intermediate constraint can be written

$$P(x(t_1)) = y_{via},$$

where $y_{via}$ is some fixed point in $\mathbb{R}^p$, and is sometimes referred to as a *via-point constraint*. In this situation, the constraint in $(\mathcal{P})_{via,s}$ writes

$$P(x(t_1)) = sP(\overline{x}(t_1)) + (1-s)y_{via},$$

and the cost in $(\mathcal{P})_{pen,\varepsilon}$ writes

$$J_\varepsilon(u) = \int_0^{t_f} f^0(t, u(t), x(t))\, dt + \|P(x(t_1)) - y_1\|^2 / \varepsilon.$$

Besides, the jump on the adjoint vector in Proposition 3.1 becomes (component-wise)

$$\begin{array}{ll} p_{\sigma(i)}(t_1^+) - p_{\sigma(i)}(t_1^-) = \beta_i & \text{for all } i \in [\![1, p]\!], \\ p_{\sigma(i)}(t_1^+) - p_{\sigma(i)}(t_1^-) = 0 & \text{for all } i \in [\![p+1, n]\!], \end{array}$$

and the jump in Proposition 3.2 becomes

$$\begin{array}{ll} p_{\sigma(i)}(t_1^+) - p_{\sigma(i)}(t_1^-) = -2p^0(x_{\sigma(i)} - y_i)/\varepsilon & \text{for all } i \in [\![1, p]\!], \\ p_{\sigma(i)}(t_1^+) - p_{\sigma(i)}(t_1^-) = 0 & \text{for all } i \in [\![p+1, n]\!]. \end{array}$$

A variant is to choose penalization parameters $\varepsilon_i$ depending on the indices under consideration. Here, for simplicity, we keep the same penalization parameter $\varepsilon$ for all indices.

### 3.2.3 Shooting functions for $(\mathcal{P})_{via,s}$ and $(\mathcal{P})_{pen,\varepsilon}$

Propositions 3.1 and 3.2 state that the optimal solutions of the problems $(\mathcal{P})_{via,s}$ and $(\mathcal{P})_{pen,\varepsilon}$ must be sought over the set of trajectories satisfying the necessary conditions of the Pontryagin maximum principle. We will now explain in detail how it can be reduced to finding the zeros of some shooting function.

**Shooting function for $(\mathcal{P})_{pen,\varepsilon}$.** In Proposition 3.2 the maximisation condition implies that, under some conditions [1], the control can be written as a function of the time, the state $x$ and the costate $p$ : $u(t) = u(t, x(t), p(t))$. Let us denote $z = (x, p)$. The dynamics of $z$ can therefore be written under the form $\dot{z}(t) = F(t, z(t))$. Let $z(t, z_0) = (x(t, z_0), p(t, z_0))$ be the solution of the Cauchy problem $\dot{z}(t) = F(t, z(t))$ with the initial condition $z(0, z_0) = z_0$ and a jump at time

1. A usual assumption is to assume that a Legendre condition is satisfied, namely that the hessian matrix $\frac{\partial^2 H}{\partial u^2}(t, x, u, p, p^0)$ is negative definite. Such a condition enables to express the control (at least locally), as a function of $x$ and $p$.

$t_1$ given by

$$x(t_1^+, z_0) - x(t_1^-, z_0) = 0,$$
$$p(t_1^+, z_0) - p(t_1^-, z_0) = -\frac{p^0}{\varepsilon} \cdot \frac{\partial g}{\partial x}(t_1, x(t_1))^T g(t_1, x(t_1)).$$

For short, let us denote $H(t_f) = \max_{v \in \mathcal{U}} H(t_f, v, z(t_f, (x_0, p(0))), p^0)$ and let us define the function

$$G_\varepsilon : \left\{ \begin{array}{ccc} \mathbb{R}^n \times \mathbb{R} & \to & \mathbb{R}^{n+1} \\ (p(0), t_f) & \mapsto & \left[ \begin{array}{c} x(t_f, (x_0, p(0))) - x_f \\ H(t_f) \end{array} \right] \end{array} \right. \tag{3.18}$$

Finding a trajectory satisfying the necessary conditions of Pontryagin maximum principle boils down to finding a zero of the function $G_\varepsilon$, that is an initialization of the costate $p(0)$ and a final time $t_f$ ($n+1$ unknowns) such that the terminal condition $x(t_f) = x_f$ and the transversality condition $H(t_f, x(t_f), u(t_f), p(t_f), p^0) = 0$ are satisfied ($n+1$ equations).

**Shooting function for $(\mathcal{P})_{via,s}$** . Note that in the Pontryagin maximum principle for $(\mathcal{P})_{pen,\varepsilon}$, the jump at time $t_1$ is given by

$$p(t_1^+) - p(t_1^-) = -\frac{p^0}{\varepsilon} \cdot \partial_x g(t_1, x(t_1))^T g(t_1, x(t_1)) \, .$$

Hence, once the initialization of the costate $p(0)$ is made, the dynamics of $z = (x, p)$ is determined up to the final time.

In $(\mathcal{P})_{via,s}$, the jump at time $t_1$ is given by $p(t_1^+) - p(t_1^-) = dg(t_1, x(t_1))^T \beta$, where $\beta \in \mathbb{R}^p$ is a new unknown of the problem. However, there are also $p$ additional equations to fulfill to find a trajectory satisfying Pontraygin's necessary conditions, namely

$$g(t_1, x(t_1)) = sg(t_1, \overline{x}(t_1)).$$

As explained in Chapter 2, solving an optimal control problem by an indirect method boilds down to finding the zeros of a shooting function. In that case, the shooting problem consists in finding a zero of a shooting function $G_s$. More precisely, it consists in finding an initialization of the costate $p(0)$, a final time $t_f$ and a vector $\beta \in \mathbb{R}^p$ ($n+1+p$ unknowns) such that $x(t_f) = x_f$, the transversality condition $H(t_f) = 0$ and the intermediate constraint $g(t_1, x(t_1)) = sg(t_1, \overline{x}(t_1))$ are satisfied ($n+1+p$ equations).

## 3.3 Application to the attitude control of a rigid body

### 3.3.1 The attitude control problem

Let us recall first the attitude equations for a rigid body, as expressed in the Introduction

$$\left\{ \begin{array}{rcl} \dot{\theta}(t) & = & \frac{\sin\varphi(t)}{\cos\psi(t)} q(t) + \frac{\cos\varphi(t)}{\cos\psi(t)} r(t) \\ \dot{\psi}(t) & = & \cos\varphi(t) \cdot q(t) - \sin\varphi(t) \cdot r(t) \\ \dot{\varphi}(t) & = & p(t) + \sin\varphi(t) \tan\psi(t) \cdot q(t) + \cos\varphi(t) \tan\psi(t) \cdot r(t) \\ \dot{p}(t) & = & a_1 q(t) r(t) + \sum_{j=1}^m u_j(t) b_j^1 \\ \dot{q}(t) & = & a_2 p(t) r(t) + \sum_{j=1}^m u_j(t) b_j^2 \\ \dot{r}(t) & = & a_3 p(t) q(t) + \sum_{j=1}^m u_j(t) b_j^3. \end{array} \right.$$

In what follows, we will denote $\omega$ the (euclidian) norm of the angular velocity vector $\overrightarrow{\omega} = (p, q, r)$. Therefore $\omega$ is zero if and only if the three components $p$, $q$ and $r$ are zero.

Our goal is to steer the system from an initial state $x_0$ to a final state $x_f$ while minimizing a combination of the fuel consumption and the final time

$$J(u) = \int_0^{t_f} \sum_{j=1}^{m} |u_j(t)|\, dt + \frac{t_f}{2}, \tag{3.19}$$

and cancelling the angular velocity at some fixed intermediate time $t_1$, i.e., $\omega(t_1) = 0$. We will explain at the beginning of Section 3.4 why this may be of interest in practice. Note that in that example, the constraint writes as a via-point constraint as in Remark 3.1.

### 3.3.2 Continuation procedure

**Computation of $(\overline{x}(\cdot), \overline{u}(\cdot))$.** A first difficulty, not mentionned so far, is that the resolution of $(\mathcal{P})_0$ by an indirect method can already be hard. For instance, when considering a $L^1$ cost, as in (3.19), the underlying shooting function is known to have a very small domain of convergence, as explained in Chapter 2.

In chapter 2, we introduced the continuation parameter $\alpha \in [0, 1]$, and for each $\alpha \in [0, 1]$, we defined the cost

$$\alpha \int_0^{t_f} \sum_{j=1}^{m} u_j(t)^2\, dt + (1-\alpha) \int_0^{t_f} \sum_{j=1}^{m} |u_j(t)|\, dt + \frac{t_f}{2}$$

When $\alpha = 0$, one recognizes the cost (3.19). When $\alpha = 1$, the cost is stricly convex in the controls, and writes

$$\int_0^{t_f} \sum_{j=1}^{m} u_j(t)^2\, dt + \frac{t_f}{2},$$

for which the initialization of the induced shooting method is much easier, see Chapter 2 for more details on this issue.

We perform a first continuation, solving a sequence of optimal control problems, for values of $\alpha$ decreasing from 1 to 0.

**Resolution of $(\mathcal{P})_{via,0}$.** Once $(\overline{x}(\cdot), \overline{u}(\cdot))$, a solution of $(\mathcal{P})_0$, is computed with the first continuation on the parameter $\alpha$, it can be used to initialize the second continuation on $\varepsilon$, considering the penalized cost

$$J_\varepsilon(u) = \int_0^{t_f} \sum_{j=1}^{m} |u_j(t)|\, dt + \frac{t_f}{2} + \frac{1}{\varepsilon} \left\| \omega(t_1) \right\|^2, \tag{3.20}$$

starting from a high value $\varepsilon_{\max}$, and decreasing progressively $\varepsilon$ until we reach a threshold $\varepsilon_{\min}$. For each $\varepsilon \in [\varepsilon_{\min}, \varepsilon_{\max}]$, the resolution of $(\mathcal{P})_{pen,\varepsilon}$ is done by finding a zero of the shooting function $G_\varepsilon$ of Equation (3.18).

Once $(\mathcal{P})_{pen,\varepsilon_{\min}}$ is solved, provided $\varepsilon_{\min}$ is small enough, it provides a good enough initialization to tackle the original problem of interest $(\mathcal{P})_{via,0}$, and we end up our procedure by solving a last shooting problem with the continuation parameter $s$ taken equal to 0.

Figure 3.1 summarizes the numerical procedure we just described.

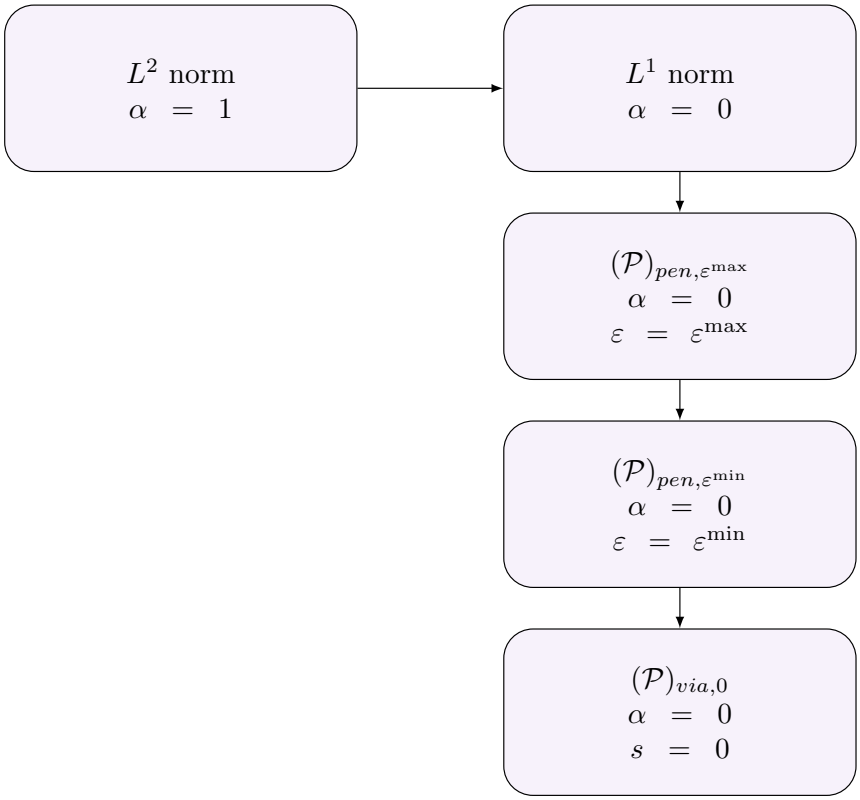



Figure 3.1 – Continuation procedure to solve $(\mathcal{P})_{via,0}$.

**Remark 3.2: Order of the continuations**

The procedure schematized on Figure 3.1 performs first a continuation on the parameter $\alpha$ and then a continuation on the parameter $\varepsilon$. We also tried to perform first the continuation on $\varepsilon$ and then on $\alpha$, but experimentally observed that the whole procedure was slower to finish.

**A convergence result.** To justify that our procedure is theoretically sound, we give now a convergence result. Namely, we show that, for the attitude control problem with the $L^1$ cost (3.19), the solutions of $(\mathcal{P})_{pen,\varepsilon}$ converge to a solution of $(\mathcal{P})_{via,0}$ when $\varepsilon$ goes to 0.

**Proposition 3.3.** – *Assume that $(\mathcal{P})_{via,0}$ has a unique solution $(\tilde{x},\tilde{u})$ defined on the time interval $[0,\tilde{t}_f]$. Let $(u^\varepsilon,x^\varepsilon)$ be a sequence of solutions for $(\mathcal{P})_{pen,\varepsilon}$, defined on $[0,t_f^\varepsilon]$. Then, when $\varepsilon$ goes to 0,*

- *$t_f^\varepsilon$ converges to $\tilde{t}_f$,*
- *$u^\varepsilon$ converges weakly to $\tilde{u}$,*
- *$x^\varepsilon$ converges uniformly to $\tilde{x}$*

*Proof.* First, we show that the sequence $t_f^\varepsilon$ is bounded. By optimality of the trajectory $(u^\varepsilon,x^\varepsilon)$, one has

$$J_\varepsilon(u^\varepsilon) \leqslant J_\varepsilon(\tilde{u}),$$

that is,

$$\int_0^{t_f} \sum_{j=1}^{m} |u_j^\varepsilon(t)|\,dt + \frac{t_f^\varepsilon}{2} + \frac{1}{\varepsilon}\left\|\omega^\varepsilon(t_1)\right\|^2 \leqslant \int_0^{t_f} \sum_{j=1}^{m} |\tilde{u}_j(t)|\,dt + \frac{\tilde{t}_f}{2} + \frac{1}{\varepsilon}\left\|\tilde{\omega}(t_1)\right\|^2.$$

Exploiting the fact that $\tilde{\omega}(t_1) = 0$, we get that

$$\int_0^{t_f} \sum_{j=1}^{m} |u_j^\varepsilon(t)|\,dt + \frac{t_f^\varepsilon}{2} + \frac{1}{\varepsilon}\left\|\omega^\varepsilon(t_1)\right\|^2 \leqslant \int_0^{t_f} \sum_{j=1}^{m} |\tilde{u}_j(t)|\,dt + \frac{\tilde{t}_f}{2}. \tag{3.21}$$

Hence, the sequence $t_f^\varepsilon$ is bounded, and up to a subsequence, it converges to some $T$.

The sequence $(u^\varepsilon)$ is bounded in $L^\infty([0,T],[0,1]^m)$ (if $t_f^\varepsilon \leqslant T$, we extend $u^\varepsilon$ to 0 on the interval $[t_f^\varepsilon, T]$, if $t_f^\varepsilon > T$, we restrict $u^\varepsilon$ to $[0,T]$), and therefore, up to a subsequence, $(u^\varepsilon)$ converges weakly in $L^2([0,T],[0,1]^m)$ to some control $u^\star$. For all $\varepsilon > 0$, $u^\varepsilon$ belongs to the set

$$\mathcal{V} = \left\{v \in L^2([0,T],\mathbb{R}^m) \text{ s.t. } \forall i \in [\![1,m]\!],\ v_i(\cdot) \in [0,1] \text{ a.e.}\right\}$$

This set is strongly closed and convex, and is therefore weakly closed. Thus, $u^\star \in \mathcal{V}$, and is admissible for the system (3.1). Let us denote $x^\star$ its associated trajectory. It is a classical result (see for instance [Tré00]) that for a control-affine system, if a control sequence $(u^\varepsilon)$ converges weakly in $L^2([0,T],\mathbb{R}^m)$ to a control $u^\star$, then the associated sequence of trajectories $(x^\varepsilon)$ converges uniformly to $x^\star$, associated to $u^\star$.

Besides, for all $\varepsilon > 0$, we have $x_\varepsilon(0) = x_0$ and $x_\varepsilon(t_f^\varepsilon) = x_f$, hence, taking the limit when $\varepsilon$ goes to 0, we have

$$x^\star(0) = x_0, \quad x^\star(T) = x_f.$$

From (3.21), we also get

$$0 \leqslant \left\|\omega^\varepsilon(t_1)\right\|^2 \leqslant \varepsilon\left(\int_0^{t_f} \sum_{j=1}^{m} |\tilde{u}_j(t)|\,dt + \frac{\tilde{t}_f}{2}\right),$$

and taking the limit when $\varepsilon$ goes to 0, we get that $\left\|\omega^\star(t_1)\right\|^2 = 0$. Hence, $(u^\star, x^\star)$ is a solution of $(\mathcal{P})_{via,0}$, and by uniqueness, we have $u^\star = \tilde{u}$, $x^\star = \tilde{x}$ and $T = \tilde{t}_f$.

□

Actually, when initializing the resolution of the optimal control $(\mathcal{P})_{via,0}$, we also use the adjoint vector $p^\varepsilon(\cdot)$ coming from the resolution of $(\mathcal{P})_{pen,\varepsilon}$ when $\varepsilon$ goes to zero to initialize the initial value $p(0)$ and the Lagrange multiplier $\beta$. Indeed, as explained previously, the shooting function for $(\mathcal{P})_{via,s}$, $G_{s=0}$, takes a Lagrange multiplier $\beta$ as an argument. Following Remark 3.1, we use the heuristic that an approximation for $\beta_i$ can be $-2p^0(x^\varepsilon_{\sigma(i)} - y_i)/\varepsilon$.

Of course, in order to have a complete justification of the procedure, we should also show a convergence property for the adjoint vector $p^\varepsilon(\cdot)$ and for the jump at time $t_1$. This study is more difficult than the proof of Proposition 3.3, and has yet to be undertaken.

## 3.4 Numerical results

In this section, we are going to illustrate the procedure previously introduced. For the sake of continuity, we will take the same numerical values that in the last section of Chapter 2, namely

$$(\theta_0, \psi_0, \varphi_0, p_0, q_0, r_0) = (0.04, 0.06, 7.7, -0.027, 0, 0),$$
$$(\theta_f, \psi_f, \varphi_f, p_f, q_f, r_f) = (0.63, 0.82, 7.0, -0.008, 0, 0).$$

We also chose the time $t_1$ to be equal to 8.5 seconds. Indeed, in practice, the state of the launcher at the beginning of the ballistic flight is inherited from the previous phases of the flight, and the angular velocity may take high values. It can therefore be useful to start the ballistic phase by controlling the three angular velocities to zero, while letting the orientation angles $\theta$, $\psi$ and $\varphi$ evolve freely.

On Figure 3.2, we display the evolution of the angular velocity (in degrees per second) at different stages of the procedure. In black, we plot the angular velocity for the solution of $(\mathcal{P})_{pen,\varepsilon^{\max}}$. As $\varepsilon^{\max}$ (we started at $\varepsilon^{\max} = 100$) is chosen large enough, it is so close to the $L^1$-optimal angular velocity $\overline{\omega}$ that both curves would overlap. In dark red, we show the angular velocity at the end of the continuation on $\varepsilon$, and with a dotted style, the angular velocity at some intermediate stage of the continuation on $\varepsilon$. When $\varepsilon$ reaches $\varepsilon^{\min}$ (taken equal to $2 \times 10^{-6}$), the solution of $(\mathcal{P})_{pen,\varepsilon^{\min}}$ provides a good enough starting point to initialize the shooting problem $(\mathcal{P})_{s,0}$ with success. The angular velocity for $(\mathcal{P})_{s,0}$ is plotted in light red. Note that except during the first seconds, the curves corresponding to $(\mathcal{P})_{pen,\varepsilon^{\min}}$ and $(\mathcal{P})_{s,0}$ are almost indistinguishable.

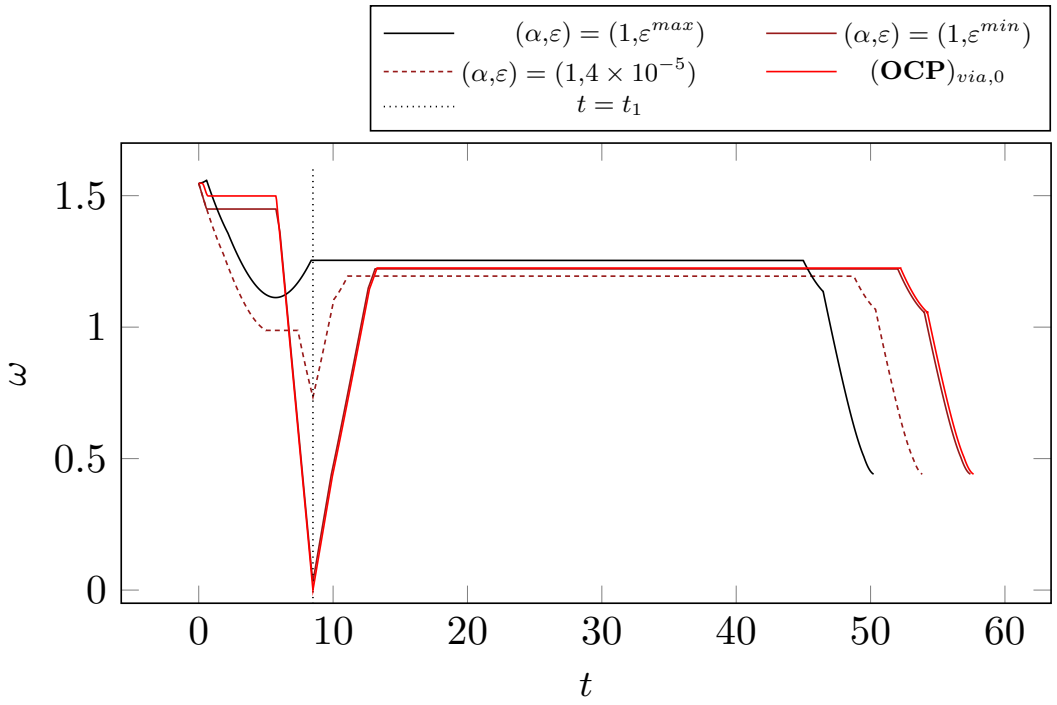



Figure 3.2 – Evolution of the angular velocity during the continuation.

On Figure 3.3, we also display the evolution of the 6 components of the optimal solution of $(\mathcal{P})_{s,0}$, and on Figure 3.4 we show the associated control. Note that compared to the controls for the resolution of (**OCP**) (represented on Figure 2.6), the number of switching times has increased. They were 6 switchings for the controls corresponding to the resolution of (**OCP**), and 16 switchings for the controls of Figure 3.4. This is one of the numerical difficulty in the resolution of $(\mathcal{P})_{s,0}$, as switching times can be hard to catch with an indirect method.

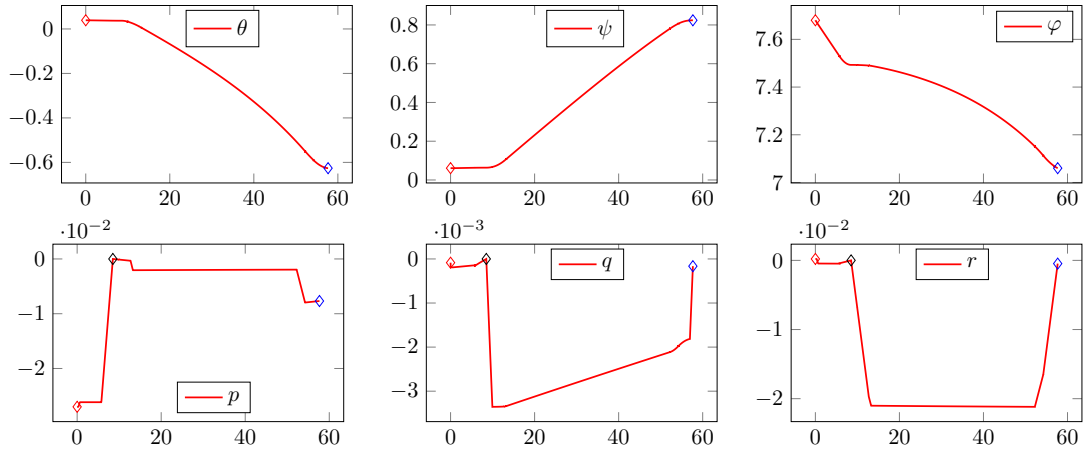



Figure 3.3 – Solution of $(\mathcal{P})_{via,0}$, steering the system from $x_0$ (◇) from $x_f$ (◇), satisfying a via-point constraint (◇).

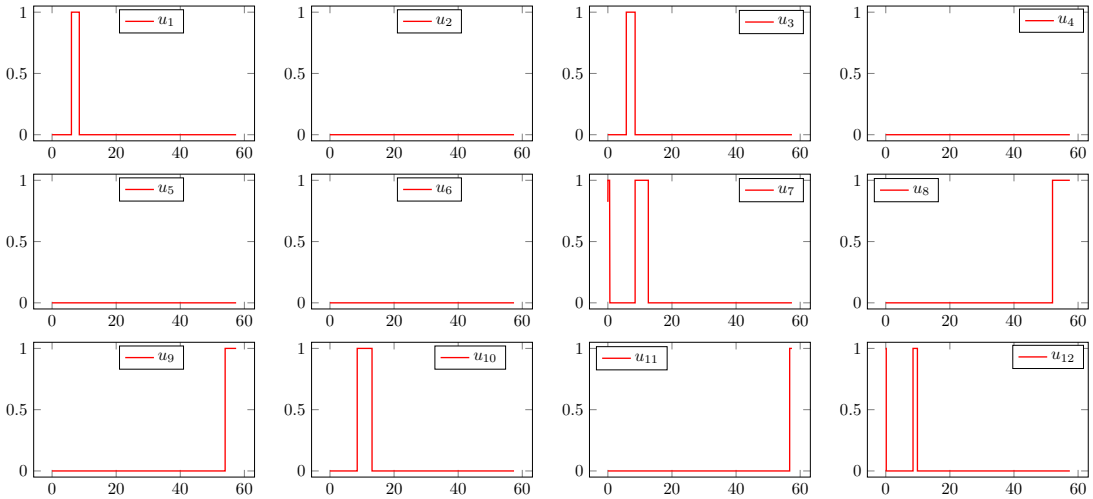



Figure 3.4 – Controls for the resolution of $(\mathcal{P})_{via,0}$.

**Numerical difficulty.** The numerical procedure described so far, relying on the combination of continuation techniques and indirect methods, allows us to solve with high accuracy the problem $(\mathcal{P})_{s,0}$. We emphasize again that solving numerically this problem would be too hard to be tackled directly.

We tried to apply the same procedure to an attitude control problem with more than one intermediate constraint: at times $t_1$ and $t_2$, we wish to enforce constraints under the form

$$g_1(t_1, x(t_1)) = 0,\ g_2(t_2, x(t_2)) = 0.$$

A natural generalization of our approach would be to penalize the constraints in the cost

$$\int_0^{t_f} f^0(t, u(t), x(t))\, dt + \frac{1}{\varepsilon_1} \|g_1(t_1, x(t_1))\|^2 + \frac{1}{\varepsilon_2} \|g_2(t_2, x(t_2))\|^2,$$

and let the pair $\varepsilon = (\varepsilon_1, \varepsilon_2)$ go to zero.

However, because of the increasing number of switching times arising when enforcing such constraints, even the continuation on $\varepsilon$ sometimes failed to reach a satisfactory level.

In order to be able to generically optimize any complete ballistic flight, it was crucial for the CNES to have a tool dealing with any given number of intermediate constraints. This is the focus of Appendix A, where we give the description of a software, designed and implemented for the CNES, combining a direct method with an interior-point algorithm.

## 3.5 Conclusion of this chapter

No matter the way the intermediate constraint is taken into account, the maximum principle on which the indirect method relies states that the adjoint vector is discontinuous at the time of the constraint. A continuation procedure has been designed in order to exactly enforce the constraint. The procedure benefits from the high accuracy of the underlying Newton method. However, when we experimentally tried to apply the procedure to the attitude control problem with more than one via-point constraint, the aforementioned procedure sometimes failed to converge.

For this reason, in the software that we designed for the CNES (see Appendix A), we proceeded in a slightly different way by combining direct methods with an interior-point algorithm able to tackle any given ballistic phase where the number of via-point constraints is up to the choice of the user.

Note that in the theoretical study we undertook in Section 3.2 and Section 3.3, two open problems remain:

- It is still unclear why the continuation procedure on the parameter $\varepsilon$ gives much better results than the continuation on $s$.
- The convergence result of Proposition 3.3 is conjectured to be true as well for the sequence of adjoint vectors $(p^\varepsilon)_{\varepsilon>0}$.

These two issues are left open.

Chapter 4

# Redundancy implies robustness for bang-bang control strategies

## Contents


**3.1 Introduction of the chapter . . . . . 52**

**3.2 Optimal control formulation . . . . . 54**

3.2.1 Hybrid maximum principle. . . . . 54

3.2.2 PMP for $(\mathcal{P})_{via,s}$ and $(\mathcal{P})_{pen,\varepsilon}$ . . . . . 56

3.2.3 Shooting functions for $(\mathcal{P})_{via,s}$ and $(\mathcal{P})_{pen,\varepsilon}$ . . . . . 59

**3.3 Application to the attitude control of a rigid body . . . . . 60**

3.3.1 The attitude control problem . . . . . 60

3.3.2 Continuation procedure . . . . . 61

**3.4 Numerical results . . . . . 64**

**3.5 Conclusion of this chapter . . . . . 66**


In the previous chapters, we applied optimal control theory to several attitude control problems. In those chapters, the dynamics of the state was given by a differential equation coming from physical laws that did not account for the presence of uncertainties, model errors or perturbations.

However, in view of aerospace applications, being able to design a control system that deals with uncertainties is of crucial importance. It is the main concern of this chapter, where we give an algorithm to control a system even with deviations from the target, identify a criterion to measure the robustness of a control and suggest a way to make a nominal control more robust. Note that our approach applies to any given nonlinear control system with a cost resulting in a bang-bang control.

## 4.1 Introduction of the chapter

### 4.1.1 Overview of the method

To introduce the subject, we explain our approach on the control problem consisting of steering the finite-dimensional nonlinear control system

$$\dot{x}(t) = f(t, x(t), u(t)), \tag{4.1}$$

from a given $x(0) = x_0$ to the target point $x(t_f) = x_f$, with a scalar control $u$ that can only switch between two values, say 0 and 1. The general method, as well as all assumptions, will be written in details in a further section.

Let $E(x_0, t_f, u) = x(t_f)$ be the end-point mapping, where $x(\cdot)$ is the solution of (4.1) starting at $x(0) = x_0$ and associated with the control $u$. One aims at finding a bang-bang control $u$, defined on $[0, t_f]$ for some final time $t_f > 0$, such that $E(x_0, t_f, u) = x_f$.

Many problems impose to implement only bang-bang controls, i.e., controls saturating the constraints but not taking any intermediate value. These are problems where only external actions of the kind on/off can be applied to the system.

Of course, such bang-bang controls can usually be designed by using optimal control theory (see [LM67b, PBGM62, Tré05b]). For instance, solving a minimal time control problem, or a minimal $L^1$ norm as in [CFPT13], is in general a good way to design bang-bang control strategies. However, due to their optimality status, such controls often suffer from a lack of robustness with respect to uncertainties, model errors, deviations from the target. Moreover, when the Pontryagin maximum principle yields bang-bang controls, such controls have in general a minimal number of switchings: in dimension 3 for instance, it is proved in [KS89, Kup87, Sch88] (see also [BC03b, BFT05, Tré12] for more details on this issue) that, locally, minimal time trajectories of single-input control-affine systems have generically two switchings. Taking into account the free final time, this makes three degrees of freedom, which is the minimal number to generically make the trajectory reach a target point in $\mathbb{R}^3$, i.e., to solve three (nonlinear) equations.

In these conditions, a natural idea is to add redundancy to such bang-bang strategies, by enforcing the control to switch more times than necessary. These additional switching times are introduced by *needle-like variations*, as in the classical proof of the Pontryagin maximum principle (see [LM67b, PBGM62]).

We recall that a needle-like variation $\pi_1 = (t_1, \delta t_1, u_1)$ of a given control $u$ is the perturbation $u_{\pi_1}$ of the control $u$ given by

$$u_{\pi_1}(t) = \begin{cases} u_1 & \text{if} \quad t \in [t_1, t_1 + \delta t_1], \\ u(t) & \text{otherwise,} \end{cases} \tag{4.2}$$

where $t_1 \in [0, t_f]$ is the time at which the spike variation is introduced, $\delta t_1$ is a real number of small absolute value that stands for the duration of the variation, and $u_1 \in [0, 1]$ is some arbitrary element of the set of values of controls. When $\delta t_1 < 0$, one replaces the interval $[t_1, t_1 + \delta t_1]$ with $[t_1 + \delta t_1, t_1]$ in (4.2). We represent on the Figure 4.1 a needle-like variation.

It is well known that, if $|\delta t_1|$ is small enough, the control $u_{\pi_1}$ is admissible (that is, the associated trajectory solution of (4.1) is well-defined on $[0, t_f]$) and generates a trajectory $x_{\pi_1}(\cdot)$, which can be viewed as a perturbation of the nominal trajectory $x(\cdot)$ associated with the control $u$, and which steers the control system to the final point

$$E(x_0, t_f, u_{\pi_1}) = E(x_0, t_f, u) + |\delta t_1|\, v_{\pi_1}(t_f) + o(\delta t_1), \tag{4.3}$$

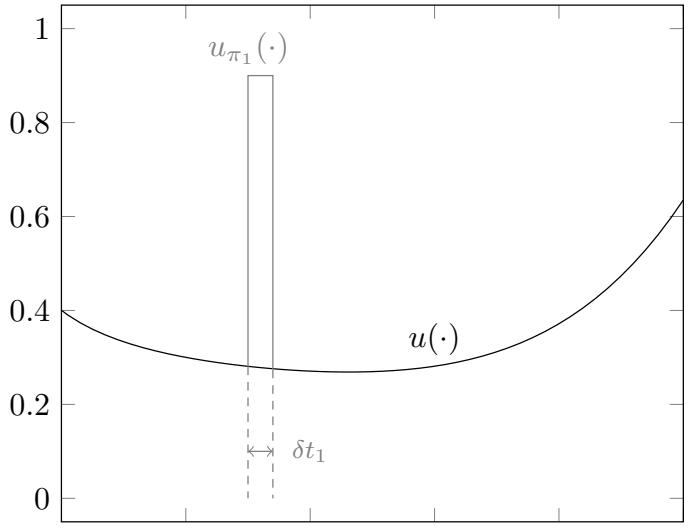



Figure 4.1 – Needle-like variation $u_{\pi_1}$ of a control $u$.

where the so-called variation vector $v_{\pi_1}(\cdot)$ is the solution of some Cauchy problem related to a linearized system along $x(\cdot)$ (see [LM67b, PBGM62, ST10b] and Proposition 4.1). Recall that the *first Pontryagin cone* $K(t_f)$ is the smallest closed convex cone containing all variation vectors $v_{\pi_1}(t_f)$; it serves as a local convex estimate of the set of reachable points at time $t_f$ (with initial point $x_0$).

Assume that the nominal control $u$, which steers the system from $x_0$ to the target point $x_f$, is bang-bang and switches $N$ times between the extreme values 0 and 1 over the time interval $[0, t_f]$. We denote by $\mathcal{T} = (t_1, \ldots, t_N)$ the vector consisting of its switching times $0 < t_1 < \cdots < t_N < t_f$. Then the control $u$ can equivalently be represented by the vector $\mathcal{T}$, provided one makes precise the value of $u(t)$ for $t \in (0, t_1)$. One can also add new switching times: for instance if $u(t) = 0$ for $t \in (0, t_1)$, given any $s_1 \in (0, t_1)$, the needle-like variation $\pi_1 = (s_1, \delta s_1, 1)$ (with $|\delta s_1|$ small enough) is a bang-bang control having two new switching times at $s_1$ and $s_1 + \delta s_1$.

In what follows, we designate a bang-bang control either by $u$ or by the set $\mathcal{T} = (t_1, \ldots, t_N)$ of its switching times. This is with a slight abuse because we should also specify the value of $u$ along the first bang arc. But we will be more precise, rigorous and general in a further section. The end-point mapping is then reduced to the switching times, and one has $E(x_0, t_f, \mathcal{T}) = x_f$. A variation $\delta\mathcal{T} = (\delta t_1, \ldots, \delta t_N)$ of the switching times generates $N$ variation vectors $(v_1(t_f), \ldots, v_N(t_f))$, and the corresponding bang-bang trajectory reaches at time $t_f$ the point (see Figure 4.2, where two variations vectors are displayed, for two switching times $t_1$ and $t_2$)

$$E\left(x_0, t_f, \mathcal{T} + \delta\mathcal{T}\right) = x_f + \delta t_1 \cdot v_1(t_f) + \cdots + \delta t_N \cdot v_N(t_f) + \mathrm{o}(\|\delta\mathcal{T}\|).$$

Therefore the end-point mapping $E$ is differentiable with respect to $\mathcal{T}$, and

$$\frac{\partial E}{\partial \mathcal{T}}(x_0, t_f, \mathcal{T}) \cdot \delta\mathcal{T} = \delta t_1 \cdot v_1(t_f) + \cdots + \delta t_N \cdot v_N(t_f). \tag{4.4}$$

Notice that compared to (4.3), the absolute values disappear. We will prove this result in details further in the chapter. In particular, the range of this differential is the first Pontryagin cone

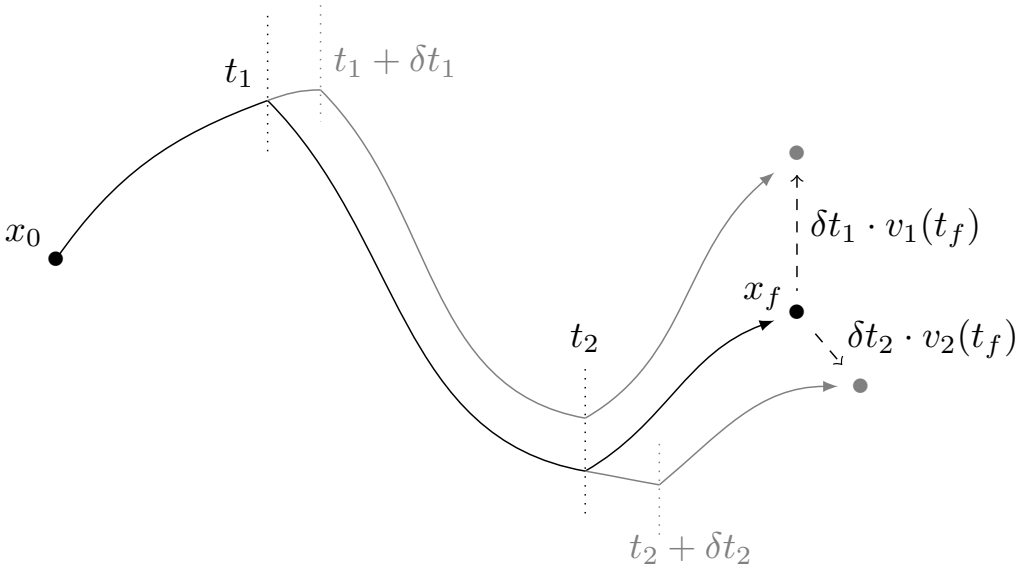



Figure 4.2 – Changing the switching times induces a displacement at the final time.

$K(t_f)$ (see also [ST10b]). Obviously, the more switching times (i.e., degrees of freedom), the more accurate the approximation of the reachable set.

We now add *redundant* switching times $(s_1, \dots, s_\ell)$ for some $\ell \in \mathbb{N}$ in order to generate more degrees of freedom to solve the control problem

$$E\left(x_0, t_f, (t_1, \dots, t_N, s_1, \dots, s_\ell)\right) = x_f.$$

We order the times in the increasing order and we still denote by $\mathcal{T}$ the vector of all switching times.

**Redundancy creates robustness.** We will see further that these redundant switching times contribute to make the trajectory robust to external disturbances or model uncertainties, we will develop a method to tune the switching times in order to absorb these perturbations and steer the system to the desired target $x_f \in \mathbb{R}^n$.

Here, in this still informal introduction, we show how to use the additional switching times to make the system reach targets $x_f + \delta x_f$ in a neighborhood of $x_f$. The idea is to solve the nonlinear system of equations

$$E(x_0, t_f, \mathcal{T} + \delta\mathcal{T}) = x_f + \delta x_f.$$

Using (4.4), we propose to solve, at the first order,

$$\frac{\partial E}{\partial \mathcal{T}}(x_0, t_f, \mathcal{T}) \cdot \delta\mathcal{T} = \delta x_f, \tag{4.5}$$

which makes $n$ equations with $N + \ell$ degrees of freedom. We assume that $N + \ell$ is (possibly much) larger than $n$ and that the matrix in (4.5) is surjective. Then one can solve (4.5) by using

the *Moore-Penrose pseudo-inverse* $\left(\frac{\partial E}{\partial \mathcal{T}}\right)^\dagger$ of $\frac{\partial E}{\partial \mathcal{T}}$ (see [GVL13] and Appendix C, or see [Beu65a, Beu65b] for a theory in infinite dimension), which yields the solution of minimal Euclidean norm

$$\delta\mathcal{T} = \left(\frac{\partial E}{\partial \mathcal{T}}\right)^\dagger \cdot \delta x_f,$$

and we have

$$\|\delta\mathcal{T}\|_2 \leqslant \frac{\|\delta x_f\|_2}{\sigma_{\min}}, \tag{4.6}$$

where $\sigma_{\min}$ is the smallest positive singular value of $\frac{\partial E}{\partial \mathcal{T}}$. This estimate gives a natural measure for robustness, that we will generalize.

The two main contributions of this chapter are:

- the idea of adding redundant switching times in order to make a nominal bang-bang control more robust, while keeping it as being bang-bang;
- the design of a practical tracking algorithm, consisting of solving an overdetermined nonlinear system by least-squares, thus identifying a robustness criterion that we optimize.

They are developed in a rigorous and general context in the core of the chapter.

### 4.1.2 State of the art on robust control design

There is an immense literature on robust control theory, with many existing methods in order to efficiently control a system subjected to uncertainties and disturbances. Whereas there are many papers on $\mathcal{H}_2$ and $\mathcal{H}_\infty$ methods, except a few contributions in specific contexts, we are not aware of any general theory allowing one to tackle perturbations by using only bang-bang controls. This is the focus of this chapter.

Let us however shortly report on robustness methods when one is not bound to design bang-bang controls. In [KC99], a path-tracking algorithm with bang-bang controls is studied, for a double integrator and a wheeled robot. The technique relies heavily on the expression of the equations and does not apply to more general systems. In [SV94], the authors build a robust minimal time control for spacecraft's attitude maneuvers by canceling the poles of some transfer function. A remarkable fact is that the robustified control presents more switchings than the minimal time control. In this case, the robustness is evaluated as the maximum amplitude on a Bode diagram (see also [LW92] and [WSL93] for similar works). In [YL00], the authors observe that bang-bang controls are intrinsically not robust, and use pieces of singular trajectories (hence, not bang-bang) to overcome this issue.

In the $\mathcal{H}_2$ and $\mathcal{H}_\infty$ theories, control systems are often written in the frequency domain using the Laplace transform. For a transfer matrix $G(s)$, the two classical measures for performance are (see [DGKF89, ZDG96]) the $\mathcal{H}_2$ norm and the $\mathcal{H}_\infty$ norm respectively:

$$\|G\|_2 = \left(\frac{1}{2\pi}\int_{-\infty}^{+\infty} \mathrm{Trace}(G(j\omega)G(j\omega)^*)d\omega\right)^{1/2} \qquad \text{and} \qquad \|G\|_\infty = \sup_{\omega\in\mathbb{R}} \overline{\sigma}(G(j\omega)),$$

where $\overline{\sigma}(G)$ is the largest singular value of $G$.

In the linear quadratic theory, the question of optimal tracking has been widely addressed: given a reference trajectory $\xi(\cdot)$, we track it with a solution of some control system $\dot{x}(t) =$

$f(x(t), u(t))$, minimizing a cost of the form

$$\int_0^{t_f} \left( \|x(t) - \xi(t)\|_W^2 + \|u(t)\|_U^2 \right) \, dt + \|x(t_f) - \xi(t_f)\|_Q^2,$$

with weighted norms (see [AM71, KS72, Tré05b]). The first term in the integral measures how close one is to the reference trajectory, the second one measures a $L^2$ norm of the control (energy), and the third one accounts for the distance at final time between the reference trajectory $\xi(\cdot)$ and $x(\cdot)$. Then, the control can be expressed as a feedback function of the error $x(t) - \xi(t)$, involving the solution of some Riccati equation. In [dNDL13, Kha92], the authors investigate the question of stabilizing around a slowly time-varying trajectory. They also introduce uncertainties on the model and study the sensitivity of the system to those uncertainties. In the case of the existence of a delay on the input, a feedback law is proposed. In [Lin07, TSL09], uncertainties $p$ are introduced in a linear system $\dot{x}(t) = A(p)x(t) + Bu(t)$, and a tracking algorithm is suggested, under matching conditions on the uncertainties or not (see also [ADDJ91] for a survey on robust control for rigid robots).

In the late 1970's, $\mathcal{H}_\infty$ control theory developed. The control system is often described by a plant $G$ and a controller $K$. Then, the dependency of the error $z$ (to be minimized) on the input $v$ can be written as $z = F(G, K)v$. The $\mathcal{H}_\infty$ control problem consists of finding the best controller $K$ such that the $\mathcal{H}_\infty$ norm of the matrix $F(G, K)$ is minimized: $\|F(G, K)\|_\infty = \sup_{\omega \in \mathbb{R}} \overline{\sigma}(F(G, K)(j\omega))$. It can be interpreted as the maximum gain from the input $v$ to the output $z$. This criterion was introduced in order to deal with uncertainties on the model (on the plant $G$). In [Zam81], the author introduced the notion and highlighted the connection with robustness. In [DGKF89], a link is shown between the existence of such a controller and conditions on the solutions of two Riccati equations. Following a notion introduced in [Gah92], the linear matrix inequality (LMI) approach was introduced in [GA94], and used in [ANTT04, AN06] to solve the $\mathcal{H}_\infty$ synthesis. The Riccati equations are replaced with Riccati inequalities, whose set of solutions parameterizes the $\mathcal{H}_\infty$ controllers (see also [BGFB94] for the use of LMIs in control theory). The papers [DS81, MG92, XdSC92] present design procedures in this context to elaborate the feedback controller $K$. In [GFL96], the theory is extended to systems with parameters uncertainties and state delays, as well as in [XSCZ06], with stochastic uncertainty.

In many optimal control problems, the application of the Pontryagin maximum principle leads to bang-bang control strategies, and the classical $\mathcal{H}_2$ and $\mathcal{H}_\infty$ theories were not designed for such a purpose. But the optimal trajectories are in general not robust. Adding needle-like variations is therefore a way to improve robustness, and is the main motivation of this chapter. Of course, the method applies to any bang-bang control strategy, not necessarily optimal.

The approach that we suggest in this chapter combines an off-line treatment of the control strategies, with a feedback algorithm based on the structure of the control. We emphasize here that this algorithm preserves the bang-bang structure of the control. It consists of applying a nominal control strategy (that needs to be computed *a priori*), and adjusting it in real time, allowing one to track a nominal trajectory. The off-line method takes a solution of the control problem and makes it more robust by adding additional switching times (i.e., *redundancy*), which can be seen as additional degrees of freedom. Note that our analysis is done in the state space, without needing to consider the frequency domain. A key ingredient to the method is the use of needle-like variations.

### 4.1.3 Structure of this chapter

The chapter is organized as follows. In Section 4.2, we develop an algorithm to steer a perturbed system to the desired final point. The method is similar to the one presented in Section 4.1.1, except that we need to consider a backward problem. Indeed, the final point is fixed, and perturbations appear all along the trajectory. Besides, our measure for robustness comes out naturally in view of (4.6). Having identified the robustness criterion, we show in Section 4.3 how to add redundant switching times, leading one to solve a finite-dimensional nonlinear optimization problem. In Section 4.4, we provide some numerical illustrations on the attitude control problem of a 3-dimensional rigid body.

## 4.2 Tracking algorithm

**Setting.** In this chapter, we consider the control system

$$\dot{x}(t) = f(t, x(t), u(t)), \tag{4.7}$$

where $f$ is a smooth function $\mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, the state $x(\cdot) \in \mathbb{R}^n$, the control $u(\cdot) \in L^\infty([0, t_f]; \Omega)$, and $\Omega$ is the subset of $\mathbb{R}^m$: $[a_1, b_1] \times \cdots \times [a_m, b_m]$. We make two additional hypothesis: the controls we consider are "bang-bang", with a finite number of switching times:

$$\begin{array}{ll} (H_1) & \forall i \in [\![1, m]\!],\ u_i(t) \in \{a_i, b_i\}, \text{ a.e.} \\ (H_2) & \forall i \in [\![1, m]\!],\ u_i \text{ does not chatter.} \end{array}$$

A control is chattering when it switches infinitely many times over a compact time interval (see [ZTC16a, Ful63]). Therefore, our method does not apply to those controls. However, when the solution of an optimal control problem chatters, provided that it is possible, one could consider a sub-optimal solution, with only a finite number of switching times.

In the context of optimal control, we will denote the cost under the form

$$C(u) = \int_0^{t_f} f^0(t, x(t), u(t))\, dt. \tag{4.8}$$

We recalled in the introduction the (classical) definitions of the end-point mapping, of a needle-like variation (4.2) and the expansion of the end-point mapping subject to a needle-like variation (4.3).

### 4.2.1 Reduced end-point mapping

In this subsection, we give the definition of the reduced end-point mapping, and show a differentiability property.

Let us consider a bang-bang control $u(\cdot)$, and its associated trajectory $x(\cdot)$. For the sake of simplicity, we make the additional assumption that for every switching time $t_j$, one and only one component of the control commutes. Therefore, provided we specify the initial value of each component, the control $u$ is entirely characterized by the switching times of its components and can be represented by a vector:

$$((u_{10}, \ldots, u_{m0}), (t_1, i_1), \ldots, (t_N, i_N), t_f) \in \Omega \times \mathbb{R}^{2N+1},$$

where $u_{i0} \in \{a_i, b_i\}$ is the initial value for the control $u_i(\cdot)$ ($i \in [\![1, m]\!]$), $N$ is the total number of switching times, $t_f$ is the final time, and $i_j$ is the component of the control that switches

at time $t_j$. As this representation entirely characterizes the control, we will use indistinctly the notation $u$ and $((u_{10},\dots,u_{m0}),(t_1,i_1),\dots,(t_N,i_N),t_f)$ to speak about the control whose components switch at the times $t_j$. In the literature, $((t_1,i_1),\dots,(t_N,i_N))$ is often called a switching sequence.

**Remark 4.1:**

Had we wanted to allow simultaneous switching of multiple components, we would need to consider controls represented by:

$$((u_{10},\dots,u_{m0}),(t_1,\mathcal{I}_1),\dots,(t_N,\mathcal{I}_N),t_f),$$

where $\mathcal{I}_j \subset [\![1,m]\!]$ represents the set of components that switch at time $t_j$.

**Definition 4.1 (Reduced end-point mapping).** – *We define the* reduced end-point mapping *by*

$$E(x_0,(u_{10},\dots,u_{m0}),(t_1,i_1),\dots,(t_N,i_N),t_f) = x_u(x_0,t_f),$$

*where $u$ is the control represented by $((u_{10},\dots,u_{m0}),(t_1,i_1),\dots,(t_N,i_N),t_f)$, and $x_u(x_0,t_f)$ is the associated state at time $t_f$, starting at $x_0$.*

Note that in [MBKK05, MO04], the authors also reduce a bang-bang control to its switching points, in order to formulate an optimization problem in finite-dimension.

In the following, when writing this reduced end-point mapping, we may consider that the initial point $x_0$ is fixed, as well as the way the components of the control switch (i.e., we consider that the N-tuple $(i_1,\dots,i_N)$ is fixed), the initial values $u_{i0}$ and the final time $t_f$. In this context, we may forget them in the notations, and denote the reduced end-point mapping by

$$E(t_1,\dots,t_N) = x_u(t_f).$$

A remarkable fact is that the reduced end-point-mapping is differentiable. Compared to the expansion (4.3) with respect to a needle-like variation, the sign of $\delta t$ does not matter. For the sake of completeness, we give the proof at the end of the chapter.

**Proposition 4.1.** – *The reduced end-point mapping is differentiable, and*

$$dE(t_1,\dots,t_N) = \begin{pmatrix} v_1(t_f) & \cdots & v_N(t_f)\end{pmatrix} \in \mathcal{M}_{n,N}(\mathbb{R}),$$

*where $v_j(\cdot)$ ($j \in [\![1,N]\!]$) is the solution of the Cauchy problem, defined for $t \geqslant t_j$:*

$$\dot{v}_j(t) = \frac{\partial f}{\partial x}(t,x(t),u(t))v_j(t)$$

$$v_j(t_j) = \begin{cases} f(t_j,x(t_j),(\dots,a_{i_j},\dots)) - f(t_j,x(t_j),u(t_j^+)) & \text{if } u_{i_j} \text{ switches from } a_{i_j} \text{ to } b_{i_j}. \\ f(t_j,x(t_j),(\dots,b_{i_j},\dots)) - f(t_j,x(t_j),u(t_j^+)) & \text{if } u_{i_j} \text{ switches from } b_{i_j} \text{ to } a_{i_j}. \end{cases}$$

*The notation $(\dots,a_{i_j},\dots)$ (resp. $(\dots,b_{i_j},\dots)$) is used to show a difference with $u(t_j^+)$ (resp. $u(t_j^-)$) on the $i_j$-th component only.*

**Remark 4.2:**

In the special case of a control-affine system, as the attitude control system (3.3.1) studied in this thesis

$$\dot{x}(t) = f_0(x(t)) + \sum_{j=1}^{m} u_j(t) f_j(x(t)),$$

the initial condition on $v_j$ can be written much more easily:

$$v_j(t_j) = \begin{cases} (a_{i_j} - b_{i_j}) f_{i_j}(x(t_j)) & \text{if } u_{i_j} \text{ switches from } a_{i_j} \text{ to } b_{i_j}. \\ (b_{i_j} - a_{i_j}) f_{i_j}(x(t_j)) & \text{if } u_{i_j} \text{ switches from } b_{i_j} \text{ to } a_{i_j}. \end{cases}$$

### 4.2.2 Absorbing perturbations

As explained in the introduction of the chapter, we present here a closed-loop method to actually steer the system towards a point $x_f$, with bang-bang controls, even in the presence of perturbations.

First, for the sake of simplicity, we will explain how to control the system to some point $x_f + \delta x_f$. We will see that this idea can be adapted for our purpose of controlling a perturbed trajectory, by simply reversing the time.

**Perturbations on the final point.** We briefly generalize the problem introduced in the introduction. Let

$$\overline{u} = \left((u_{10}, \dots, u_{m0}), \left(\bar{t}_1, i_1\right), \dots, \left(\bar{t}_N, i_N\right), t_f\right) \in \Omega \times \mathbb{R}^{2N+1}$$

be a control such that $x_{\overline{u}}(t_f) = x_f$. That is, using the definition of Subsection 4.2.1, we have that

$$E(x_0, (u_{10}, \dots, u_{m0}), \left(\bar{t}_1, i_1\right), \dots, \left(\bar{t}_N, i_N\right), t_f) = x_f.$$

Or, considering that the final time $t_f$, the initial point $x_0$, the components $(i_1, \dots, i_N)$ and the initial values $(u_{10}, \dots, u_{m0})$ are fixed,

$$E(\bar{t}_1, \dots, \bar{t}_N) = x_f.$$

Let $\delta x_f$ be some perturbation of the final point $x_f$. We look for a vector $\delta\mathcal{T} = (\delta t_1, \dots, \delta t_N)$ so that the system reaches the target point $x_f + \delta x_f$:

$$E(\bar{t}_1 + \delta t_1, \dots, \bar{t}_N + \delta t_N) = x_f + \delta x_f.$$

As we have shown in Proposition 4.1 the differentiability of the reduced end-point mapping, we can write

$$E(\bar{t}_1 + \delta t_1, \dots, \bar{t}_N + \delta t_N) = E(\bar{t}_1, \dots, \bar{t}_N) + dE(\bar{t}_1, \dots, \bar{t}_N) \cdot \delta\mathcal{T} + o(\|\delta\mathcal{T}\|).$$

At order one, the solution is given by the solution of the linear equation

$$dE(\bar{t}_1, \dots, \bar{t}_N) \cdot \delta\mathcal{T} = \delta x_f.$$

It is natural to target the final point $x_f + \delta x_f$ while shifting the switching times as little as possible. That is, we look for the solution of minimal euclidian norm of the previous equation,

which is given by $\delta\mathcal{T} = dE(\bar{t}_1, \dots, \bar{t}_N)^\dagger \cdot \delta x_f$.

Therefore, we have shown how to compute, at order one, the correction to apply to control the system to some point $x_f + \delta x_f$: it boils down to solving a least-squares problem. Let us keep in mind that our definitive goal is to control systems that are perturbed all along their trajectory, to a fixed final point $x_f$. In other words, from a perturbed point $x(t) + \delta x(t)$ at some time $t \in [0, t_f)$, we want to absorb the perturbation $\delta x(t)$ and still reach the final point $x_f$. Even if this is a slightly different setting, we show that we can apply the same idea if we look at a *backward problem*.

**Absorbing a perturbation at time $t$.** Let $(\overline{x}(\cdot), \overline{u}(\cdot))$ be a nominal solution of the control system (4.7). We assume that when applying in practice the control $\overline{u} = \overline{\mathcal{T}}$, because of model uncertainties and perturbations, we observe a perturbed trajectory $x_{per}(t) = \overline{x}(t) + \delta x(t)$.

Let $t \in [0, t_f]$. Starting from the perturbed point $\overline{x}(t) + \delta x(t)$, which stands as a new initial point, we want to reach the final point $x_f$ in time $t_f - t$. Hence, we look for a control $\overline{u} + \delta u$ such that

$$E(\overline{x}(t) + \delta x(t), \overline{u} + \delta u, t_f - t) = x_f.$$

Assume for a moment that the perturbation of the control $\delta u$ is small in $L^\infty$ norm. Then, at least formally, one can write

$$E(\overline{x}(t), \overline{u}, t_f - t) + \frac{\partial E}{\partial x_0}(\overline{x}(t), \overline{u}, t_f - t) \cdot \delta x(t) + \frac{\partial E}{\partial u}(\overline{x}(t), \overline{u}, t_f - t) \cdot \delta u + o(\|\delta x(t)\| + \|\delta u\|) = x_f.$$

Therefore, at order one, we look for a solution of the (linear) equation

$$\frac{\partial E}{\partial x_0}(\overline{x}(t), \overline{u}, t_f - t) \cdot \delta x(t) + \frac{\partial E}{\partial u}(\overline{x}(t), \overline{u}, t_f - t) \cdot \delta u = 0. \tag{4.9}$$

However, we do not want, in this chapter, to apply small perturbations in the $L^\infty$ norm, as they would not result in bang-bang controls (However, this is similar to what is done while performing a Ricatti procedure to stabilize a system or track a reference trajectory). Nevertheless, reducing the end-point mapping to the switching times enables us to preserve the bang-bang structure: in the formalism previously introduced, we need to solve the nonlinear system of equations

$$E(\overline{x}(t) + \delta x(t), \overline{\mathcal{T}} + \delta\mathcal{T}, t_f - t) = x_f.$$

The equation (4.9) becomes

$$\frac{\partial E}{\partial \mathcal{T}}(\overline{x}(t), \overline{\mathcal{T}}, t_f - t) \cdot \delta\mathcal{T} = -\frac{\partial E}{\partial x_0}(\overline{x}(t), \overline{\mathcal{T}}, t_f - t) \cdot \delta x(t), \tag{4.10}$$

where the expression $\partial E / \partial \mathcal{T}$ is given by Proposition 4.1.

**A backward problem.** Solving this equation requires the computation of the partial differential $\partial E / \partial x_0$ at the initial point $\bar{x}(t)$. We will see now that it can be overcome by introducing a backward problem. Of course, the two formulations are equivalent.

**Definition 4.2 (Backward end-point mapping).** — *Let $u = (t_1, \dots, t_N)$ be a bang-bang control, and $t \in [0, t_f]$. We define the backward end-point mapping by*

$$\tilde{E}(t, t_1, \dots, t_N) = \tilde{x}(t_f - t),$$

*where $\tilde{x}(\cdot)$ is the solution to the Cauchy problem*

$$\begin{aligned}\dot{\tilde{x}}(t) &= -f(t_f - t, \tilde{x}(t), u(t_f - t)),\\ \tilde{x}(0) &= x_f.\end{aligned}$$

Note that for the nominal trajectory $(\overline{x}(\cdot), \overline{u}(\cdot))$, we have that

$$\tilde{E}(t, \bar{t}_1, \ldots, \bar{t}_N) = \overline{x}(t).$$

Indeed, we have in this case that $\overline{x}(t) = \tilde{x}(t_f - t)$: if we integrate the nominal system backward, starting from the point $x_f$ during a time period $t_f - t$, we end up at point $\overline{x}(t)$.

**Remark 4.3:**

Let $t \in [0, t_f]$, and $j$ be the smallest index such that $\bar{t}_j > t$ (with the convention that $j = N+1$ if $t > t_N$). Then, note that $\bar{t}_1, \ldots, \bar{t}_{j-1}$ do not play any role in the computation of $\tilde{E}(t, \bar{t}_1, \ldots, \bar{t}_N)$. The differential of $\tilde{E}$ can be computed with the Proposition 4.1. It is a matrix of size $n \times (N - j + 1)$.

In this context, the problem of adjusting the system back towards $x_f$ writes: at time $t$, find $(t_j, \ldots, t_N)$ such that

$$\tilde{E}(t, t_1, \ldots, t_N) = x_{per}(t). \tag{4.11}$$

We see that reversing the time, we place ourselves in the setting previously described of aiming at a perturbed final point. Therefore, we have the following proposition.

**Proposition 4.2.** – *At order one in $\delta x$, the solution of minimal norm of the problem (4.11) is given by $\overline{\mathcal{T}} + \delta\mathcal{T}$, with*

$$\delta\mathcal{T} = d\tilde{E}(t, \overline{\mathcal{T}})^{\dagger} \cdot \delta x(t), \tag{4.12}$$

*where $d\tilde{E}(t, \overline{\mathcal{T}})^{\dagger}$ denotes the pseudo-inverse of $d\tilde{E}(t, \overline{\mathcal{T}})$. Moreover, we have the estimate*

$$\|\delta\mathcal{T}\|_2 \leqslant \frac{1}{\sigma_{\min}(t)} \|\delta x(t)\|_2, \tag{4.13}$$

*where $\sigma_{\min}(t)$ is the smallest positive singular value of $d\tilde{E}(t, \overline{\mathcal{T}})$.*

*Proof.* The scheme of the proof has already been exposed previously in the chapter. However, we write it extensively here. Let $\delta\mathcal{T} = \mathcal{T} - \overline{\mathcal{T}}$. The problem writes

$$\tilde{E}(t, \overline{\mathcal{T}} + \delta\mathcal{T}) = x_{per}(t).$$

According to Proposition 4.1, the backward end-point mapping is differentiable (and we also know how to compute its derivative), so

$$\begin{aligned}\tilde{E}(t, \overline{\mathcal{T}} + \delta\mathcal{T}) &= \tilde{E}(t, \overline{\mathcal{T}}) + d\tilde{E}(t, \overline{\mathcal{T}}) \cdot \delta\mathcal{T} + o(\|\delta\mathcal{T}\|)\\ &= \overline{x}(t) + d\tilde{E}(t, \overline{\mathcal{T}}) \cdot \delta\mathcal{T} + o(\|\delta\mathcal{T}\|).\end{aligned}$$

So, at order one, the problem writes

$$d\tilde{E}(t, \overline{\mathcal{T}}) \cdot \delta\mathcal{T} = \delta x(t). \tag{4.14}$$

It is well known (see [AK02] for instance), that the solution of minimal norm of this equation

is $\delta\mathcal{T} = d\tilde{E}(t,\overline{\mathcal{T}})^\dagger \cdot \delta x(t)$. Besides, let $\sigma_{\max}(t) > \cdots > \sigma_{\min}(t) > 0$ denote the positive singular values of $d\tilde{E}(t,\overline{\mathcal{T}})$. We have that $\left\|d\tilde{E}(t,\overline{\mathcal{T}})^\dagger\right\|_2 = 1/\sigma_{\min}(t)$ ($\|\cdot\|_2$ for a matrix denotes the induced norm corresponding to the euclidean norm), so that

$$\begin{aligned}\|\delta\mathcal{T}\|_2 &= \left\|d\tilde{E}(t,\overline{\mathcal{T}})^\dagger \cdot \delta x(t)\right\|_2 \\ &\leqslant \left\|d\tilde{E}(t,\overline{\mathcal{T}})^\dagger\right\|_2 \cdot \|\delta x(t)\|_2 \\ &\leqslant \frac{\|\delta x(t)\|_2}{\sigma_{\min}},\end{aligned}$$

which concludes the proof. □

**Remark 4.4: Relative error estimate**

In Proposition 4.2, we show the absolute error estimate

$$\|\delta\mathcal{T}\|_2 \leqslant \frac{1}{\sigma_{\min}(t)} \|\delta x(t)\|_2 ,$$

where $\delta\mathcal{T}$ is a solution of the equation (4.12). However, one may want in some cases to have instead a relative error estimate. It holds

$$\frac{\|\delta\mathcal{T}\|_2}{\|\mathcal{T}\|_2} \leqslant \frac{\sigma_{max}(t)}{\sigma_{\min}(t)} \cdot \frac{\|\delta x(t)\|_2}{\|x(t)\|_2}.$$

The quantity $\frac{\sigma_{max}(t)}{\sigma_{\min}(t)}$ is the condition number (with respect to the Euclidian norm) of the matrix $d\tilde{E}(t,\overline{\mathcal{T}})$. We give in Appendix C more details on the condition number of a matrix.

**Remark 4.5:**

We have the relation that, for all vector of switching times $\mathcal{T}$

$$E(\tilde{E}(t,\mathcal{T}),\mathcal{T},t_f - t) = x_f.$$

Differentiating this equality with respect to $\mathcal{T}$, we have that, for all $\delta\mathcal{T}$

$$\frac{\partial E}{\partial x_0}(\tilde{E}(t,\mathcal{T}),\mathcal{T},t_f - t) \cdot d\tilde{E}(t,\overline{T}) \cdot \delta\mathcal{T} + \frac{\partial E}{\partial \mathcal{T}}(\tilde{E}(t,\mathcal{T}),\mathcal{T},t_f - t) \cdot \delta\mathcal{T} = 0.$$

Replacing the second term by its value in (4.10), it follows that

$$\frac{\partial E}{\partial x_0}(\tilde{E}(t,\mathcal{T}),\mathcal{T},t_f - t) \cdot d\tilde{E}(t,\mathcal{T}) \cdot \delta\mathcal{T} = \frac{\partial E}{\partial x_0}(\tilde{E}(t,\mathcal{T}),\mathcal{T},t_f - t) \cdot \delta x(t).$$

It is easy to show that $\partial E/\partial x_0$ can be expressed as the resolvent of a linearized system. Therefore, the matrix $\partial E/\partial x_0$ is invertible, and the equations (4.10) and (4.14) are equivalent. But solving (4.14) only requires to compute the derivative of $\tilde{E}$. This is what we do in the following.

**Remark 4.6:**

Note that it might not always be possible to find a solution to the equation $d\tilde{E}(t,\overline{\mathcal{T}})\cdot\delta\mathcal{T}=\delta x(t)$. This may happen for instance if $t>t_{N-n+1}$, i.e., we do not have enough degrees of freedom left to absorb the perturbation $\delta x(t)\in\mathbb{R}^n$. However, we can still give a meaning to the equation $d\tilde{E}\cdot\delta\mathcal{T}=\delta x(t)$. We look for a solution to the least-square problem:

$$\min_{\delta\mathcal{T}\in\mathbb{R}^N}\left\|d\tilde{E}(t,\bar{t}_1,\dots,\bar{t}_N)\cdot\delta\mathcal{T}-\delta x(t)\right\|_2^2,$$

for which $\delta\mathcal{T}=d\tilde{E}(t,\bar{t}_1,\dots,\bar{t}_N)^\dagger\cdot\delta x(t)$ is still the solution of minimal norm (see [AK02]). We see here emerging the idea that the number of switching times (i.e., degree of freedom) left at time $t$, is going to be an important factor to track the system back towards the final point $x_f$.

**Numerical algorithm.** At time $t$, Equation (4.12) provides us with a formula to adjust the control so that the perturbed trajectory eventually reaches $x_f$. But it certainly does not enable us to face perturbations that would happen after time $t$. In order to absorb perturbations all along the trajectory, we suggest the following algorithm: Let $\mathcal{T}$ be an initial control. Given an integer $s$ and a subdivision $0<\tau_1<\cdots<\tau_s<t_f$ of the interval $[0,t_f]$, we adjust the control at each $\tau_i$ for all $i\in[\![1,s]\!]$. That is, for each $i\in[\![1,s]\!]$, we measure the drift $\delta x(\tau_i)=x_{per}(\tau_i)-x_{ref}(\tau_i)$, and compute the differential of the backward end-point mapping $d\tilde{E}(\tau_i,\bar{t}_1,\dots,\bar{t}_N)$. We deduce from (4.12) that the correction to apply is then $\delta\mathcal{T}=d\tilde{E}(\tau_i,\bar{t}_1,\dots,\bar{t}_N)^\dagger\cdot\delta x(\tau_i)$. We then update the control by considering the new vector of switching times $\mathcal{T}+\delta\mathcal{T}$.

**Algorithm 2** Tracking algorithm to absorb perturbations

Choose an integer $s$ and a subdivision $(\tau_1,\dots,\tau_s)$.
Set $t=0$.
Set $x_{ref,0}=x_0$ ▷ Initial conditions
$\overline{\mathcal{T}}$ ▷ Initial switching times
**for** $i=0$, $i<N$, $i=i+1$ **do**
    Integrate the ideal system $f$ from $t$ to $\tau_i$, with initial conditions $x_{ref,0}$.
    Measure the drift $\delta x(\tau_i)=x_{per}(\tau_i)-x_{ref}(\tau_i)$.
    Compute the differential of the backward end-point mapping $d\tilde{E}(\tau_i,\bar{t}_1,\dots,\bar{t}_N)$.
    Compute the correction $\delta\mathcal{T}=d\tilde{E}(\tau_i,\bar{t}_1,\dots,\bar{t}_N)^\dagger\cdot\delta x\tau_i)$.
    Apply the correction $\overline{\mathcal{T}}\leftarrow\overline{\mathcal{T}}+\delta\mathcal{T}$.
    **if** $\exists j$ s.t. $\bar{t}_{j+1}<\bar{t}_j$ **then**
        "Stop". Interchanging of switching times.
    **end if**
    $x_{ref,0}\leftarrow x_{per}(\tau_i)$.
    $t\leftarrow\tau_i$.
**end for**

**Remark 4.7:**

When computing the correction $\mathcal{T} + \delta\mathcal{T}$, it may happen that the new switching times are not ordered, i.e., there exists some integer $j \in [\![1, N-1]\!]$ such that $t_{j+1} < t_j$. In this case, we consider that the correction is not physically acceptable, and we reject it at line 12 of Algorithm 2. (Note that in some cases, we may want to continue the integration of the system even if two switching times are not ordered. In that case, we can always use the last admissible control, where all the switching times are ordered.)

**Remark 4.8:**

The computation of the differential $d\tilde{E}(t, \bar{t}_1, \dots, \bar{t}_N)$ is done via the integration of a system of ordinary differential equations, which can be done efficiently and quickly using numerical integrators. However, the size of the system (as well as the time required to compute the pseudo-inverse) directly depends on the number of switching times $N$ and on the state dimension $n$.

## 4.3 Promoting robustness

Intuitively, we want to say that a control is robust whenever the correction $\delta\mathcal{T}$ required to absorb the perturbation $\delta x(t)$ is small. Since we have shown the estimate $\|\delta\mathcal{T}\|_2 \leqslant \|\delta x(t)\|_2 / \sigma_{\min}(t)$, a robust trajectory is then one for which the values of $1/\sigma_{\min}(t)$ remain small along the trajectory.

**Definition 4.3.** – *We define the following cost, that we will use to characterize the robustness of a trajectory*

$$C_r(t_1, \dots, t_N) = \int_0^{t_N} \frac{1}{\sigma_{\min}(t)^2}\, dt. \tag{4.15}$$

**Remark 4.9: Variations of the cost**

In the previous definition, the upper bound in the integral is $t_N$, because for $t > t_N$, the backward end-point mapping derivative $d\tilde{E}(t, t_1, \dots, t_N)$ is not defined, and neither is $\sigma_{\min}(t)$. For some reason, we may only want to have robustness up until some time $t^\star < t_N$. Then the previous definition would become $\int_0^{t^\star} 1/\sigma_{\min}(t)^2 dt$.
Note also that following Remark 4.4, one may have wished to define the cost

$$C_r(t_1, \dots, t_N) = \int_0^{t_N} \frac{\sigma_{\max}(t)^2}{\sigma_{\min}(t)^2}\, dt.$$

In this section, we show how the switching times of a trajectory can be chosen to build one that is more robust. We also suggest a new way to design a trajectory, by adding redundant switching times, that give us more degrees of freedom. Note also that we will start from a solution of an *optimal* control problem, because it is of high importance in practice, but the method generally applies when starting from any control, as long as it satisfies the hypothesis $(H_1)$ and $(H_2)$. Starting from an initial control such that $E(t_1, \dots, t_N) = x_f$, we look for *redundant* switching times $(s_1, \dots, s_l)$ such that $E(t_1, \dots, t_N, s_1, \dots, s_l) = x_f$, while minimizing the cost (4.15) that accounts for robustness:

$$C_r(t_1, \dots, t_N, s_1, \dots, s_l).$$

### 4.3.1 An auxiliary optimization problem

Let us consider a bang-bang trajectory (satisfying the hypothesis $(H_1)$ and $(H_2)$) of the control system (4.7), optimal for the cost (4.8). That is, $\overline{u} = ((u_{10}, \dots, u_{m0}), (\bar{t}_1, i_1), \dots, (\bar{t}_N, i_N), t_f)$ is an optimal solution of the optimization problem

$$\min_{(i_1,\dots,i_N)} \quad \min_{\substack{(t_1,\dots,t_N) \\ \text{s.t. } E(t_1,\dots,t_N) = x_f}} \quad C(t_1, \dots, t_N). \tag{4.16}$$

Let us emphasize the fact that reducing the control to its switching times enables us to reduce a problem in infinite dimension

$$\min_{\substack{u \in L^\infty([0,t_f];\Omega) \\ \text{s.t. } E(u) = x_f}} \quad C(u)$$

to a finite number of non-linear problems under non-linear constraints in finite dimension, provided we set $N$, as we left aside chattering trajectories.

In order to make the control more robust we suggest to solve the following problem. We fix the components of the control $(i_1, \dots, i_N)$, and we introduce the cost that accounts for the robustness of a trajectory:

$$\min_{\substack{(t_1,\dots,t_N) \\ \text{s.t. } E(t_1,\dots,t_N) = x_f}} \quad \lambda_1 C(t_1, \dots, t_N) + \lambda_2 C_r(t_1, \dots, t_N),$$

where $\lambda_1$ and $\lambda_2$ are two parameters, chosen to give more or less importance to the different costs. For instance, if $\lambda_1 \gg \lambda_2$, the solution is close to the initial one $(\bar{t}_1, \dots, \bar{t}_N)$.

### 4.3.2 Redundancy creates robustness

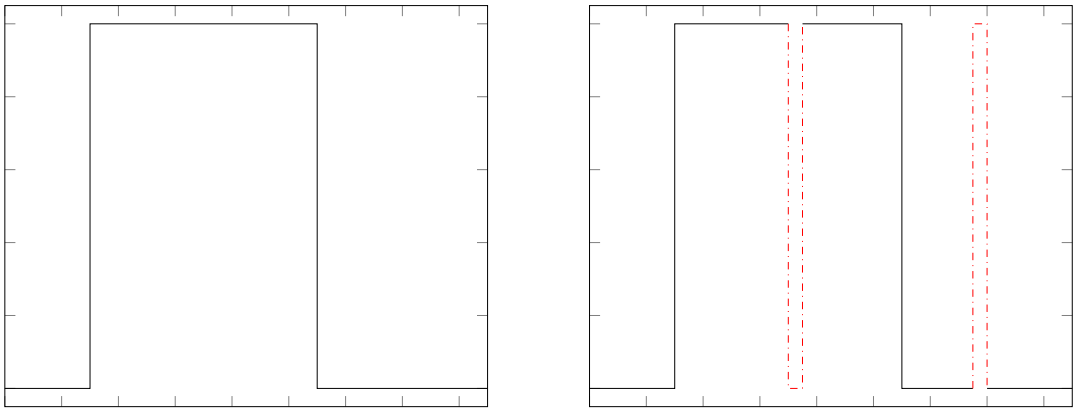
Let us consider a control $u = ((u_{10}, \dots, u_{m0}), (t_1, i_1), \dots, (t_N, i_N), t_f)$. In order to reduce the optimization space, we will consider in the following subsection that the initial control values $(u_{10}, \dots, u_{m0})$, the components $(i_1, \dots, i_N)$ and the final time $t_f$ are fixed, so we will forget them in the notations.

We propose here to go further in order to improve the robustness of the corresponding trajectory. We do so by adding needles to some components of the control. By needle, we mean a short impulse on one of the control. Let us denote by $l$ the number of needles we are willing to add. It means that we look for additional switching times $[(s_1, s_2), \dots, (s_{2l-1}, s_{2l})]$ and components of the control $(j_1, \dots, j_l)$, so that for all $i \in [\![1, l]\!]$, $(s_{2i-1}, s_{2i})$ are switching times for the $j_i$-th components of the control (see Figure 4.3). It aims at giving us more degrees of freedom while trying to absorb perturbations $\delta x$ by moving the switching times $(\mathcal{T}, \mathcal{S}) = (t_1, \dots, t_N, (s_1, s_2), \dots, (s_{2l-1}, s_{2l}))$. Thus, we are solving the optimization problem

$$\min_{(j_1,\dots,j_l)} \quad \min_{\substack{(\mathcal{T},\mathcal{S}) \\ \text{s.t. } E(\mathcal{T},\mathcal{S}) = x_f}} \quad \lambda_1 C(\mathcal{T}, \mathcal{S}) + \lambda_2 C_r(\mathcal{T}, \mathcal{S}). \tag{4.17}$$

**Remark 4.10:**

If the original bang-bang control strategy $\bar{u}$ does not come from an optimization process, that is there is no cost $C$ associated with it, we can still consider problem (4.17) but with $\lambda_1 = 0$.



(a) Initial control. (b) Adding four switching times i.e., 2 needles.

Figure 4.3 – Principle of adding needles.

Let us denote by $\overline{\mathcal{T}}$ the solution of problem (4.16), and by $(\mathcal{T}^\star, \mathcal{S}^\star)$ the solution of problem (4.17). Then, we have that

$$C(\overline{\mathcal{T}}) \leqslant C(\mathcal{T}^\star, \mathcal{S}^\star).$$

It means that the solution $(\mathcal{T}^\star, \mathcal{S}^\star)$ is sub-optimal with respect to the initial cost $C$. However, this sub-optimality comes with a gain in terms of robustness. Besides, the loss of optimality (and therefore gain in robustness) can be controlled by the choice of the coefficients $\lambda_1$ and $\lambda_2$.

This problem is a mixed problem, with integer variables (the components $(j_1, \dots, j_l)$), and continuous variables (the switching times $(t_1, \dots, t_N, (s_1, s_2), \dots, (s_{2l-1}, s_{2l}))$). However, if the components are fixed, we only have to solve a non-linear problem subject to non-linear constraints in finite dimension

$$\begin{array}{ll} \min_{(\mathcal{T},\mathcal{S})} & \lambda_1 C(\mathcal{T},\mathcal{S}) + \lambda_2 C_r(\mathcal{T},\mathcal{S}). \\ \text{s.t. } & E(\mathcal{T},\mathcal{S}) = x_f \end{array} \tag{4.18}$$

We used an interior-point algorithm to solve (4.18). In [XA00, ZA15], gradient-based algorithms are shown to be effective to solve such problems, when the sequence of indices $(j_1, \dots, j_l)$ is fixed. Therefore a "naïve" way to proceed, if $m$ denotes the number of components of the control, is to solve $m^l$ optimization problems, which is extremely costly if $m$ or $l$ is big. A compromise has to be found between the potential benefit in robustness and the computational cost. Such a compromise will however depend on the particular problem at hand, so we do not elaborate too much on this issue and give an example in Section 4.4. Let us cite [CHSC08, CHS+09], where the authors parametrize an optimal control problem (for the time-minimal and $L^1$ problem) with the switching times of the controls. They simplify its complex structure by fixing the number of switching times, and wonder how many switching times are required to obtain a cost close to the optimal one : the result is striking as 2 or 3 may be enough. However, they know from an *a priori* study the value of the optimal $L^1$ or time-minimal cost, and therefore can stop adding switching times when reaching a given percentage of this optimal value of the criterion. In our problem, we do not know what is the optimal value of the criterion we identified to quantify the robustness of a trajectory. It becomes necessary to find another way to decide how many needles to add.

One could consider tackling directly Problem 4.17, a combinatorial optimization problem (which is a class of problem known to be hard to solve). Recent years have seen the development

of advanced numerical procedures to deal with the combinatorial nature of those problem at a reasonable computational cost. We give more details on this issue at the end of this section.

**Remark 4.11:**

Let us make here a remark on the ordering of the switching times. In the vector $(\mathcal{T}, \mathcal{S})$ are stored the switching times $t_i$ and $s_i$ that represent the control $u$. Those swicthing times are not necessarily ordered during or after the optimization process, so let $\mathbb{T} = (\tau_1, \dots, \tau_{N+2l})$ denote the ordered equivalent to $(\mathcal{T}, \mathcal{S})$. So far, we have made the implicit assumption that when we perform the numerical integration of the system, the switching times are ordered: $\tau_{i+1} - \tau_i \geqslant 0$ for all $i \in [\![0, N+2l-1]\!]$. We recall that our goal is to absorb perturbations $\delta x$. As explained in Subsection 4.2.2, we compute at order one the correction to apply $\delta\mathbb{T} = dE(\mathbb{T})^\dagger \cdot \delta x$. At this point, we could have that $\mathbb{T} + \delta\mathbb{T}$ does not satisfy this ordering property. Then, we consider that $\mathbb{T} + \delta\mathbb{T}$ is not admissible, and an estimate like (4.13) would not hold.

In the following, in order to guarantee that we do not have an interchanging of the switching times (at least for small perturbations), we add an additional constraint whilst elaborating the robustified trajectory $(u(\cdot), x(\cdot))$ at (4.17):

$$\tau_{i+1} - \tau_i \geqslant \eta \quad \text{for all} \quad i \in [\![0, N+2l-1]\!], \tag{4.19}$$

for some $\eta > 0$, where $\mathbb{T} = (\tau_1, \dots, \tau_{N+2l})$ denotes the re-ordering of the vector $(\mathcal{T}, \mathcal{S})$. In that way, we ensure that two consecutive switching times ($\mathcal{T}$ and $\mathcal{S}$ combined) are at least distant of $\eta$. Thus, if $\delta x$ is small enough, the elements of the vector $\mathbb{T} + dE(\mathbb{T})^\dagger \cdot \delta x$ remain in ascending order. Besides, such a constraint is often highly justified in practice, for instance if a physical system has to spend some minimum time $\eta$ before it switches to another mode. For example, in Section 4.4, the attitude control of a rigid body is studied. In real life, because of robustness issues and mechanical constraints, nozzles on a space launcher have indeed a minimum activation time.

**Remark 4.12:**

Let $t_f$ denote the final time. If $\eta$ is the minimal time between two switchings in (4.19), then the total number of switchings $N + 2l$ has an upper bound of $\lfloor t_f/\eta \rfloor$.

The elaboration of a robust trajectory in (4.17) can be seen as an optimal control problem of switched-mode dynamical system. A recent survey on switched systems can be found in [ZA15]. This theory deals with control systems where the dynamics can only take a finite number of modes. To determine the command law, one has to determine the switching times, as well as the different modes of the system. If the modes are fixed (in our case, it means that the components $(i_1, \dots, i_N, j_1, \dots, j_l)$ are fixed), it is often called a timing-optimization problem ; if not, a scheduling optimization problem. In [Pic99, Sus00], necessary conditions are derived, for trajectories of hybrid systems considering a fixed sequence of modes of finite length (in our setting, it corresponds to the Problem (4.18)). In [AE14, War12], the authors develop numerical algorithms to solve both the timing and the scheduling problems. Their techniques rely heavily on gradient-like methods. However, the latter problem is much more complex because of its discrete nature: indeed the procedure needs to account for both continuous and discrete control variables, and can therefore be seen as a combinatorial optimization problem. Note that the paper [AE14] deals with dwell time constraints. It consists in imposing a threshold $\eta$ between two consecutive switching times which is the constraint we introduced at (4.19). Let us also mention other techniques to solve scheduling optimization problems, like zoning algorithms [SC05], or

relaxation methods, where discrete variables are temporarly relaxed into continuous variables [BD05].

## 4.4 Numerical results

In order to illustrate the results of Sections 4.2 and 4.3, we consider the problem of interest in this thesis : the attitude control of a rigid body. However, in order to keep a reasonnable run-time, we will only consider the part for the angular velocity in (3.3.1), and for thrusters on the launcher.

Let $\vec{\omega} = (p, q, r)$ be the angular velocity of the body with respect to a frame fixed on the body. We recall the Euler's equation (established in the Introduction chapter) for the angula velocity of a rigid body, subjected to torques $(b^1, \dots, b^m)$, writes:

$$I\dot{\vec{\omega}} = I\vec{\omega} \wedge \vec{\omega} + \sum_{k=1}^{m} b^k.$$

In the case when the axes of the body frame are the axes of inertia of the body, the matrix $I$ is diagonal: $I = \mathrm{diag}(I_x, I_y, I_z)$. The controlled Euler's equations can then be reduced to

$$\dot{\vec{\omega}}(t) = f\left(\vec{\omega}(t), u(t)\right),$$

where for $1 \leqslant k \leqslant m$, $u_k(t) \in \{0, 1\}$ almost everywhere, and the function $f$ describing the dynamics writes:

$$f(p, q, r, u_1, u_2, u_3, u_4) = \begin{cases} \alpha_1 qr + \sum_{k=1}^{m} b_1^k u_k \\ \alpha_2 pr + \sum_{k=1}^{m} b_2^k u_k \\ \alpha_3 pq + \sum_{k=1}^{m} b_3^k u_k \end{cases}, \tag{4.20}$$

with $\alpha_1 = (I_y - I_z)/I_x$, $\alpha_2 = (I_z - I_x)/I_y$ and $\alpha_3 = (I_y - I_x)/I_z$. This is with a slight abuse in the notations, because we still denote by $b^k$ the normalized vector $(b_1^k/I_x, b_2^k/I_y, b_3^k/I_z)$.

The controllability of such a system has been studied in Chapter 1. Let us mention here the papers [KT99, OS92, Win63], that implement, in the special case of the stabilization of a rigid spacecraft, methods to stabilize the spacecraft towards the point $(0, 0, 0)$, but once again, the controls used are not bang-bang. Note that (4.20) is a control-affine system, and therefore, Remark 4.2 applies.

In the following, we consider the numerical values $\alpha_1 = 1$, $\alpha_2 = -1$, $\alpha_3 = 1$, $b^1 = [2, 1, 0.3]$, $b^2 = [-2, -1, -0.3]$, $b^3 = [0, 0, 1]$ and $b^4 = [0, 0, -1]$, and initial and final conditions $x_0 = (0, 0, 0)$ and $x_f = (0.4, -0.3, 0.4)$.

We start by building an optimal trajectory for the $L^1$ cost $\int_0^{t_f} \sum_{j=1}^{4} |u_j(t)| dt + t_f/2$ (the presence of $t_f$ ensures us not to obtain a trajectory with infinite final time). It amounts to minimizing the consumption of the launcher. The resolution of such a problem with a $L^1$ cost can be numerically challenging. We explained in Chapter 2 how a continuation procedure can be implemented to tackle this numerical difficulty. In the following subsection we recall very briefly the principle of such a method.

### 4.4.1 Computing the nominal trajectory

The nominal trajectory, optimal for the $L^1$ cost, is computed with a continuation procedure. The idea of such a procedure is to solve first an "easier" problem, and deform it step by step to solve the targeted problem. We introduce the continuation parameter $\lambda \in [0, 1]$, and we consider

the optimal control problem $(\mathcal{P}_\lambda)$ of steering the system (4.20) from $x_0$ to $x_f$, by minimizing the cost

$$\lambda \int_0^{t_f} \sum_{i=1}^{4} |u_j(t)|^2 \, dt + (1-\lambda) \int_0^{t_f} \sum_{i=1}^{4} |u_j(t)| \, dt + t_f.$$

When $\lambda = 0$, we recognize our problem. For some $\lambda \in [0,1]$, solving problem $(\mathcal{P}_\lambda)$ is done by finding the zeros of a shooting function that results from the application of Pontryagin maximum principle. Solving a shooting problem is done with Newton like methods. Such methods are highly sensitive to their initialization, that can be very difficult, especially in the case of the minimization of the $L^1$ norm $\int_0^{t_f} |u(t)|dt$. The continuation procedure is introduced to overcome this difficulty.

For $\lambda = 1$, the cost is stricly convex in the controls, and writes

$$\int_0^{t_f} \sum_{i=1}^{4} |u_j(t)|^2 \, dt + t_f,$$

for which the initialization of the induced shooting method is much easier. Therefore, we solve a sequence of optimal control problems, for values of $\lambda$ decreasing from 1 to 0. The result of the shooting problem for some $\lambda \in ]0,1]$ serves as the initialization of another problem with $\lambda' < \lambda$.

### 4.4.2 Robustifying the nominal trajectory

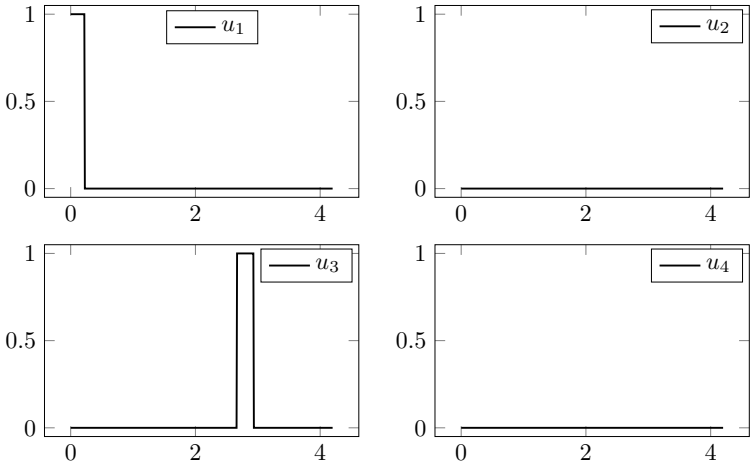
From this $L^1$ - minimal trajectory, represented on Figure 4.4, with three switching times that we denote $(t_1, t_2, t_3)$ we build a new trajectory by solving the problem (4.17) with 3 needles (i.e., $l = 3$), $\lambda_1 = \lambda_2 = 1$, and taking $\eta = 0.05$ in Equation (4.19). As explained in Remark 4.6, we see that it is worthwile to have the additional switching times available as long as possible. That is, we force the additional switchings to occur after $t_3$. Keeping in mind Equation (4.19), this constraint can be written:

$$t_{i+1} - t_i \geqslant \eta \quad (\forall i \in [\![1,3]\!]), \quad s_1 - t_3 \geqslant \eta, \quad s_{i+1} - s_i \geqslant \eta \quad (\forall i \in [\![1,6]\!]).$$

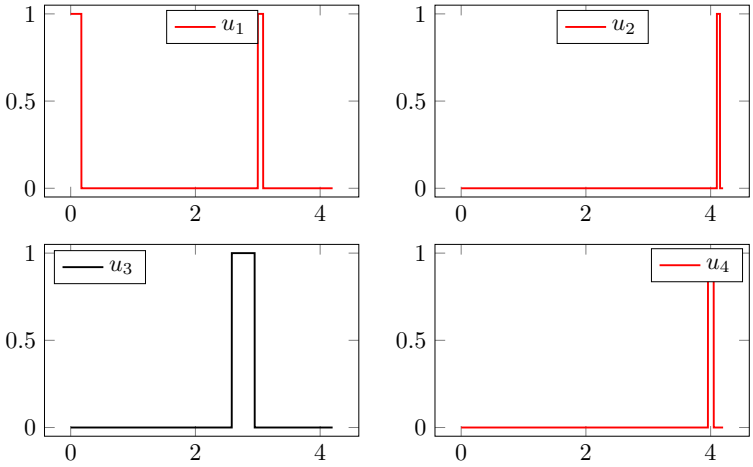
We find that the optimal triplet is $(j_1, j_2, j_3) = (1,4,2)$, for which we have $C = 0.77$ and $C_r = 2.22$. We found this optimal triplet by exploring the $4^3 = 64$ possibilities. We then used the heuristic that this solution would make a good choice to start looking for the solution with 4 needles (as it would have been to costly to examine the $4^4 = 256$ possibilities). However we could not make the cost dicrease significantly (the best cost we found was $C_r = 2.07$). This heuristic is very similar to what is used in Branch and Bound methods. Besides, as an element of comparison, the optimal couple when adding only two needles is $(j_1, j_2) = (1,4)$, for which $C_r = 4.25$, and the optimal solution when adding only on needle is $j_1 = 2$, for which $C_r = 30.28$. Thus, we notice a substantial improvement when increasing the number of needles from 1 to 2 and from 2 to 3, whereas it seems less profitable to add a fourth one. We therefore stopped at 3 needles. The controls are displayed on Figure 4.4, and the components 1, 2 and 4, on which needles have been added, are represented in red.

In order to represent perturbations, we consider that the principal moments of inertia can vary, causing the coefficients $\alpha_1$, $\alpha_2$ and $\alpha_3$ to vary. Thus we consider the perturbed dynamics

$$f_{per}(t,p,q,r,u_1,u_2,u_3,u_4) = \begin{cases} \alpha_1^{per,\varepsilon}(t)qr + \sum_{k=1}^{4} b_1^k u_k \\ \alpha_2^{per,\varepsilon}(t)pr + \sum_{k=1}^{4} b_2^k u_k \\ \alpha_3^{per,\varepsilon}(t)pq + \sum_{k=1}^{4} b_3^k u_k \end{cases}, \tag{4.21}$$




(a) Controls for the minimal $L^1$ trajectory. $C = 0.49$




(b) Controls with three needles. $C = 0.77$, $C_r = 2.22$

Figure 4.4 – Improving the robustness of a trajectory adding needles. We lose optimality with respect to the consumption in order to gain robustness.

so that $\varepsilon$ models the size of the perturbation. More precisely, we take $\alpha_i^{per,\varepsilon}(t) = \alpha_i + \varepsilon h_i(t)$, where $h_i(\cdot)$ is some periodic function satisfying $\|h_i\|_\infty \leqslant 1$ (note that the exact expression of $h_i$ is not relevant here, as it is supposed to model any perturbation of the $\alpha_i$). We denote by $x_{per}$ the solution of the Cauchy problem

$$\begin{aligned} \dot{x}(t) &= f_{per}(t, x(t), u(t)), \\ x(0) &= x_0. \end{aligned}$$

We denote by $x_{cor}$ the corrected trajectory computed with our algorithm. We show, on Figure 4.6, the three trajectories, for $\varepsilon = 0.78$ and a cost $C_r = 2.22$. We can see the perturbed trajectory $x_{per}$ drifting away from the reference trajectory $x_{ref}$ and away from the final point

$x_f$, whereas the corrected trajectory $x_{cor}$ eventually reaches a point very close to $x_f$. Actually, for the trajectories represented on Figure 4.6, we have that $\|x_{cor}(t_f) - x_f\| / \|x_f\| = 5.5 \times 10^{-3}$, whereas $\|x_{per}(t_f) - x_f\| / \|x_f\| = 1.3 \times 10^{-1}$. Our algorithm has indeed been able to adjust the perturbed trajectory back towards $x_f$.

One may wonder how this method behaves with respect to the choice of $\varepsilon$. As explained in Remark 4.7, we stop if two switching times are interchanged, that is, if $\delta T$ is too big, as the initial vector of switching times satisfies a gap property (4.19). Actually, this is not strictly true, as we could have a "big" correction that does not change the ascending order of the switching times, for instance if we shift all the switching times in the same direction. However, we experimentally notice that the cost $C_r$ *has an impact on the size of the perturbation we are able to absorb.*

We build several trajectories, for which we apply our algorithm for increasing values of $\varepsilon$, until the algorithm fails as explained in Remark 4.7, for some $\varepsilon_{max}$. We plot on Figure 4.5 the value of $\varepsilon_{max}$ with respect to the cost $C_r$ (that is, for a given cost $C_r$, $\varepsilon_{\max}$ is the smallest value for which there is an interchanging of switching times). Even if the curve is not decreasing (for the reason explained above), we can see that *having a low cost $C_r$ enables us to absorb bigger perturbations.*

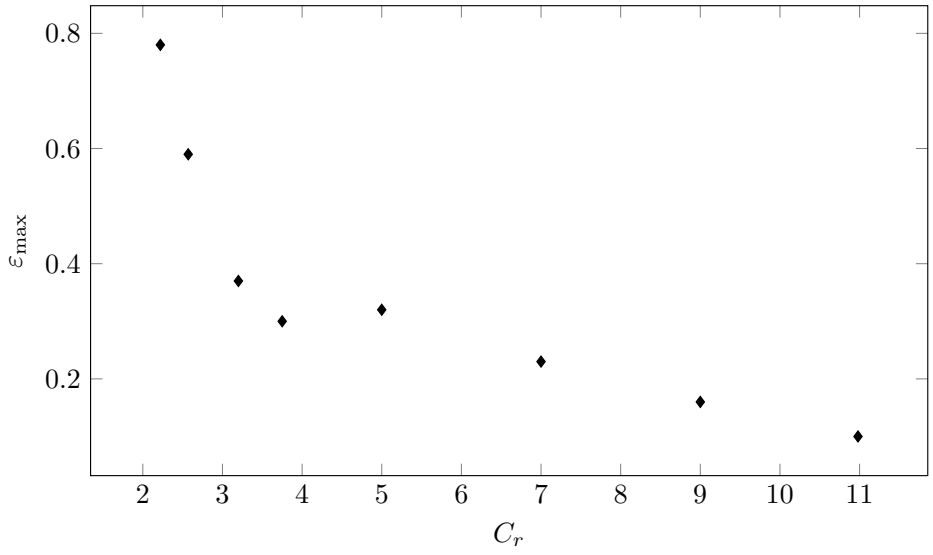



Figure 4.5 – Size of the maximal perturbation absorbed with respect to the robustness of a trajectory

On Figure 4.7, we show the relative error $\|x(t_f) - x_f\|/\|x_f\|$ for the perturbed $x_{per}$ and corrected $x_{cor}$ trajectories, for several values of $\varepsilon$. As we apply order one corrections, we see that our method shows better results for small values of $\varepsilon$, but also gives very satisfactory results for larger values of $\varepsilon$.

## 4.5 Proof of Proposition 4.1

In order to prove the differentiability of the end-point mapping, we start with the differentiability with respect to one component. The proof relies heavily on the expansion (4.3), that we recall first. For the sake of completness, we will also give the proof of this result.

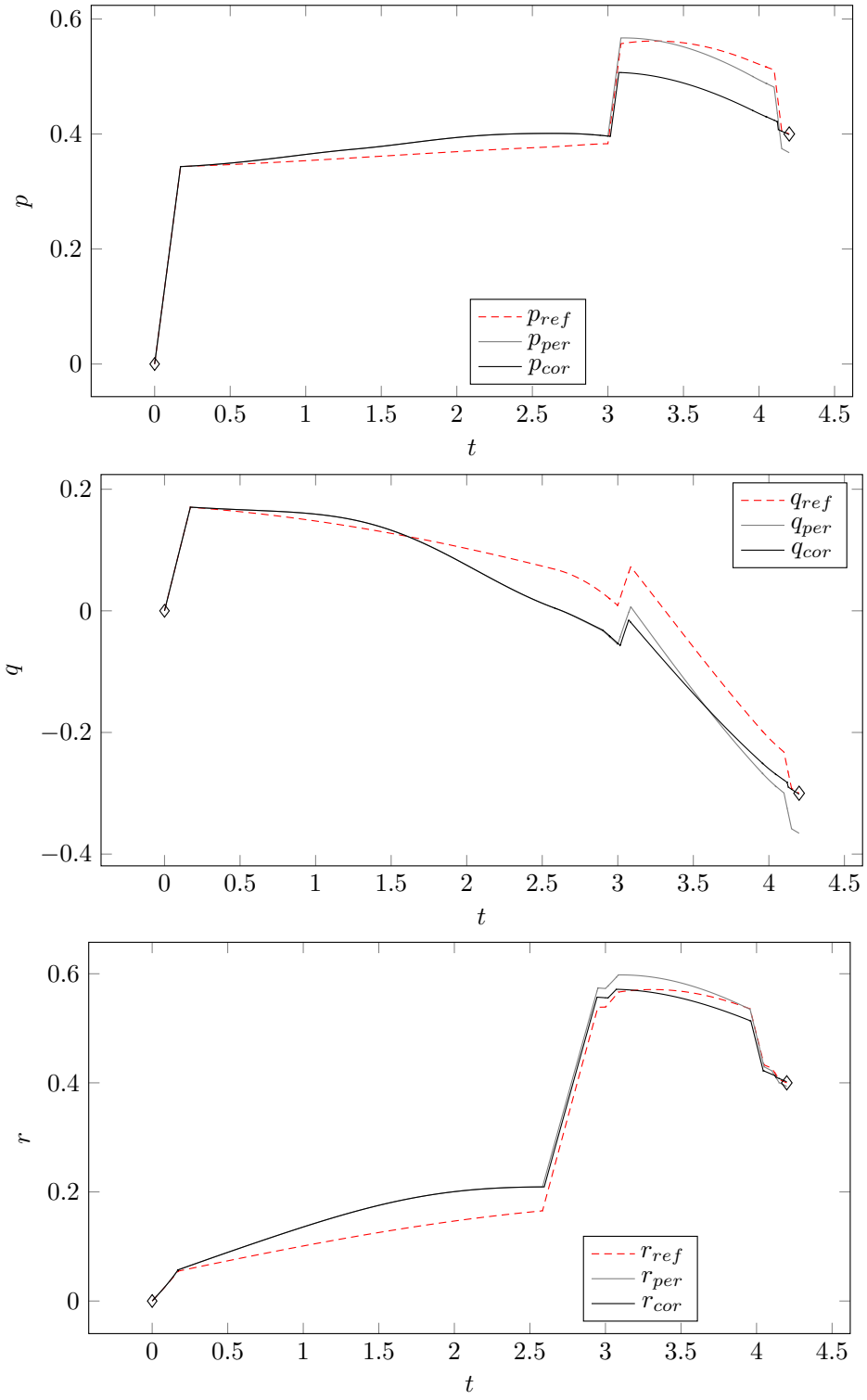



Figure 4.6 – Reference, perturbed and corrected trajectories for $\varepsilon = 0.78$, $C_r = 2.22$.

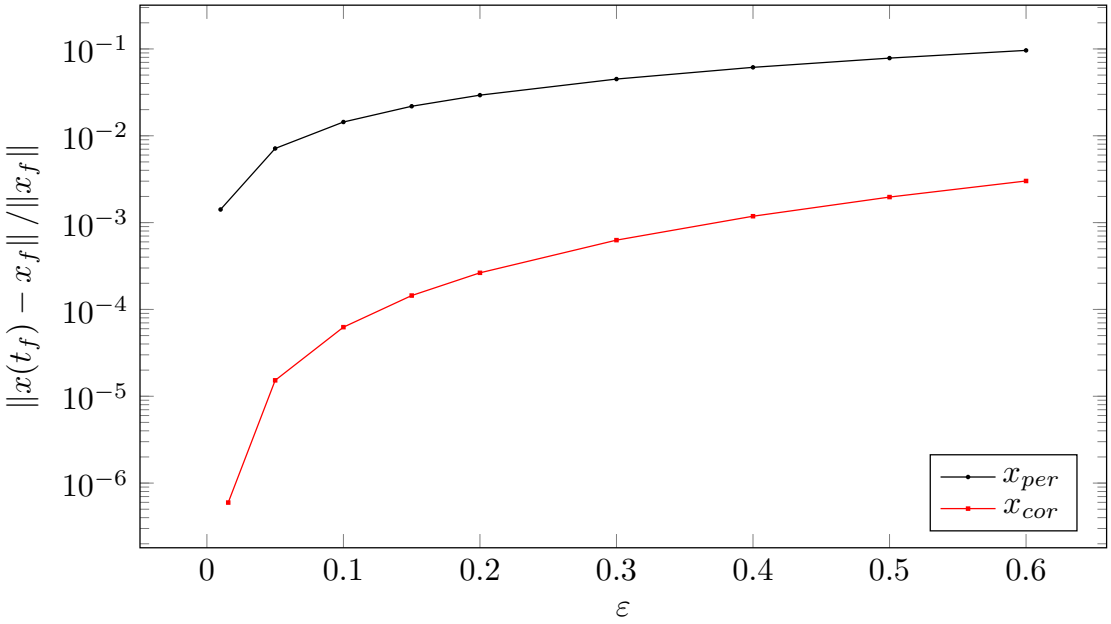



Figure 4.7 – Tracking results for several values of $\varepsilon$.

**Lemma 4.1.** – *Let $t_1 \in [0, t_f[$, and let $u_{\pi_1}(\cdot)$ be a needle-like variation of $u(\cdot)$, with $\pi_1 = (t_1, \delta t_1, u_1)$. Then*

$$x_{\pi_1}(t_f) = x(t_f) + |\delta t_1|\, v_{\pi_1}(t_f) + \mathrm{o}(\delta t_1),$$

*where $v_{\pi_1}(\cdot)$ is the solution of a Cauchy problem on $[t_1, t_f]$*

$$\begin{aligned} \dot{v}_{\pi_1}(t) &= \frac{\partial f}{\partial x}(t, x(t), u(t)) v_{\pi_1}(t), \\ v_{\pi_1}(t_1) &= f(t_1, x(t_1), u_1) - f(t_1, x(t_1), u(t_1)). \end{aligned}$$

*Proof.* Let $\pi_1 = (t_1, \delta t_1, u_1)$ be a needle-like variation. The trajectory $x_{\pi_1}(\cdot)$ is a solution of the equation

$$\begin{aligned} x_{\pi_1}(t_f) &= x(0) + \int_0^{t_f} f(t, x_{\pi_1}(t), u_{\pi_1}(t))\, dt \\ &= x(0) + \int_0^{t_1} f(t, x_{\pi_1}(t), u(t))\, dt + \int_{t_1}^{t_f} f(t, x_{\pi_1}(t), u_{\pi_1}(t))\, dt \\ &= x(t_1) + \underbrace{\int_{t_1}^{t_1+\delta t_1} f(x_{\pi_1}(t), u_{\pi_1}(t))\, dt}_{:=A} + \underbrace{\int_{t_1+\delta t_1}^{t_f} f(x_{\pi_1}(t), u_{\pi_1}(t))\, dt}_{:=B} \end{aligned}$$

For almost very point $t_1 \in [0, t_f[$,

$$\begin{aligned} A &= \int_{t_1}^{t_1+\delta t_1} f(x_{\pi_1}(t), u_{\pi_1}(t))dt \\ &= \delta t_1 \cdot f(\overline{x}(t_1), u_1) + o(\eta_1) \end{aligned}$$

For the second term, splitting the integral,

$$\begin{aligned} B &= \int_{t_1+\delta t_1}^{t_f} f(x_{\pi_1}(t), u_{\pi_1}(t))\, dt \\ &= \int_{t_1+\delta t_1}^{t_f} f(x_{\pi_1}(t), u(t))\, dt \\ &= \int_{t_1}^{t_f} f(x_{\pi_1}(t), u(t))dt - \int_{t_1}^{t_1+\delta t_1} f(x_{\pi_1}(t), u(t))\, dt \\ &= \int_{t_1}^{t_f} f(x_{\pi_1}(t), u(t))\, dt - \delta t_1 \cdot f(x(t_1), u(t_1)) + o(\delta t_1) \end{aligned}$$

Thus,

$$\begin{aligned} x_{\pi_1}(t_f) &= x(t_1) + \delta t_1 \left(f(x(t_1), u_1) - f(x(t_1), u(t_1))\right) + \int_{t_1}^{t_f} f(x_{\pi_1}(t), u(t))dt + o(\delta t_1) \\ &= x(t_f) + \delta t_1 v_{\pi_1}(t_1) + \int_{t_1}^{t_f} \left(f(x_{\pi_1}(t), u(t)) - f(x(t), u(t))\right) dt + o(\delta t_1). \end{aligned}$$

But also, following the definition of Lemma 4.1,

$$v_{\pi_1}(t_f) = v_{\pi_1}(t_1) + \int_{t_1}^{t_f} \frac{\partial f}{\partial x}(x(t), u(t)) v_{\pi_1}(t)\, dt.$$

Joining the previous inequalities together, we get that

$$\begin{aligned} \left| \frac{x_{\pi_1}(t_f) - x(t_f)}{\delta t_1} - v_{\pi_1}(t_f) \right| &= \left| \int_{t_1}^{t_f} \left( \frac{f(x_{\pi_1}(t), u(t)) - f(x(t), u(t))}{\delta t_1} - \frac{\partial f}{\partial x}(x(t), u(t)) v_{\pi_1}(t) \right) dt + o(1) \right| \\ &\leqslant \int_{t_1}^{t_f} \left| \frac{\partial f}{\partial x}(x(t), u(t)) \cdot \left( \frac{x_{\pi_1}(t) - x(t)}{\delta t_1} - v_{\pi_1}(t) \right) \right| dt + o(1) \end{aligned}$$

Let $\varepsilon > 0$. For $\delta t_1$ small enough, we have

$$\left| \frac{x_{\pi_1}(t_f) - x(t_f)}{\delta t_1} - v_{\pi_1}(t_f) \right| \leqslant \varepsilon + \int_{t_1}^{t_f} C \left| \frac{x_{\pi_1}(t) - x(t)}{\delta t_1} - v_{\pi_1}(t) \right| dt$$

Then, thanks to Grönwall inequality,

$$\left| \frac{x_{\pi_1}(t_f) - x(t_f)}{\delta t_1} - v_{\pi_1}(t_f) \right| \leqslant \varepsilon e^{C(t_f - t_1)}$$

which is equivalent to $x_{\pi_1}(t_f) = x(t_f) + \delta t_1 v_{\pi_1}(t_f) + \mathrm{o}(\delta t_1)$. □

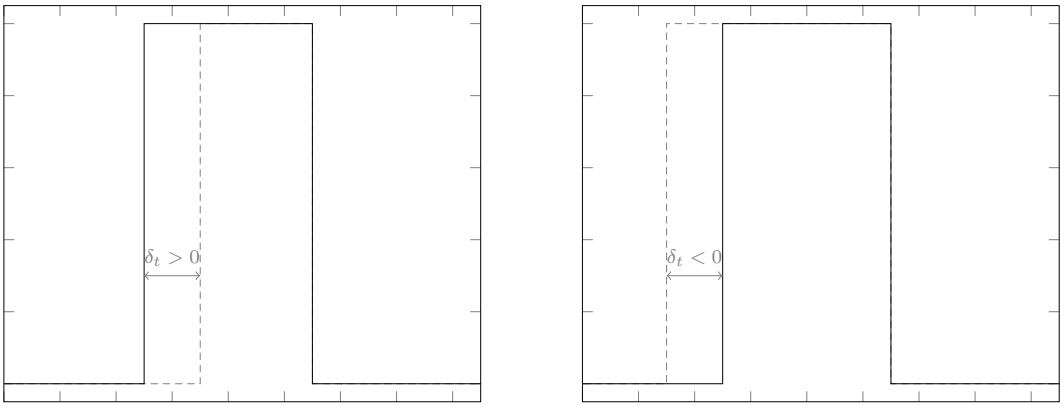
**Proposition 4.3.** – *We denote by $u$ the control $(t_1, \ldots, t_N, t_f)$ and $x(\cdot)$ the associated tra-*

*jectory of the control system. Let $\delta t_1 \in \mathbb{R}$ be small enough. Then*

$$E(t_1 + \delta t_1, t_2, \dots, t_N, t_f) = E(t_1, \dots, t_N, t_f) + \delta t_1 \cdot v_1(t_f) + o(\delta t_1),$$

*where $v_1(\cdot)$ is the solution of the Cauchy problem on $[t_1, t_f]$:*

$$\begin{aligned} \dot{v}_1(t) &= \frac{\partial f}{\partial x}(t, x(t), u(t))v_i(t), \\ v_1(t_1) &= \begin{cases} f(t_1, x(t_1), (\dots, a_{i_1}, \dots)) - f(t_1, x(t_1), u(t_1^+)) & \text{if } u_{i_1} \text{ switches from } a_{i_1} \text{ to } b_{i_1}. \\ f(t_1, x(t_1), (\dots, b_{i_1}, \dots)) - f(t_1, x(t_1), u(t_1^-)) & \text{if } u_{i_1} \text{ switches from } b_{i_1} \text{ to } a_{i_1}. \end{cases} \end{aligned}$$




(a) $\delta t_1 > 0$. (b) $\delta t_1 < 0$.

Figure 4.8 – Shifting an opening time is equivalent to add a needle.

*Proof.* Assume that at time $t_1$ the control $u_{i_1}$ switches from $a_{i_1}$ to $b_{i_1}$, and that $\delta t_1 > 0$. Let us define the needle-like variation $\pi = (t_1, \delta t_1, a_{i_1})$ for the $i_1$-th component of the control. Then, the control $u_\pi$ is represented by the vector $(t_1 + \delta t_1, \dots, t_N, t_f)$ (figure 4.8): adding the needle-like variation $\pi$ to the $i_1$-th component, with value $a_{i_1}$ and length $\delta t_1$ is equivalent to shifting the opening time to $t_1 + \delta t_1$. Thus, we have that $u(t_1^+)_{i_1} = b_{i_1}$ and $u_\pi(t_1^+)_{i_1} = a_{i_1}$. Hence, we obtain that, according to lemma 4.1

$$x_\pi(t_f) = x(t_f) + \delta t_1 \cdot v_1(t_f) + o(\delta t_1), \tag{4.22}$$

where $v_1(\cdot)$ is the solution of the Cauchy problem:

$$\begin{aligned} \dot{v}_1(t) &= \frac{\partial f}{\partial x}(t, x(t), u(t))v_1(t), \\ v_1(t_1) &= f(t_1, x(t_1), u_\pi(t_1^+)) - f(t_1, x(t_1), u(t_1^+)) \\ &= f(t_1, x(t_1), (\dots, a_{i_1}, \dots)) - f(t_1, x(t_1), (\dots, b_{i_1}, \dots)). \end{aligned}$$

(Between $u_\pi(t_1^+)$ and $u(t_1^+)$, only the $i_1$-th component differs.)

If $\delta t_1 < 0$, define the variation $\pi = (t_1, \delta t_1, 1)$ for the $i_1$-th component of the control. Then again, the control $u_\pi$ is represented by the vector $(t_1 + \delta t_1, \dots, t_N, t_f)$ (figure 4.8). Thus, we

have that $u(t_1^-)_j = a_{i_1}$ and $u_\pi(t_1^-)_{i_1} = 1$. Thanks to lemma 4.1, we obtain that

$$x_\pi(t_f) = x(t_f) - \delta t_1 \cdot w_1(t_f) + o(\delta t_1), \tag{4.23}$$

where $w_1(\cdot)$ is the solution of the Cauchy problem:

$$\begin{aligned}
\dot{w}_1(t) &= \frac{\partial f}{\partial x}(t, x(t), u(t)) w_1(t), \\
w_1(t_1) &= f(t_1, x(t_1), u_\pi(t_1^-)) - f(t_1, x(t_1), u(t_1^-)) \\
&= f(t_1, x(t_1), (\ldots, b_{i_1}, \ldots)) - f(t_1, x(t_1), (\ldots, a_{i_1}, \ldots)) \\
&= -v_1(t_1).
\end{aligned}$$

Thus, by uniqueness we have $w_1 = -v_1$, and from (4.22) and (4.23), we obtain:

$$x_\pi(t_f) = x(t_f) + \delta t_1 \cdot v_1(t_f) + o(\delta t_1).$$

We can proceed the exact same way if at $t_1$, the control $u_{i_1}$ switches from $b_{i_1}$ to $a_{i_1}$ □

The general result at proposition 4.1 follows by an immediate iteration.

## 4.6 Conclusion of the chapter and perspectives

Starting with the expansion of the end-point mapping with respect to a needle like variation, we have shown in this chapter how redundant switching times can be added in order to make a control more robust, for general control systems of the form $\dot{x}(t) = f(t, x(t), u(t))$. Those additional switching times can be seen as extra degrees of freedom used to absorb perturbations. A potential application is to start from a bang-bang solution of an optimal control problem, that is usually not robust, and make it more robust. A compromise is then to be found between loss of optimality and gain of robustness. This is why we have designed a measure of robustness, as follows.

In the presence of a perturbation $\delta x$, the correction to apply to the switching times is the solution of an equation $dE \cdot \delta\mathcal{T} = \delta x$. It is natural to try to solve this equation while shifting the switching times as little as possible. The least-squares problem formulation is then the appropriate setting to find the solution of minimal (euclidian) norm of the previous equation, and it is given by $\delta\mathcal{T} = dE^\dagger \cdot \delta x$, for which we have the norm estimation $\|\delta\mathcal{T}\|_2 \leqslant \|\delta x\|_2 / \sigma_{\min}$. This enabled us to identify the measure for robustness:

$$\int \frac{1}{\sigma_{\min}(t)^2}\, dt.$$

The numerical example studied in Section 4.4 remains academic, and was used to legitimize the theoretical ideas explained previously. In a future work, we aim at applying the method to the complete (and more complex) attitude control system of a three-dimensional rigid body presented in the Introduction, for which we wish to control the angular velocity, as well as the orientation with respect to a fixed reference frame as written at equation (3.3.1). To the three velocity variables will be added three angles to parametrize the orientation of the body. Thus, a challenge will come from the dimension of the state space (6), as well as the potentially bigger number of needle-like variations required to robustify a trajectory.

Chapitre 5

# Combination of Direct Methods and Homotopy in Numerical Optimal Control : Application to the Optimization of Chemotherapy in Cancer

## Contents


**4.1 Introduction of the chapter . . . . . 68**
4.1.1 Overview of the method . . . . . 68
4.1.2 State of the art on robust control design . . . . . 71
4.1.3 Structure of this chapter . . . . . 73
**4.2 Tracking algorithm . . . . . 73**
4.2.1 Reduced end-point mapping . . . . . 73
4.2.2 Absorbing perturbations . . . . . 75
**4.3 Promoting robustness . . . . . 80**
4.3.1 An auxiliary optimization problem . . . . . 81
4.3.2 Redundancy creates robustness . . . . . 81
**4.4 Numerical results . . . . . 84**
4.4.1 Computing the nominal trajectory . . . . . 84
4.4.2 Robustifying the nominal trajectory . . . . . 85
**4.5 Proof of Proposition 4.1 . . . . . 87**
**4.6 Conclusion of the chapter and perspectives . . . . . 92**


The setting of this chapter differs from the previous ones, and we will consider an optimal control problem in infinite dimension modelling the evolution of cell populations. We will use two elements previously introduced in this work, namely direct methods and continuation techniques, and combine them to solve the problem at hand.

Even if it does not directly concern aerospace applications, we claim that our approach could be used quite generically, each time it is possible to simplify enough a control problem to start the numerical resolution.

## 5.1 Introduction of the chapter

The motivation for this work is the article [PCLT17], itself initiated by [LLC+13]. In the former, the subject was the theoretical and numerical analysis of an optimal control problem coming from oncology. Through chemotherapy, it consists of minimizing the number of cancer cells at the end of a given therapeutic window. The underlying model was an integro-differential system for the time-evolution of densities of cancer and healthy cells, structured by their continuous level of resistance to chemotherapeutic drugs. The model took into account cell proliferation and death, competition between the cells, and the effect of chemotherapy on them. The optimal control problem also incorporated constraints on the doses of the drugs, as well as constraints on the tumor size and on the healthy tissue.

In [PCLT17], the numerical resolution of the optimal control problem was made through a direct method, thanks to a discretization both in time and in the phenotypic variable. It led to a complex nonlinear constrained optimization problem, for which even efficient algorithms will fail for large discretization parameters because they require a good initial guess. To overcome this, the idea was to perform (with AMPL and IPOPT, see below) a continuation on the discretization parameters, starting from low values (*i.e.*, a coarse discretization) for which the optimization algorithm converges regardless of the starting point.

A clear optimal strategy emerged from these numerical simulations when the final time was increased. It roughly consists of first using as few drugs as possible during a long first phase to avoid the emergence of resistance. Cancer cells would hence concentrate on a sensitive phenotype, allowing for an efficient short second phase with the maximum tolerated doses.

The model of [PCLT17] did not include *epimutations*, namely heritable changes in DNA expression which are passed from one generation of cells to the others, which are believed to be very frequent in the life-time of a tumor. Our aim here is to numerically address the optimal control problem with the epimutations modeled through diffusion operators (Laplacians), in order to test the robustness of the optimal strategy.

However, the previous numerical technique already failed (even without Laplacians) to get fine discretizations when the final time is very large : the optimization stops converging when the discretization parameters are large. The values reached for the discretization in time were enough to observe the optimal structure, in particular all the arcs that were expected for theoretical reasons.

The addition of Laplacians significantly increases the run-time and again fails to work once the discretization parameters are too large when the final time itself is large, and some arcs become difficult to observe. We thus have to find an alternative method to see whether the optimal strategy found in [PCLT17] is robust with respect to adding the effect of epimutations.

This chapter is devoted to the presentation of a method which, up to our knowledge, is new. In our case, it provides a significant improvement in run-time and precision, and shows that the optimal strategy keeps an analogous structure when epimutations are considered. The method relies on the two following steps :

- first, simplify the optimal control problem up to a point where we can show that, thanks to a Pontryagin Maximum Principle (PMP) in infinite dimension, the optimal controls are bang-bang and thus can be reduced to their switching times, which are very easy to estimate numerically. This is equivalent to setting several coefficients to 0 in the model.
- second, perform a continuation on these parameters on the optimization problems obtained with a direct method, starting from the simplified problem all the way back to the full optimal control problem.

It allows us to start the homotopy method on this simplified optimization problem with an already fine discretization, actually much finer than the maximal values which could be obtained

with the previous homotopy method. We also believe that the theoretical result obtained for the simplified optimal control problem can serve as the starting step for many other optimal control problems of related models in mathematical biology.

**Numerical optimal control and novelty of the approach.** Discretizing the time variable, control and state variables to approximate a control problem for an ODE (which is an optimization problem in infinite dimension) by a finite-dimensional optimization problem has now become the most standard way of proceeding. These so-called direct methods thus lead to using efficient optimization algorithms, for example through the combination of automatic differentiation softwares (such as the modeling language AMPL, see [FGK02]) and expert optimization routines (such as the open-source package IPOPT, see [WB06b]).

Another approach is to use indirect methods, where the whole process relies on a PMP, leading to a shooting problem on the adjoint vector. Numerically, one thus needs to find the zeros of an appropriate function, which is usually done through a Newton-like algorithm. For a comparison of the advantages and drawbacks of direct and indirect methods, we refer to the survey [Tré12].

For both direct and indirect methods, the numerical problem shares at least the difficulty of finding an initial guess leading to convergence of the optimization algorithm or the Newton algorithm, respectively (it is well known that Newton algorithms can have a very small domain of convergence). To tackle this issue in the case of indirect methods, it is very standard to use homotopy techniques, for instance to simplify the problem so that one can have a good idea for a starting point as in [CHT12, CHT17], or to change the cost in order to benefit from convexity properties, as in [GH06, CDG12]. Besides, when studying optimal control problem for ODE systems, a common approach is the use of so-called hybrid methods, in order to take advantage from the better convergence properties of the direct method and the high accuracy provided by the indirect method. We refer to [Tré12, BNPvS93, Pes94, vSB92] for further developments on this subject.

We have found the combination of direct methods and continuation (such as the one done in [PCLT17]) to be much less common in the literature, see however [BNPvS93]. For a mathematical investigation of why continuation methods are mathematically valid, see [Tré12].

It is however believed that direct methods typically lead to optimization problems with several local minima [Tré12], as it could happen for the starting problem (with low discretization), which has yet no biological meaning. This implies one important drawback of a continuation on discretization parameters with direct methods : although the algorithm will quickly converge in such cases, one cannot *a priori* exclude that one will get trapped in local minima that are meaningless, with the possibility for such trapping to propagate through the homotopy procedure.

Our approach of simplifying the optimal control problem so that it can be analyzed with theoretical tools such as a PMP is a way to address the previous problem and to decrease the computation time. The simplified optimal control problem, once approximated by a direct method, will indeed efficiently be solved even with a very refined discretization. Therefore, another original aspect of our work, due to the complex PDE structure of the model, is the use of the PMP in view of building an initial guess for the direct method, in contrast with the hybrid approach we described for ODE systems, where direct methods serve to initialize shooting problems.

More generally, we advocate for the strategy of trying to simplify the problem, testing whether a PMP can provide a good characterization of the optimal controls. Then continuation with direct methods are performed to get back to the original and more difficult one. We believe that this can always be tried as a possible strategy to solve any optimal control problem (ODE or PDE) numerically.

**Outline of the chapter.** The chapter is organized as follows. Section 5.2 is devoted to a detailed presentation of the optimal control problem and the results that were obtained in [PCLT17]. Section 5.3 presents the simplified optimal control problem together with the application of a Pontryagin Maximum Principle in infinite dimension which almost completely determines the optimal controls. In Section 5.4, we thoroughly explain how direct methods for the optimal control of PDEs and continuations can be combined to solve a given PDE optimal control problem. We then combine these techniques and the result of Section 5.3 to build an algorithm solving the complete optimal control problem. In Section 5.5 the numerical simulations obtained thanks to the algorithm are presented. Finally, we will give some perspectives in Section 5.6 before concluding in Section 5.7.

## 5.2 Modeling Approach and Optimal Control Problem

### 5.2.1 Modeling Approach

Let us first explain the modeling approach, which is based on the classical logistic ODE

$$\frac{dN}{dt} = (r - dN)\, N.$$

In this setting, individuals $N(t)$ have a net selection rate $r$, together with an additional death term $dN$ increasing with $N$ : the more individuals, the more death due to competition for resources and space.

If the individuals have different selection and death rates $r(x)$ and $d(x)$ depending on a continuous variable $x$ which we will call *phenotype* (the size of the individual, for example), then a natural extension to the previous model is to study the density of individuals $n(t,x)$ of phenotype $x$, at time $t$, satisfying the integro-differential equation

$$\frac{\partial n}{\partial t}(t,x) = \big(r(x) - d(x)\rho(t)\big)n(t,x),$$

where

$$\rho(t) := \int n(t,x)\, dx.$$

At this stage, individuals do not change phenotype over time, nor can they give birth to offspring with different phenotypes. Accounting for such a possibility consists in modeling random mutations (respectively random epimutations), *i.e.*, heritable changes in the DNA (respectively heritable changes in DNA expression). The model is complemented with a diffusion term and takes the form

$$\frac{\partial n}{\partial t}(t,x) = \big(r(x) - d(x)\rho(t)\big)n(t,x) + \beta\Delta n(t,x),$$

together with Neumann boundary conditions if $x$ lies in a bounded domain, thus becoming a non-local partial differential equation because of the integral term $\rho$.

Such so-called *selection-mutation* models are actively studied as they represent a suitable mathematical framework for investigating how selection occurs in various ecological scenarios [D$^+$04, DJMP05, Per06], thus belonging to the branch of mathematical biology called adaptive dynamics. When $\beta = 0$, the previous model indeed leads to asymptotic selection : $n$ converges to a sum of Dirac masses located on the set of phenotypes on which $\frac{r}{d}$ reaches its maximum [PCLT17, Per06]. In particular, if this set is reduced to a singleton $x_0$ it holds that $\frac{n(t,\cdot)}{\rho(t)}$ weakly converges to a Dirac at $x_0$ as $t$ goes to $+\infty$.

### 5.2.2 The Optimal Control Problem

The model considered in this chapter is an extension of the one studied in [PCLT17] by the addition of epimutations (it is believed that mutations occur on a too long time-scale and are consequently neglected [CLC16]). It describes the dynamics of two populations of cells, healthy and cancer cells, which are both structured by a trait $x \in [0,1]$ representing resistance to chemotherapy, which ranges from sensitiveness ($x = 0$) to resistance ($x = 1$). $x$ is taken to be a continuous variable because resistance to chemotherapy can be correlated to biological characteristics which are continuous, see [CLC16] for more details. Chemotherapy is modeled by two functions of time $u_1$ and $u_2$, standing for the rate of administration of cytotoxic drugs and cytostatic drugs, respectively. The first type of drug actively kills cancer cells, while the second slows down their proliferation.

The system of equations describing the time-evolution of the density of healthy cells $n_H(t,x)$ and cancer cells $n_C(t,x)$ is given by

$$\begin{aligned}
\frac{\partial n_H}{\partial t}(t,x) &= \left[\frac{r_H(x)}{1+\alpha_H u_2(t)} - d_H(x) I_H(t) - u_1(t)\mu_H(x)\right] n_H(t,x) + \beta_H \Delta n_H(t,x),\\
\frac{\partial n_C}{\partial t}(t,x) &= \left[\frac{r_C(x)}{1+\alpha_C u_2(t)} - d_C(x) I_C(t) - u_1(t)\mu_C(x)\right] n_C(t,x) + \beta_C \Delta n_C(t,x),
\end{aligned}$$

starting from an initial condition $(n_H^0, n_C^0)$ in $C([0,1])^2$, with Neumann boundary conditions in $x = 0$ and $x = 1$.

Let us describe in more details the different terms and parameters appearing above, with the functions $r_H$, $r_C$, $d_H$, $d_C$, $\mu_H$ $\mu_C$ all continuous and non-negative on $[0,1]$, with $r_H$, $r_C$, $d_H$, $d_C$ positive on $[0,1]$.

- The terms $\frac{r_H(x)}{1+\alpha_H u_2(t)}$, $\frac{r_C(x)}{1+\alpha_C u_2(t)}$ stand for the selection rates lowered by the effect of the cytostatic drugs, with
$$\alpha_H < \alpha_C.$$
- The non-local terms $d_H(x)I_H(t)$, $d_C(x)I_C(t)$ are added death rates to the competition inside and between the two populations, with
$$I_H := a_{HH}\rho_H + a_{HC}\rho_C,\ I_C := a_{CC}\rho_C + a_{CH}\rho_H$$
and as before
$$\rho_i(t) = \int_0^1 n_i(t,x)\,dx,\ i = H, C.$$
We make the important assumption that the competition inside a given population is greater than between the two populations :
$$a_{HC} < a_{HH}, \quad a_{CH} < a_{CC}.$$
- The terms $\mu_H(x)u_1(t)$, $\mu_C(x)u_1(t)$ are added death rates due to the cytotoxic drugs. Owing to the meaning of $x = 0$ and $x = 1$, $\mu_H$ and $\mu_C$ are taken to be decreasing functions of $x$.
- The terms $\beta_H \Delta n_H(t,x)$ and $\beta_C \Delta n_C(t,x)$ model the random epimutations, with their rates $\beta_H$, $\beta_C$ such that
$$\beta_H < \beta_C,$$
because cancer cells mutate faster than healthy cells.

Finally, for a fixed final time $T$ we consider the optimal control problem (denoted in short by

(**OCPPDE$_1$**)) of minimizing the criterion

$$\lambda_0 \frac{1}{T} \int_0^T \rho_C(s)\, ds + (1-\lambda_0)\rho_C(T) \tag{5.1}$$

as a function of the $L^\infty$ controls $u_1$, $u_2$ subject to $L^\infty$ constraints for the controls and two state constraints on $(\rho_H, \rho_C)$, for all $0 \leqslant t \leqslant T$ :

- The maximum tolerated doses cannot be exceeded :

$$0 \leqslant u_1(t) \leqslant u_1^{max}, \;\; 0 \leqslant u_2(t) \leqslant u_2^{max}.$$

- The tumor cannot be too big compared to the healthy tissue :

$$\frac{\rho_H(t)}{\rho_H(t)+\rho_C(t)} \geqslant \theta_{HC}, \tag{5.2}$$

  with $0 < \theta_{HC} < 1$.

- Toxic side-effects must remain controlled :

$$\rho_H(t) \geqslant \theta_H \rho_H(0), \tag{5.3}$$

  with $0 < \theta_H < 1$.

Optimal control problems applied to cancer therapy have started being considered long ago, see [SL15] for a complete presentation. However, the usual way of taking resistance into account is to consider that cells are either resistant or sensitive, leading to ODE models, as for example in [CBB92, KS06, LS06, LS14, Car17]. Considering both a continuous modeling of resistance and the effect of chemotherapy is more recent, as in [PCLT17, CLC16, LLH+13, GLGL14, LCDH15]. We also mention some cases where an additional space variable is considered [LLC+13, LLC+15].

**Remark 5.1:**

Note that in the definition of the cost (5.1), the choice of $\lambda_0$ depends on the relative importance one wishes to give to the terms $\rho_C(T)$ and $\int_0^T \rho_C(s)\, ds/T$. By chosing $\lambda_0 = 0$ as in [PCLT17], the criterion to minimize becomes $\rho_C(T)$ and can be of interest in practice. In that case, even if the cost does no longer account for the evolution of $\rho_C(\cdot)$ over the time interval $[0,T]$, the size of the tumour cannot be too big as it remains controlled by the constraint (5.2) :

$$\frac{\rho_H(t)}{\rho_H(t)+\rho_C(t)} \geqslant \theta_{HC}.$$

### 5.2.3 Previous Results for $\lambda_0 = 0$

In [PCLT17], we studied this system and the optimal control problem both theoretically and numerically in the case of selection exclusively, namely for $\beta_H = \beta_C = 0$, while minimizing the number of tumour cells at final time, i.e. with $\lambda_0 = 0$ in the cost (5.1).

First, we proved that for constant controls (*i.e.*, constant doses), the generic behavior is the convergence of both densities to Dirac masses. When these doses are high, the model thus reproduces the clinical observation that high doses usually fail at controlling the tumor size on the long run. They might indeed initially lead to a decrease of the overall cancerous population.

However, this is the consequence of only the sensitive cells being killed, while the most resistant cells are selected (in our mathematical framework, this corresponds to the cancer cell density concentrating on a resistant phenotype). Further treatment is then inefficient and the tumor starts growing again.

As for the optimal control problem which is our focus in this work, the main findings without diffusion were the following : when the final time $T$ becomes large, the optimal controls acquire some clear structure which is made of two main phases.

- First, there is a long phase with low doses of drugs ($u_1 = 0$ with our parameters), along which the constraint (5.2) quickly saturates. At the end of this first long arc, both densities have concentrated on a sensitive phenotype.
- Then, there is a second short phase, which is the concatenation of two arcs. The first one is a free arc (no state constraint is saturated) along which $u_1 = u_1^{max}$ and $u_2 = u_2^{max}$, with a quick decrease of both cell numbers $\rho_H$ and $\rho_C$, up until the constraint on the healthy cells (5.3) saturates. The last arc is constrained on (5.3) with boundary controls ($u_2 = u_2^{max}$ with our parameters), allowing for a further decrease of $\rho_C$.

In other words, the optimal strategy is to let the cell densities concentrate on sensitive phenotypes so that the full power of the drugs can efficiently be used. This strategy is followed as long as the healthy tissue can endure it, and then lower doses are used to keep on lowering $\rho_C$ while still satisfying the toxicity constraint.

## 5.3 Resolution of a Simplified Model

### 5.3.1 Simplified Model for one Population with no State Constraints

We here introduce the simpler optimal control problem. Its precise link with the initial optimal control ($\mathbf{OCPPDE_1}$) will be explained in Section 5.4. It is based on the equation

$$\frac{\partial n_C}{\partial t}(t,x) = \left[\frac{r_C(x)}{1+\alpha_C u_2(t)} - d_C(x)\rho_C(t) - \mu_C(x)u_1(t)\right] n_C(t,x), \tag{5.4}$$

starting from $n_C^0$, where $\rho_C(t) = \int_0^1 n_C(t,x)\,dx$. We denote by ($\mathbf{OCPPDE_0}$) the optimal control problem

$$\min_{(u_1,u_2)\in\mathcal{U}} \rho_C(T) \tag{5.5}$$

where $\mathcal{U}$ is the space of admissible controls

$$\mathcal{U} := \{(u_1,u_2) \in L^\infty([0,T],\mathbb{R}) \text{ such that } 0 \leqslant u_1 \leqslant u_1^{max},\ 0 \leqslant u_2 \leqslant u_2^{max},\ \text{a.e. on } [0,T]\}\,.$$

Note that we choose $\lambda_0 = 0$ in the cost (5.1), in order for the Pontryagin Maximum Principle to yield an exploitable result.

### 5.3.2 A Maximum Principle in Infinite Dimension

**General statement.** Let $T$ be a fixed final time, $X$ be a Banach space and $n_0 \in X$, $U$ be a separable metric space. We also consider two mappings $f : [0,T]\times X\times U \to X$ and $f^0 : [0,T]\times X\times U \to \mathbb{R}$.

We consider the optimal control problem of minimizing an integral cost, with a free final state $n(T)$ :

$$\inf_{u\in\mathcal{U}} J(u(\cdot)) := \int_0^T f^0(t, n(t), u(t))\, dt,$$

where $y(\cdot)$ is the solution [1] of

$$\dot{n}(t) = f(t, n(t), u(t)), \quad n(0) = n_0.$$

In [LY12, Chapter 4], necessary conditions for optimality are presented, for such problems (they are actually presented in [LY12] in a more general setting, but for the sake of simplicity, we restrict ourselves to the material required to solve ($\mathbf{OCPPDE_0}$)). The set of these conditions is referred to as a Pontryagin Maximum Principle (PMP).

Under appropriate regularity assumptions on $f$ and $f^0$, it states that any optimal pair $(\overline{n}(\cdot), \overline{u}(\cdot))$ must be such that there exists a nontrivial pair $(p^0, p(\cdot)) \in \mathbb{R} \times C([0,T], X)$ satisfying

$$p^0 \leqslant 0, \tag{5.6}$$

$$\dot{p}(t) = -\frac{\partial H}{\partial n}(t, \overline{n}(t), \overline{u}(t), p^0, p(t)), \tag{5.7}$$

$$H(t, \overline{n}(t), \overline{u}(t), p^0, p(t)) = \max_{v\in U} H(t, \overline{n}(t), v, p^0, p(t)), \tag{5.8}$$

where the Hamiltonian $H$ is defined as $H(t, n, u, p, p^0) := p^0 f^0(t, n, u) + \langle p, f(t, n, u)\rangle$.

**Remark 5.2:**

If the final state is free, (5.6) can be improved to $p_0 < 0$ [a] and we have the additional transversality condition :

$$p(T) = 0. \tag{5.9}$$

Besides, if the final state were fixed, there would be additional assumptions to check in order to apply the PMP, assumptions that are automatically fulfilled whenever $n(T)$ is free. We refer to [LY12, Chapter 4 - Section 5] for more details on this issue.

*a.* An extremal in the PMP is said to be normal (resp. abnormal) whenever $p^0 \neq 0$ (resp. $p^0 = 0$). Here, it means that there is no abnormal extremal.

**Application to the problem ($\boldsymbol{OCPPDE_0}$).** By applying the PMP, we derive the following theorem on the optimal control structure.

**Theorem 5.1.** – *Let $(n_C(\cdot), u(\cdot))$ be an optimal solution for ($\boldsymbol{OCPPDE_0}$). There exists $t_1 \in [0, T[$ and $t_2 \in [0, T[$ such that*

$$u_1(t) = u_1^{max} \mathbb{1}_{[t_1, T]}, \qquad u_2(t) = u_2^{max} \mathbb{1}_{[t_2, T]}.$$

*Démonstration.* Let us define $U := \{u = (u_1, u_2)$ such that $0 \leqslant u_1 \leqslant u_1^{max},\ 0 \leqslant u_2 \leqslant u_2^{max}\}$. Given a function $u \in L^\infty([0,T], U)$, the associated solution of the equation (5.4) belongs to

1. Note that the evolution equation has to be understood in the mild sense

$$n(t) = n_0 + \int_0^t f(s, n(s), u(s))\, ds.$$

$C([0,T],C(0,1))$, which can be seen as a subset of $C([0,T],L^2(0,1))$. We define $X := L^2(0,1)$.

First, as the initial number of cells is prescribed, we notice that minimizing the cost $\rho_C(T)$ is equivalent to minimizing the cost $\rho_C(T) - \rho_C(0)$, and it can be written under the integral form :

$$\begin{aligned}\rho_C(T) - \rho_C(0) &= \int_0^T \rho_C'(t)\,dt \\ &= \int_0^T \int_0^1 \partial_t n_C(t,x)\,dx\,dt \\ &= \int_0^T \int_0^1 \left[\frac{r_C(x)}{1+\alpha_C u_2(t)} - d_C(x)\rho_C(t) - \mu_C(x)u_1(t)\right] n_C(t,x)\,dx\,dt\end{aligned}$$

Thus, in view of applying the PMP, we define the function $f^0 : X \times U \to \mathbb{R}$ by

$$f^0(n,u_1,u_2) := \int_0^1 \left[\frac{r_C(x)}{1+\alpha_C u_2} - d_C(x)\rho - \mu_C(x)u_1\right] n(x)\,dx,$$

where $\rho := \int_0^1 n$, and the Hamiltonian is then defined by

$$H(n,u_1,u_2,p,p^0) := p^0 f^0(n,u_1,u_2) + \int_0^1 p(x)\left[\frac{r_C(x)}{1+\alpha_C u_2} - d_C(x)\rho - \mu_C(x)u_1\right] n(x)\,dx.$$

Since $(n_C(\cdot),u(\cdot))$ is optimal, there exists a non trivial pair $(p^0,p(\cdot)) \in \mathbb{R}\times C([0,T],X)$, such that the adjoint equation (5.7) writes :

$$\frac{\partial p}{\partial t}(t,x) = -\left[\frac{r_C(x)}{1+\alpha_C u_2(t)} - d_C(x)\rho - \mu_C(x)u_1(t)\right]\cdot\left[p(t,x)+p^0\right] + \int_0^1 d(x)n(t,x)\left[p(t,x)+p^0\right]\,dx.$$

Owing to Remark 5.2, we know that $p^0 < 0$.

Let us set $\tilde{p} := p + p^0$, which satisfies

$$\frac{\partial \tilde{p}}{\partial t}(t,x) = -\left[\frac{r_C(x)}{1+\alpha_C u_2(t)} - d_C(x)\rho - \mu_C(x)u_1(t)\right]\tilde{p}(t,x) + \int_0^1 d(x)n(t,x)\tilde{p}(t,x)\,dx.$$

The transversality equation (5.9) yields $p(T,\cdot) = 0$, i.e., $\tilde{p}(T) = p^0$.

Then, in order to exploit the maximisation condition (5.8), we can split the Hamiltonian as

$$H(t,n_C(t),u_1(t),u_2(t),p(t),p^0) = -\int_0^1 p(t,x)d_C(x)\rho(t)n_C(t,x)\,dx - u_1(t)\phi_1(t) + \frac{\phi_2(t)}{1+\alpha_C u_2(t)},$$

where the two switching functions are defined as

$$\begin{aligned}\phi_1(t) &:= \int_0^1 \mu_C(x)n_C(t,x)\tilde{p}(t,x)\,dx, \\ \phi_2(t) &:= \int_0^1 r_C(x)n_C(t,x)\tilde{p}(t,x)\,dx.\end{aligned}$$

Thus, we derive the following rule to compute the controls :

- If $\phi_1(t) > 0$ (resp. $\phi_2(t) > 0$), then $u_1(t) = 0$ (resp. $u_2(t) = 0$).
- If $\phi_1(t) < 0$ (resp. $\phi_2(t) < 0$), then $u_1(t) = u_1^{max}$ (resp. $u_2(t) = u_2^{max}$).

We compute the derivative of the switching function :

$$\begin{aligned}\phi_1'(t) &= \int_0^1 \mu_C(x)\left(\partial_t n_C(t,x)\tilde{p}(t,x) + n_C(t,x)\partial_t\tilde{p}(t,x)\right)\,dx \\ &= \left(\int_0^1 \mu_C(x) n_C(t,x)\,dx\right)\cdot\left(\int_0^1 d_C(x) n_C(t,x)\tilde{p}(t,x)\,dx\right).\end{aligned}$$

We know that $\int_0^1 \mu_C(x) n_C(t,x)\,dx > 0$, so that the sign of $\phi_1'(t)$ is given by the sign of :

$$\int_0^1 d_C(x) n_C(t,x)\tilde{p}(t,x)\,dx.$$

Let us set $\psi_1(t) := \int_0^1 d_C(x) n_C(t,x)\tilde{p}(t,x)\,dx$. The same computation as before yields

$$\psi_1'(t) = \left(\int_0^1 d_C(x) n_C(t,x)\,dx\right)\psi_1(t).$$

Therefore, the sign of $\psi_1(t)$ is constant, given by the sign of

$$\begin{aligned}\psi_1(T) &= \int_0^1 d_C(x) n_C(T,x)\tilde{p}(T,x)\,dx \\ &= \int_0^1 d_C(x) n_C(T,x) p^0\,dx \\ &< 0\end{aligned}$$

since $p^0 < 0$. This implies that the function $\phi_1$ is decreasing on $[0,T]$. Since at the final time, $\phi_1(T) < 0$, we deduce the existence of a time $t_1 \in [0,T)$ such that $\phi_1(t) \geqslant 0$ on $[0,t_1]$, and $\phi_1(t) < 0$ on $[t_1,T]$. The same computation yields the same result for $\phi_2$, for some time $t_2 \in [0,T]$. □

## 5.4 The Continuation Procedure

### 5.4.1 General Principle

We here recall the principle of direct methods and of continuations for optimization problems. Together with Theorem 5.1, we then derive an algorithm to solve the problem (**OCPPDE$_1$**).

**On direct methods for PDEs.** Let us give an informal presentation of the principle of a direct method for the resolution of the optimal control of a PDE. Assume that we have some evolution equation written in a general form on $[0,T]\times[0,1]$ as

$$\frac{\partial n}{\partial t}(t,x) = f(t,n(t),u(t)) + An(t,x),\ n(0) = n^0,$$

where $T$ is a fixed time, $A$ is some operator on the state space, $f$ some function which might depend non-locally on $n$, $u$ a scalar control, $t \in [0,T]$, and $x \in [0,1]$ is the space or phenotype variable. The possible boundary conditions are contained in the operator $A$, which in our case will be the Neumann Laplacian.

Consider the optimal control problem

$$\inf_{u\in\mathcal{U}} g(n(T)),$$

where $T$ is fixed, as a function of $u \in \mathcal{U} := \{u \in L^\infty([0,T],\mathbb{R}),\ 0 \leqslant u(t) \leqslant u^{max} \text{ on } [0,T]\}$.

Further assume that we have discretized this PDE both in time and space through uniform meshes $0 < t_0 < t_1 < \ldots < t_{N_t} := T$, $0 =: x_0 < x_1 < \ldots < x_{N_x} := 1$, and that we are given some discretizations of the operator $A$ (resp. the function $f$, $g$) denoted by $A_h$ (resp. $f_h$, $g_h$), where $h := \frac{1}{N_x}$. With a Euler scheme in time, if one writes formally $n(t_i, x_j) \approx n_{i,j}$, $u(t_i) \approx u_i$ and $n_i := (n_{i,j})_{0\leqslant j\leqslant N_x}$, we are faced with the optimization problem

$$\inf_{u_i,\, 0\leqslant i\leqslant N_t} g_h\left(n_{N_t}\right),$$

subject to the constraints

$$n_{i+1,j} = n_{i,j} + hf_{h,j}(t_i, n_{i,j}, u_i) + hA_h(n_i),\ n_{i,0} = n^0(x_i),\ 0 \leqslant u_i \leqslant u^{max}$$

for all $0 \leqslant i \leqslant N_t,\ 0 \leqslant j \leqslant N_x$. Note that $f_{h,j}(t_i, n_{i,j}, u_i)$ stands for the function $f_h(t_i, n_{i,j}, u_i)$ evaluated at $x_j$.

**On continuation methods for optimization problems.** The optimal control problem of a PDE becomes a finite-dimensional optimization problem once approximated through a direct method, such as the one presented above. Let us denote $\mathcal{P}_1$ this problem. As already mentioned in the introduction, the numerical resolution of such a problem requires a good initial guess for the optimal solution. The idea of a continuation is to deform the problem to an easier problem $\mathcal{P}_0$ for which we either have a very good a priori knowledge of the optimal solution, or expect the problem to be solved efficiently.

One then progressively transforms the problem back to the original one thanks to a continuation parameter $\lambda$, thus passing through a series of optimization problems $(\mathcal{P}_\lambda)$. At each step of the procedure, the optimization problem $\mathcal{P}_{\lambda+d\lambda}$ is solved by taking the solution to $\mathcal{P}_\lambda$ as an initial guess.

### 5.4.2 From $(\mathbf{OCPPDE_1})$ to $(\mathbf{OCPPDE_0})$

Let us consider $(\mathbf{OCPPDE_1})$ and formally set the following coefficients to 0 :

$$\beta_H,\ \beta_C,\ a_{CH},\ \theta_H,\ \theta_{HC}.$$

Note that by setting $\beta_H$ and $\beta_C$ to 0, we also imply that the Neumann boundary conditions are no longer enforced.

When doing so, the equations on $n_C$ and $n_H$ are no longer coupled since the constraints do not play any role and the interaction itself (through $a_{CH}$) is switched off. Consequently, the optimal control problem with all these coefficients set to 0 is precisely $(\mathbf{OCPPDE_0})$.

We now define a family of optimal control problems $(\mathbf{OCPPDE}_\lambda)$ where $\lambda \in \mathbb{R}^5$ has each of its components between 0 and 1. It is a vector because several consecutive continuations will be performed (in an order to be chosen) on the different parameters. For $\lambda = (\lambda_i)_{0\leqslant i\leqslant 4}$, we use the subscript $\lambda$ for the parameters associated to the optimal control problem $(\mathbf{OCPPDE}_\lambda)$, and

they are defined by :

$$\beta_H^{(\lambda)} := \lambda_1 \beta_H, \quad \beta_C^{(\lambda)} := \lambda_1 \beta_C, \quad a_{CH}^{(\lambda)} := \lambda_2 a_{CH}, \quad \theta_{CH}^{(\lambda)} := \lambda_3 \theta_{CH}, \quad \theta_H^{(\lambda)} := \lambda_4 \theta_H,$$

In other words, $\lambda_1$, $\lambda_2$, $\lambda_3$ and $\lambda_4$ stand for the continuations on the epimutations rates, the interaction coefficient $a_{CH}$, the constraint (5.2) and the constraint (5.3), respectively. $\lambda_0$ accounts for the balance between the terms in the cost (5.1). Note that the parameters $\lambda_1$, $\lambda_2$, $\lambda_3$ and $\lambda_4$ are meant to be brought from 0 to 1, whereas the value of $\lambda_0$ may at the end lie in the interval $[0, 1]$.

### 5.4.3 General Algorithm

Let us now explain the general approach based on the previous considerations.

**Final objective and discretization.** Our final aim is to solve (**OCPPDE$_1$**) numerically, with $T$ large, and a very fine discretization in time ($N_t$ is taken to be large) : $T$, $N_t$ and $N_x$ are thus fixed to certain given values. To do so, we will solve successively several problems (**OCPPDE$_\lambda$**) with the same discretization paremeters. Following the general method introduced about direct methods for PDEs, numerically solving an intermediate optimal control problem (**OCPPDE$_\lambda$**) for a given $\lambda$ will mean solving the resulting optimization problem. To be more specific, we briefly explain below how the different terms are discretized. Recall that our discretization is uniform both in time $t$ and in phenotype $x$, with respectively $N_t$ and $N_x$ points.

- The non-local terms $\rho_H$, $\rho_C$ are discretized with the rectangle method :

$$\rho(t_i) = \int_0^1 n(t_i, x)\, dx \approx \frac{1}{N_x} \sum_{j=0}^{N_x - 1} n_{i,j}.$$

- The Neumann Laplacian is discretized by its classical discrete explicit counterpart :

$$\Delta n(t_i, x_j) \approx \frac{n_{i,j+1} - 2n_{i,j} + n_{i,j-1}}{(\Delta x)^2}.$$

  We manage to take $N_t$ large enough to make sure that the CFL

$$\beta_C T \frac{(N_x)^2}{N_t} < \frac{1}{2},$$

  is verified. Using an implicit discretization could allow us to get rid of the CFL condition but an implicit scheme happens to be more time-consuming. Therefore, we preferred using an explicit discretization, as our procedure enables us to discretize the equations finely enough to satisfy the CFL.

- The selection term (whose sign can be both positive or negative) is discretized through an implicit-explicit scheme to ensure unconditional stability.

**Sketch of the algorithm.**

*Step 1.* We start the continuation by solving (**OCPPDE$_0$**). Thanks to the result 5.1, finding the minimizer of the end-point mapping $(u_1, u_2) \longmapsto \rho_C(T)$ is equivalent to finding the minimizer of the application $(t_1, t_2) \longmapsto \rho_C(T)$ where $t_1$ (resp. $t_2$) are the switching times of $u_1$ (resp. $u_2$) from 0 to $u_1^{max}$ (resp. $u_2^{max}$), as introduced in Theorem 5.1.

Numerically, we can use an arbitrarily refined discretization of ($\mathbf{OCPPDE_0}$), since the resulting optimization problem has to be made on a $\mathbb{R}^2$-valued function, which leads to a quick and efficient resolution.

*Step 2.* Once ($\mathbf{OCPPDE_0}$) has been solved numerically, we get an excellent initial guess to start performing the continuation on the parameter $\lambda$. Its different components will successively be brought from 0 to 1 (except for $\lambda_0$ which will be brought from 0 to its final desired value), either directly or, when needed, through a proper discretization of the interval $[0, 1]$. The order in which the successive coefficients are brought to their actual values is chosen so as to reduce the run-time of the algorithm. The precise order and way in which the continuation has been carried out are detailed together with the numerical results in Section 5.4.

Let us make one remark on a possible further continuation : since the goal is to take large values for $T$, one might think of performing a continuation on the final time. We again emphasize that the interest and coherence of the method requires to start with a fine discretization at Step 1, but we note that it is also possible to further refine the discretization after Step 2.

## 5.5 Numerical Results

Let us now apply the algorithm with AMPL [FGK02] and IPOPT [WB06b].

For our numerical experiments, we will use the following values, taken from [LLC+13] :

$$r_C(x) = \frac{3}{1+x^2}, \quad r_H(x) = \frac{1.5}{1+x^2},$$

$$d_C(x) = \frac{1}{2}(1 - 0.3x), \quad d_H(x) = \frac{1}{2}(1 - 0.1x),$$

$$a_{HH} = 1, \quad a_{CC} = 1, \quad a_{HC} = 0.07, \quad a_{cH} = 0.01$$

$$\alpha_H = 0.01, \quad \alpha_C = 1,$$

$$\mu_H = \frac{0.2}{0.7^2 + x^2}, \quad \mu_C = \max\left(\frac{0.9}{0.7^2 + 0.6x^2} - 1, 0\right),$$

$$u_1^{max} = 2, \quad u_2^{max} = 5.$$

One can find in [PCLT17] a discussion on the choice of the functions $\mu_H$ and $\mu_C$. Also, we consider the initial data :

$$n_H(0,x) = K_{H,0} \exp\left(-\frac{(x-0.5)^2}{\varepsilon}\right), \quad n_C(0,x) = K_{C,0} \exp\left(-\frac{(x-0.5)^2}{\varepsilon}\right), \tag{5.10}$$

with $\varepsilon = 0.1$ and $K_{H,0}$ and $K_{C,0}$ are chosen such that :

$$\rho_H(0) = 2.7, \quad \rho_C(0) = 0.5.$$

The rest of the parameters (namely $\beta_H$, $\beta_C$, $\theta_H$ and $\theta_{HC}$) will depend on the case we consider, and we will specify them in what follows.

**Remark 5.3:**

Note that the initial condition (5.10) for the healthy and cancer cells - a Gaussian density centered at 0.5 - models a highly heterogeneous tumor, where resistance to the treatment is already present. Such a choice has been made because in the clinic, cytotoxic drugs are often given upfront. Our optimal strategy would therefore take place after this automatic administration of drugs.

**Remark 5.4:**

Note also that we have taken $u_1^{max}$ and $u_2^{max}$ to be slightly below their values chosen in [PCLT17] (which makes the problem harder from the applicative point of view). This is because we are here able to let $T$ take larger values, for which the final cost obtained with the optimal strategy $\rho_C(T)$ becomes too small, see below for the related numerical difficulties.
As for the epimutations rates, we have proceeded as follows : we have simulated the effect of constant doses and observed the long-time behavior. In the case $\beta_H = \beta_C = 0$, we know by [PCLT17] that both cell densities must converge to Dirac masses. With mutations, we expect some Gaussian-like approximation of these Diracs, the variance of which was our criterion to select a suitable epimutation rate in terms of modeling. It must be large enough to observe a real variability due to the epimutations, but small enough to avoid seeing no selection effects (diffusion dominates and the steady state looks almost constant).

**Test case 1 : $T = 60$ and $\lambda_0 = 0$.** We recall that this case corresponds to the example presented in [PCLT17], to which we add a diffusion term. We set the parameters for the diffusion to $\beta_H = 0.001$ and $\beta_C = 0.0001$. The coefficients for the constraints are $\theta_{HC} = 0.4$ and $\theta_H = 0.6$. For such numerical values, the optimal cost satisfies $\rho_C(T) << 1$, which can be source of numerical difficulties. To overcome this, we introduce the following trick : let us define $u_1^{max,0} = 1$ and $u_2^{max,0} = 4$. We apply the procedure described in Section 5.3 with the values $u_1^{max,0}$ and $u_2^{max,0}$. We then add another continuation step by raising them to the original desired values $u_1^{max} = 2$ and $u_2^{max} = 5$. In the formalism previously introduced, it amounts to adding two continuation parameters $\lambda_5$ and $\lambda_6$ to the vector $\lambda = (\lambda_i)_{1 \leqslant i \leqslant 4}$ (as we are interested in solving the problem for $\lambda_0 = 0$ in the cost (5.1)), we forget it in the notation of the vector $\lambda$). The parameters associated to the optimal control problem ($\mathbf{OCPPDE}_\lambda$) are then defined as :

$$u_1^{max,(\lambda)} := (1 - \lambda_5) u_1^{max,0} + \lambda_5 u_1^{max}, \quad u_2^{max,(\lambda)} := (1 - \lambda_6) u_2^{max,0} + \lambda_6 u_2^{max}.$$

More precisely, we perform the continuation in the following way, summarized in Figure 5.1 :

- First, we solve ($\mathbf{OCPPDE_0}$), with $u_1^{max,0} = 1$ and $u_2^{max,0} = 4$.
- Second, we add the interaction between the two populations, the diffusion parameters, and the constraint on the number of healthy cells. That is, the parameters $a_{CH}$, $\beta_H$, $\beta_C$ and $\theta_H$ are set to their values.
- Then, we add the constraint measuring the ratio between the number of healthy cells and the total number of cells, that is $\theta_{HC}$.
- Lastly, we raise the maximum values for the controls from $u_i^{max,0}$ to $u_i^{max}$ ($i \in \{1, 2\}$), and we solve ($\mathbf{OCPPDE_1}$) for $T = 60$.

Actually, for this set of parameters, only four consecutive resolutions are required to solve ($\mathbf{OCPPDE_1}$) starting from ($\mathbf{OCPPDE_0}$). That is, the components of the continuation vector

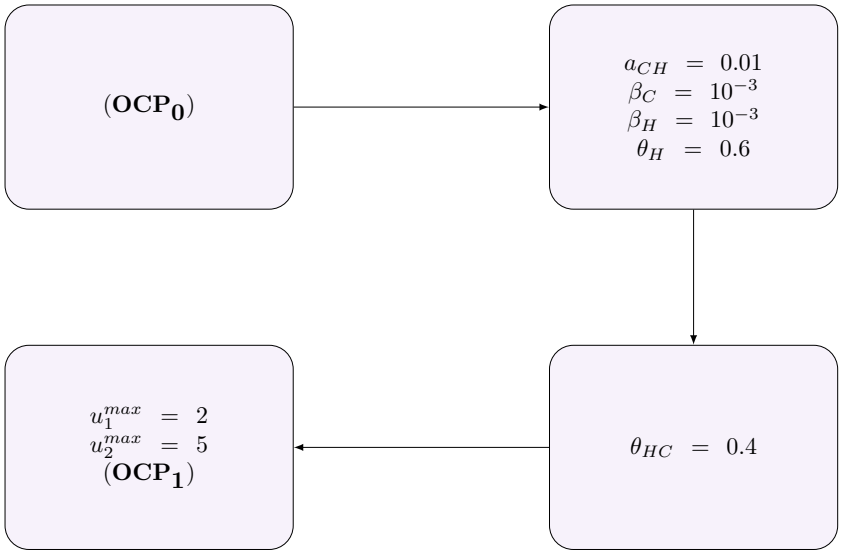



FIGURE 5.1 – Continuation procedure to solve $(\mathbf{OCPPDE_1})$ for $T = 60$.

$\lambda = (\lambda_i)_{1 \leqslant i \leqslant 6}$ are brought directly from 0 to 1, taking no intermediate value, in the order schematized on Figure 5.1. We will study further in the chapter a case for a larger final time, for which having a more refined discretization is mandatory.

On Figure 5.2, we plot the optimal controls $u_1$ and $u_2$ at the four steps of the continuation procedure. We also display the evolution of the constraint on the size of the tumor compared to the healthy tissue (5.2). We can clearly identify the emergence of the expected structure for the controls, namely a long phase along which the constraint (5.2) saturates, followed by a bang arc with $u_1 = u_1^{max}$ and $u_2 = u_2^{max}$, and a last boundary arc along which the constraint (5.3) saturates. Throughout this section, we will use a red solid line in our figures for $(\mathbf{OCPPDE_1})$, a light green solid line for $(\mathbf{OCPPDE_0})$ and colors varying from green to blue for anything referring to $(\mathbf{OCPPDE}_\lambda)$.

**Remark 5.5:**

We would like to emphasize here that our procedure enables us to use a much more refined discretization of the problem than what was done in [PCLT17]. More precisely, we discretize with $N_t = 500$ and $N_x = 20$ points in our direct method. For such a discretization, directly tackling $(\mathbf{OCPPDE_1})$ with the direct method fails.

**Remark 5.6:**

Note that the constraint $\rho_H/\rho_H(0) > 0.6$ does not saturate until the last step of the continuation, when raising the maximal value of the controls. Therefore, when we add it at the beginning of the procedure, it is not actually active.

**Test case 2 : $T = 80$ with $\lambda_0 = 0$.** Whereas one could believe that raising the final time from $T = 60$ to $T = 80$ does not much increase the difficulty of the problem, we noticed that several

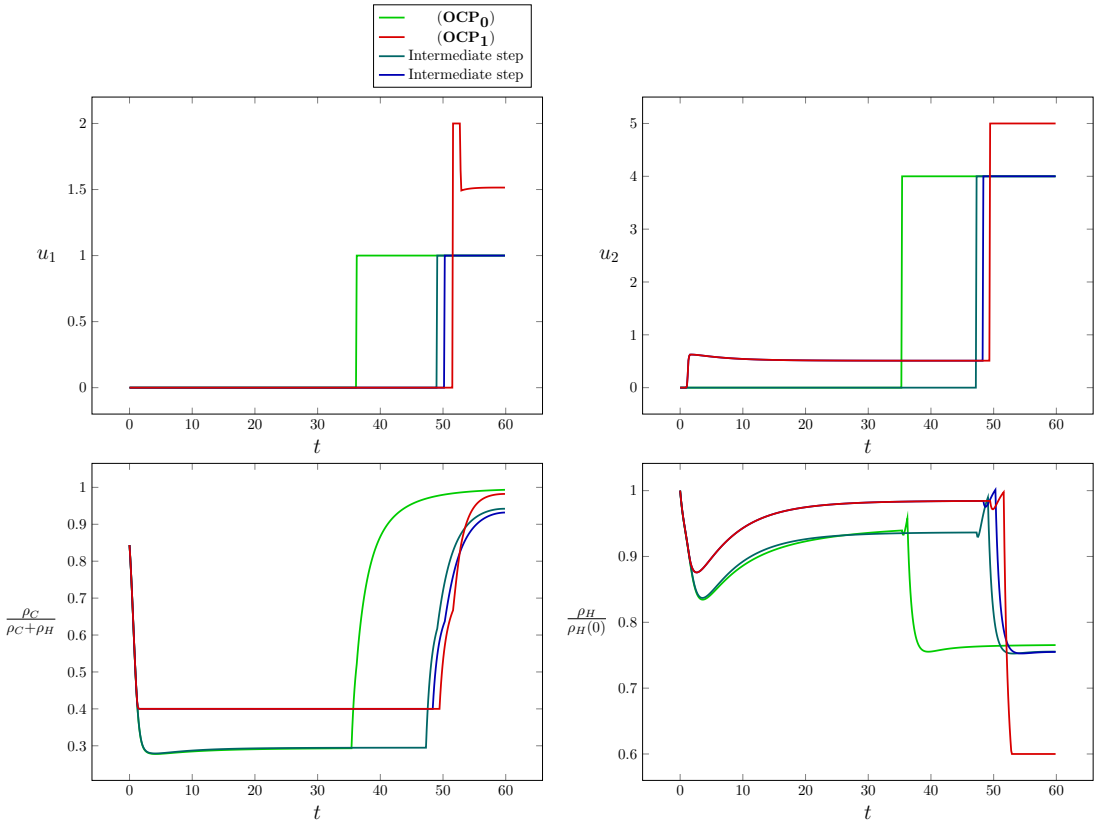



FIGURE 5.2 – Intermediate steps of the continuation procedure for the test case 1.

numerical obstacles appeared. In the following, we consider a discretization with $N_t = 250$ and $N_x = 12$ points, in order to keep the optimization run-time reasonable. Besides, in order to test the robustness of our procedure, we consider more restrictive constraints on the density of cells : we choose $\theta_H = 0.75$ in (5.3) (0.6 in the first example), and we also consider $\theta_{HC} = 0.6$ in (5.2) (0.4 in the first example). Note that setting a higher value for $\theta_{HC}$ means that the density of cancer cells is to be maintained below a lower level during the treatment.

First, we use the same numerical trick as explained in our first example, reducing the maximal value for the controls to $u_1^{max,0} = 0.7$ and $u_2^{max,0} = 3.5$. For given values of $u_1^{max}$ and $u_2^{max}$, the optimal cost $\rho_C(T)$ decreases when $T$ increases. This is why we now use smaller values of $u_1^{max,0}$ and $u_2^{max,0}$, compared to the first example where we set them to respectively 1 and 4.

We performed the continuation in the following way, summarized in Figure 5.3 :

- First, we solve ($\mathbf{OCPPDE_0}$), with $u_1^{max,0} = 0.7$ and $u_2^{max,0} = 3.5$.
- Second, we add the interaction between the two populations (via the parameter $a_{CH}$), and the constraint measuring the ratio between the number of healthy cells and the total number of cells (5.2) is introduced at the intermediate value $\theta_{HC}^{(\lambda)} = 0.3$.
- We then raise it to its final value of $\theta_{HC} = 0.6$.
- As a fourth step, we simultaneously add the constraint (5.3) on the healthy cells and raise the maximal values for the controls from $u_i^{max,0}$ to $u_i^{max}$ ($i \in \{1, 2\}$).

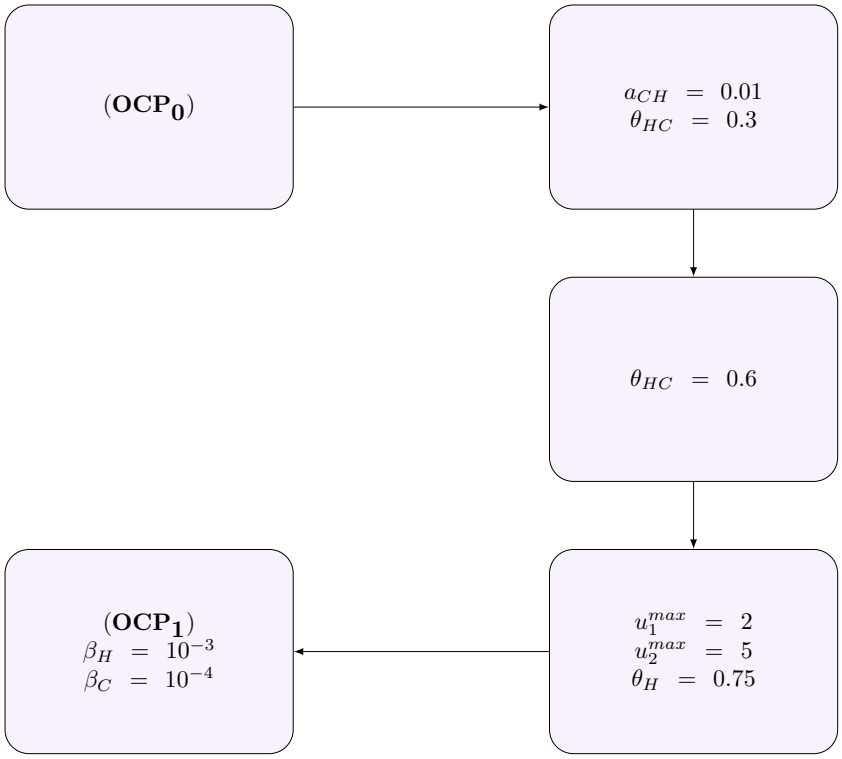



FIGURE 5.3 – Continuation procedure to solve $(\mathbf{OCPPDE_1})$ for $T = 80$.

- Lastly, we add diffusion to the model, via the parameters $\beta_H$ and $\beta_C$, and we solve $(\mathbf{OCPPDE_1})$ for $T = 80$.

At this point, we need to make two important remarks concerning this continuation procedure.

**Remark 5.7:**

The order in which we make the components of the continuation vector $\lambda = (\lambda_i)_{1 \leqslant i \leqslant 6}$ vary from 0 to 1 is different from the order we presented for $T = 60$. For instance, we noticed that the diffusion makes the problem significantly harder to solve, although the Laplacians where discretized using the simplest explicit finite-difference approximation. Therefore, we only added it at the last step of the continuation.

Whereas for $T = 60$, raising the $(\lambda_i)_{1 \leqslant i \leqslant 6}$ directly from 0 to 1 was enough to solve $(\mathbf{OCPPDE_1})$, it became necessary to use a more refined discretization for $T = 80$. This fact justifies the principle of our continuation procedure, as each step is necessary to solve the next one, and thus $(\mathbf{OCPPDE_1})$ in the end. For instance, on Figure 5.4, we display the evolution of the constraint (5.2) :

$$\frac{\rho_H(t)}{\rho_C(t) + \rho_H(t)} \geqslant \lambda_3 \theta_{HC}$$

when raising the continuation parameter $\lambda_3$ from 0 to 1. For values of $\lambda_3$ increasing from 0 to 1, the constraint (5.2) becomes more and more restrictive, but the continuation procedure enables

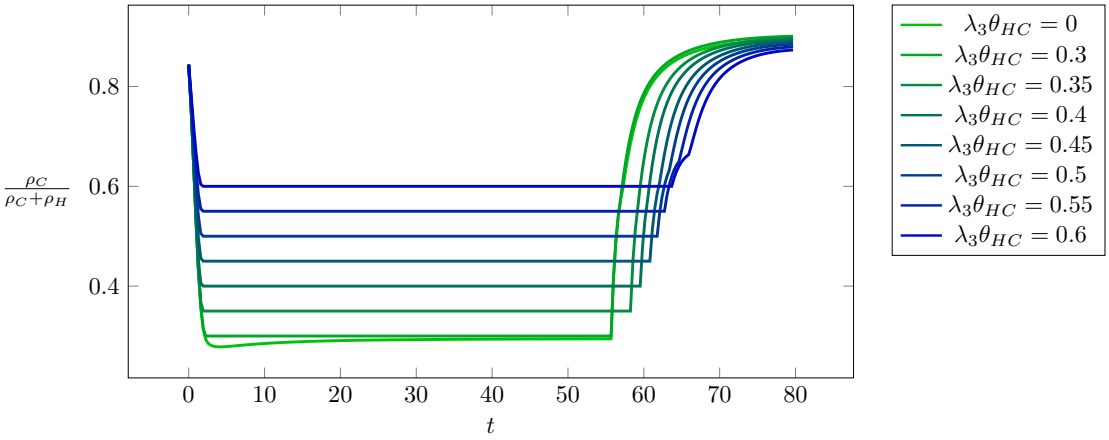



FIGURE 5.4 – Evolution of the constraint (5.2) during the continuation.

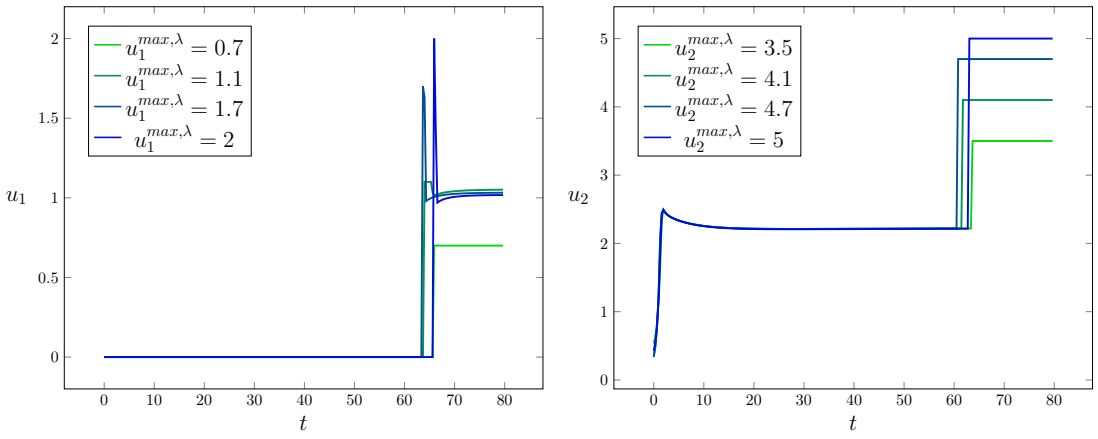



FIGURE 5.5 – Raising the maximal values $u_1^{max}$, $u_2^{max}$ for the controls.

us to reach the final value $\theta_{HC} = 0.6$. A noticeable fact is that compared to the test case 1, higher doses of cytostatic drugs are administered during the first phase. That is because, as pointed out before, the constraint (5.2) becomes more restrictive.

On Figure 5.5, we display the evolution of the controls $u_1$ and $u_2$ when raising their maximal allowed values from $(u_1^{max,0}, u_2^{max,0})$ to $(u_1^{max}, u_2^{max})$. For the sake of readability, we do not show all the steps of the continuation, but only some of them. It clearly shows how the structure of the optimal solution evolves from the simple one of $(\mathbf{OCPPDE_0})$ to the much more complex one of $(\mathbf{OCPPDE_1})$.

Finally, we display on Figure 5.6 the evolution of $n_C$, when applying the optimal strategy we found solving $(\mathbf{OCPPDE_1})$. One clearly sees that the optimal strategy has remained the same : the cancer cell population concentrates on a sensitive phenotype, around $x = 0.2$, which is the key idea to then use the maximal tolerated doses. In other words, the strategy identified in the previous work [PCLT17] is robust with respect to addition of epimutations. An important remark is that the cost obtained with the optimal strategy is higher with the mutations than without them : this is because we cannot have convergence to a Dirac located at a sensitive phenotype,

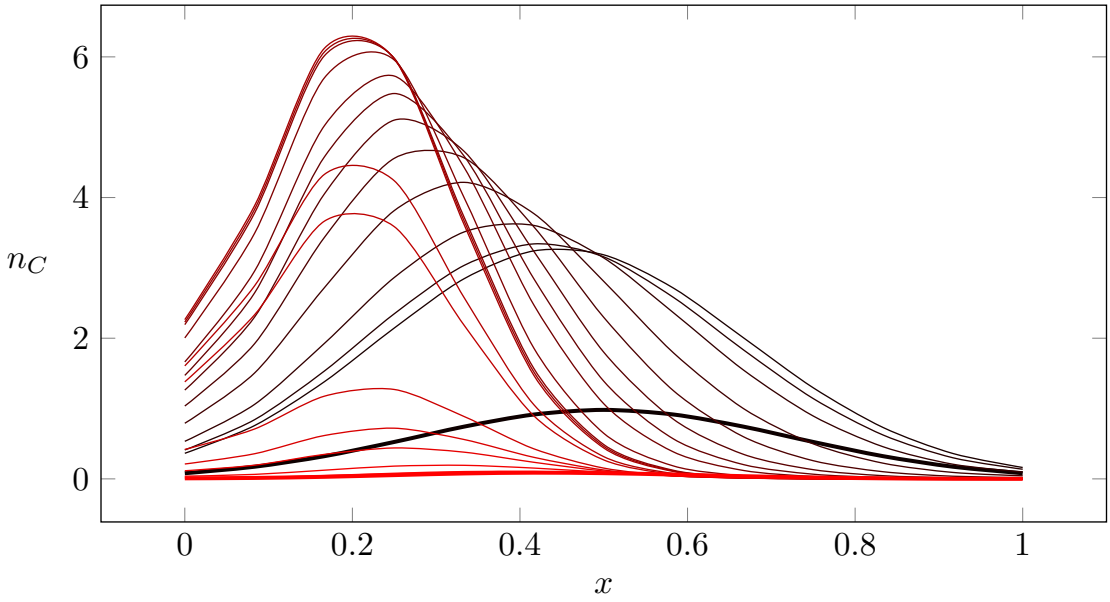



FIGURE 5.6 – Evolution of $n_C$ for the optimal solution of ($\mathbf{OCPPDE_1}$). In black with a thick line, the initial condition $n_C(0,\cdot)$, with lighter shades of red, the evolution of $n_C(t,x)$ as time increases. At final time, the population of cancer cells is drawn with a thick red line.

but to a smoothed (Gaussian-like) version of that Dirac. There will always be residual resistant cells which will make the second phase less successful.

**Further comments on the continuation principle.** A continuation procedure can be used in a wide range of applications, and one can easily imagine ways to generalize the ideas we have previously introduced. Let us illustrate our point with an example : we have presented a procedure to solve ($\mathbf{OCPPDE_1}$), for some initial conditions $n_H^0$ and $n_C^0$. Suppose that we wish to solve ($\mathbf{OCPPDE_1}$) for some different initial conditions $\tilde{n}_H^0$ and $\tilde{n}_C^0$. Biologically, this could correspond to finding a control strategy for a different tumor. A natural idea is then to use a continuation procedure to deform the problem from the initial conditions $(n_H^0, n_C^0)$ to $(\tilde{n}_H^0, \tilde{n}_C^0)$, rather than applying again the whole procedure to solve ($\mathbf{OCPPDE_1}$) with $\tilde{n}_H^0$ and $\tilde{n}_C^0$. We successfully performed some numerical tests to validate this idea : if we dispose of a set of initial conditions for which we want to solve ($\mathbf{OCPPDE_1}$), it is indeed faster to solve ($\mathbf{OCPPDE_1}$) for one of them and then perform a continuation on the initial data, rather than solving ($\mathbf{OCPPDE_1}$) for each of the initial conditions. More generally, any parameter in the model could lend itself to a continuation.

**Test case 3 : $T = 60$, for different values of $\lambda_0$.** The optimal strategy obtained with the previous objective function $\rho_C(T)$ might seem surprising, in particular because it advocates for very limited action at the beginning : giving no cytotoxic drugs and low loses of cytostatic drugs. To further investigate the robustness of this strategy, let us also consider the objective function $\lambda_0 \int_0^T \rho_C(s)\,ds + (1-\lambda_0)\rho_C(T)$ as introduced in Remark 5.1, for different values of $\lambda_0$. To ease numerical computations, we take $\beta_H = \beta_C = 0$, $u_1^{max} = 2$, $u_2^{max} = 5$, and finally $N_x = 20$,

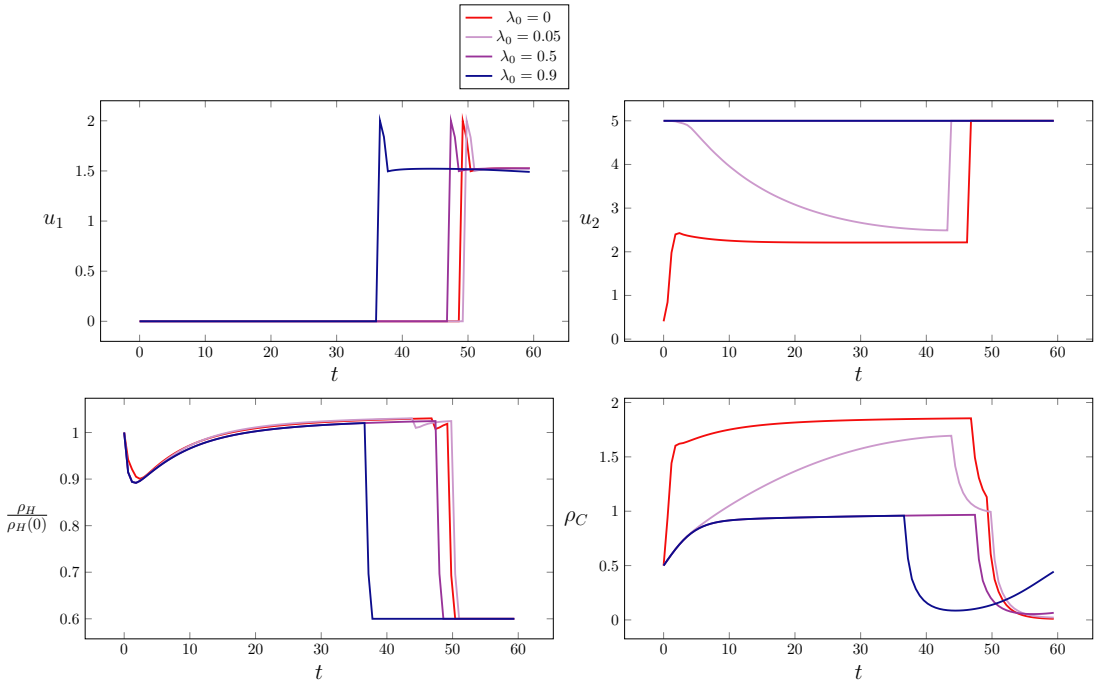



FIGURE 5.7 – Adding a term accounting for the $L^1$ norm $\int \rho_C$ in the cost.

$N_t = 100$. The results are reported on Figure 5.7.

For $\lambda_0 = 0.5$ (in purple) and $\lambda_0 = 0.9$ (in blue), the $L^1$ term is dominant in the optimization and the variations of $\rho_C$ are smaller over the interval $]0, T[$. However, although there is a significant change in the control $u_2$ which is always equal to $u_2^{max}$, $u_1$ has kept the same structure : an arc with no drugs, a short arc with maximal doses and a final arc with intermediate doses. The only (though important) difference is that the first arc is not a long one as before : for $\lambda_0 = 0.9$, the maximum dose of cytotoxic drugs is given earlier, around $t = 35$, in order to have a low $L^1$ term in the cost. However, in this case, cytotoxic drugs are given during a longer time period, making the tumor cells more resistant. This is supported by the representation of $\rho_C$ on the fourth graph of Figure 5.7, where $\rho_C$ increases during the last from $t = 65$ up to the end, because of the emergence of drug-resistant cells.

We infer from these numerical simulations that the optimal structure is inherent in the equations : there is no choice but to let the cancer cell density concentrate on a sensitive phenotype. Since at $\lambda_0 = 0.5$ and $\lambda_0 = 0.9$, the integral term dominates, we also consider other convex combinations with smaller values of $\lambda_0$, namely for $\lambda_0 = 0$ (in red) and $\lambda_0 = 0.05$ (in light purple) for which $u_2$ takes intermediate values (and even the maximum tolerated value during a short time when $\lambda_0 = 0.05$) before being equal to $u_2^{max}$, while $u_1 = 0$ on a longer arc.

## 5.6 Perspectives

**Theoretical perspectives.** A theoretical analysis of the problem (**OCPPDE$_1$**) is completely open. The first step in [PCLT17] in the absence of Laplacians was to analyse the asymptotic behavior for constant infusion of drugs, in which case the limit is the sum of Dirac masses on the

fittest phenotypes (depending on the drug). With Laplacians, however, the asymptotic analysis of the system

$$\frac{\partial n_H}{\partial t}(t,x) = \left[\frac{r_H(x)}{1+\alpha_H \bar{u}_2} - d_H(x) I_H(t) - \bar{u}_1 \mu_H(x)\right] n_H(t,x) + \beta_H \Delta n_H(t,x),$$
$$\frac{\partial n_C}{\partial t}(t,x) = \left[\frac{r_C(x)}{1+\alpha_C \bar{u}_2} - d_C(x) I_C(t) - \bar{u}_1 \mu_C(x)\right] n_C(t,x) + \beta_C \Delta n_C(t,x),$$

with constant controls $(\bar{u}_1, \bar{u}_2)$ below is not known, up to our knowledge. Actually, even the asymptotic analysis of a single equation of that type has not been tackled. Note that results are available when the functions $d_H$ and $d_C$ are independent of $x$, as in [LMM14]. The theoretical optimal control of a such a system with state constraints seems out of reach for the moment.

For epimutations with rates in reasonable ranges, we found that the optimal strategy obtained in [PCLT17] is preserved, which is a proof of its robustness. We believe that robustness can further be tested for more complicated models, with the same strategy.

For example, one may want to model longer-range mutations by a non-local alternative to the Laplacian, either through a mutation term through a Kernel [BCL15], or through a non-local operator like a fractional Laplacian [CR13]. These could both be added by continuation, on the Kernel starting from the integro-differential model, or on the fractional exponent for the fractional Laplacian, starting from the case of the (classical) Laplacian.

Another (local) possibility is to choose a more general elliptic operator. In particular, one can think of putting a drift term to model the *stress-induced adaptation* [Cov13, CLL16], namely epimutations that occur because cells actively change their phenotype in a certain direction depending on the environment created by the drug.

Finally, other objective functions can also be considered through a continuation as already introduced in the present paper : one minimizes a convex combination of $\rho_C(T)$ and the objective function of interest.

We refer to [PCLT17] for other possible generalizations of the model that might be of interest.

**Numerical perspectives.** For the numerics presented in this chapter, we used the modeling language AMPL with the interior-point solver IPOPT. Most of the time, like displayed on Figure 5.4, we were able to perform the continuation with a constant step (on Figure 5.4, two successive values of $\lambda_3 \theta_{HC}$ differ by 0.5). For computational efficiency, one may wish to use a refined procedure. For instance, in the case of convergence, one may try to increase the step in the continuation procedure. On the other hand, when solving the next optimization problem fails, the step can be decreased.

Dealing with this variability of the step could benefit from the use of the solver IPOPT with an efficient programming language, like C or C++. Note that there exist interfaces to use IPOPT designed for the following programming language : C++, C, Fortran, Java, R, Matlab. We refer to the official documentation of the IPOPT project for more details on this issue.

Besides, one could try and use a higher-order method to discretize the dynamics, for instance with Runge-Kutta schemes, and using the trapezoidal rule to discretize the terms $\rho_C$ and $\rho_H$. Again, implementing such a complex numerical method could benefit from the use of one of the previously mentionned programming languages.

## 5.7 Conclusion of the chapter

The objective of the present work was to numerically solve an optimal control generalizing the one studied in the article [PCLT17], in which epimutations were neglected. We have developed an approach which significantly reduces the computation time and improves precision, even without mutations. More precisely, by setting enough parameters to 0 in the original optimal control problem, we arrive to a situation where the problem can be tackled by a Pontryagin Maximum Principle in infinite dimension. Direct methods and continuation then allow to solve the problem of interest, with the strong improvement that we actually start the continuation with a very refined discretization.

We advocate that this approach is suitable for many complicated optimal controls problems. This would be the case as soon as an appropriate simplification leads to a problem for which precise results can be obtained by a PMP. In particular, this approach is an option to be investigated for optimal control problems which have a high-dimensional discretized counterpart.

# Conclusion and perspectives

The main goal of this PhD dissertation was to develop a mathematical framework and conceive a portable software to tackle and solve the problem of optimizing the ballistic phase for an Ariane 5 launcher. A code in C, described in Appendix A, based on a direct method and the use of an interior-point algorithm, was developed for the CNES in order to compute the solution of an optimal control problem with a $L^1$ cost, where the number of body separations and via-point constraints is up to the choice of the user. As a consequence of our process, we also studied in Chapter 3 how to combine continuation techniques and indirect methods to solve a problem with only one intermediate constraint.

Throughout this manuscript, we have made an intensive use of *continuation techniques*. Actually, each time we had to deal with a problem too hard to be addressed directly, we tried to find a deformation of the problem ending with an easier one. Therefore, we exploited the power of continuation techniques in various contexts, depending on the problem at hand. In Chapter 2, we showed how a (now classical) $L^2 \to L^1$ continuation could be used to solve the attitude control problem with minimization of the consumption. In Chapter 3, the heart of our procedure to enforce an intermediate constraint (when using an indirect method) was to penalize the constraint in the cost and do a continuation on the penalization parameter. The intention of Chapter 5 was to carry the expertise gained while applying continuation procedures to aerospace problems to the resolution of an optimal control problem in infinite dimension. We studied an integro-differential system modelling the evolution of cells populations structured by a phenotypical trait, the resistance to chemotherapeutic drugs. Again, the original control problem is highly simplified in order to apply a PMP in infinite dimension, yielding controls with a simple structure over time. From their wide range of applications, we claim that continuation techniques are robust, and can be used quite generically. Besides, the parametric deformation can take a variety of forms including change in the cost, introduction of constraints, increase in the level of discretization, modulation on the set of parameters in the differential system...

Because of flight conditions in real life, it is crucial for the Ariane 5 pilot to be based on a robust control algorithm. Perturbations, model errors can always cause the system to drift away from a planned trajectory. We found the literature on robust algorithms preserving the bang-bang structure of a control to be elusive. In Chapter 4, starting from the intuition that the switching times of a bang-bang command can be considered as degrees of freedom, we suggested an algorithm preserving this structure. Our main idea was to add bang arcs in the form of needle-like variations of the control. In this context, steering the control system to some given target starting form a perturbed point amounts to solving an overdetermined nonlinear shooting problem, what we do by developing a least-square approach. In turn, we design a criterion to measure the quality of robustness of any given bang-bang strategy, based on the singular values of some end-point mapping, and which we optimize.

**Some perspectives.** We shall now finish by giving some perspectives to this work. In Chapter 3 we already mentionned two possible continuations of the work undertaken. A deeper study of why the homotopy on $\varepsilon$ gives far better results that the homotopy on $s$ should be carried. A first step could be to focus on the accessibility set at time $t_1$, in order to diagnose a potential loss of controllability. Besides, the proof of a convergence result for the sequence of adjoint vectors $(p^\varepsilon(\cdot))_{\varepsilon>0}$ is still missing. Such a result would complete the theoretical justification of the procedure of Chapter 3.

Besides, the optimization software presented in Chapter 3 could be subject to many improvements depending on the needs of the CNES. One major axis of developement could be to include in the attitude equations the position and the velocity of the launcher. This would enable the user of the software to take into account constraints ensuring a minimal distance between bodies after a separation, which is of high importance in practice when designing a ballistic phase. This would double the dimension of the system, passing it from $\mathbb{R}^6$ to $\mathbb{R}^{12}$. Therefore, the size of the data in any underlying optimization algorithm would double as well.

As for Chapter 4, the ultimate perspective would be to test our procedure on some real-life system. Recall that in this chapter, we applied our algorithm to the reduced attitude equations in $\mathbb{R}^3$, with only the three angular velocities. Implementing the method on the complete attitude system in $\mathbb{R}^6$, or even on the system with the kinematic variables in $\mathbb{R}^{12}$, would surely be a source of challenges coming from the increased dimension, as well as from the potentially higher number of needle-like variations of the control required to robustify a given trajectory.

On the complete attitude system with position and velocity, let us mention the works [ZTC16a, ZTC16b], where it is shown that optimal trajectories for the time-optimal control problem contain singular arcs, and at the connection between bang arcs and a singular arc, the control chatters: on a compact time interval, the control switches an inifinite number of times. In [ZTC16b], the authors suggest a sub-optimal control strategy, with only a finite numer of switchings. Our approach could be combined with their work in order to design a robust way to place those switchings.

Appendix A

# A software to solve a complete ballistic phase

As announced in the Introduction and earlier in Chapter 3, one of the goals of this thesis was to design an optimization software able to optimize the trajectory of a launcher during any given ballistic flight. Because of the limitations that appeared with the procedure previously presented in Chapter 3 as soon as the number of intermediate constraints becomes greater than one, we took the decision to implement this software with a direct method. It calls an open source library based on an interior-point algorithm. Therefore, we had to implement all the routines for the cost and the constraints, as well as for their derivatives.

**Description of the software.** The details of the software are classified and are the property of the CNES. We shall however give some general elements to explain our approach. The following data has to be provided by the user of the software:

- General elements on the geometry of the launcher, that do not change during the whole ballistic phase, such as the location of the thrusters.
- An integer $\nu$ and the times $t_1, t_2, \ldots, t_\nu$ of the intermediate constraints, $t_\nu$ being also the final time.
- For each $k \in [\![1, \nu]\!]$, the number of constrained components of the state at time $t_k$, $x(t_k)$, and for each constrained component, the value of the constraint.
- For each $k \in [\![1, \nu]\!]$, the values of the inertia coefficients and for the position of the center of mass at times $t_k^+$ and $t_k^-$. As we will emphasize in Remark A.1, when a time $t_k$ corresponds to the separation of a body, the inertia coefficients and the position of the center of mass change. When planning a ballistic phase, the knowledge of the geometry of the launcher and the placement of the satellites allows to know *a priori* those values at any given moment of the future mission.

Note that if $\nu = 1$, the software can be used to solve a simple ballistic phase with only one separation, as we did in Chapter 2.

**Remark A.1: Consequence of a separation**

Some of the times $t_k$ (with $k \in [\![1, \nu]\!]$) can coincide with a rigid body being separated from the launcher. Therefore, the inertia coefficients $I_x$, $I_y$ and $I_z$ and the location of the center of mass are discontinuous at such a time $t_k$. Recall that the expression for the parameters $(a_i)_{1 \leqslant i \leqslant 3}$ is

$$a_1 = \frac{I_y - I_z}{I_x}, a_2 = \frac{I_z - I_x}{I_y}, a_3 = \frac{I_x - I_y}{I_y},$$

and the expression for the vectors $(\vec{b^j})_{1 \leqslant j \leqslant m}$ (corresponding to a force $\overrightarrow{P}$ produced at point $A_j$) is

$$\vec{b^j} = I^{-1} \overrightarrow{P} \wedge \overrightarrow{A_j G}.$$

Therefore, the numerical coefficients $a_1$, $a_2$ and $a_3$, as well as the torques vectors $\vec{b^j}$ (for $j \in [\![1, m]\!]$) are also discontinuous at such a time $t_k$. It follows that the discretization of the dynamics changes after each separation, and a routine computing those coefficients has to be called after each separation.

We give on Figure A.1 a description of the software. After the data is read, an instance of the optimization problem is created, using the routines for the cost (implemented in the file `cost.c`), the constraints (`constraint.c`) and their derivatives (`dcost.c` and `dconstraint.c`). This problem is then solved using an interior-point algorithm, and the output is displayed in the `output.txt` file.

**Numerical output.** We shall now display the output of this software, when used to optimize a ballistic phase with three separations: two satellites are put into orbit. Besides, between the two droppings of the satellites, the dual launch system also has to be separated. Both satellites are required to be separated in a spinned state, rotating at 1.5 degrees per second (0.027 radians per second) about the principal axis of inertia and in both cases, the angle $\varphi$ is left free at the separation. For the separation of the dual launch system, all 6 components of the state are prescribed.

The initial condition on the state of the launcher is

$$(\theta_0, \psi_0, \varphi_0, p_0, q_0, r_0) = (2.6, -0.17, 7.7, 0.01, 0, 0).$$

We denote $t_1$, $t_2$ and $t_3$ the times of the separations of the three bodies. Therefore, the separations of the satellites happen at time $t_1$ and $t_3$, and the separation of the dual launch system happens at time $t_2$. At time $t_1$ and $t_3$, the angle $\varphi$ is left free, and the following constraints are enforced:

$$\begin{aligned}
\theta(t_1) &= \theta(t_3) = 0.04, \\
\psi(t_1) &= \psi(t_3) = 0.06, \\
p(t_1) &= p(t_3) = -0.027, \\
q(t_1) &= q(t_3) = 0, \\
r(t_1) &= r(t_3) = 0.
\end{aligned}$$

We also impose, in order to demonstrate the robustness of our method and as it could be of interest in practice, to control the angular velocities to zero a few seconds before and after each

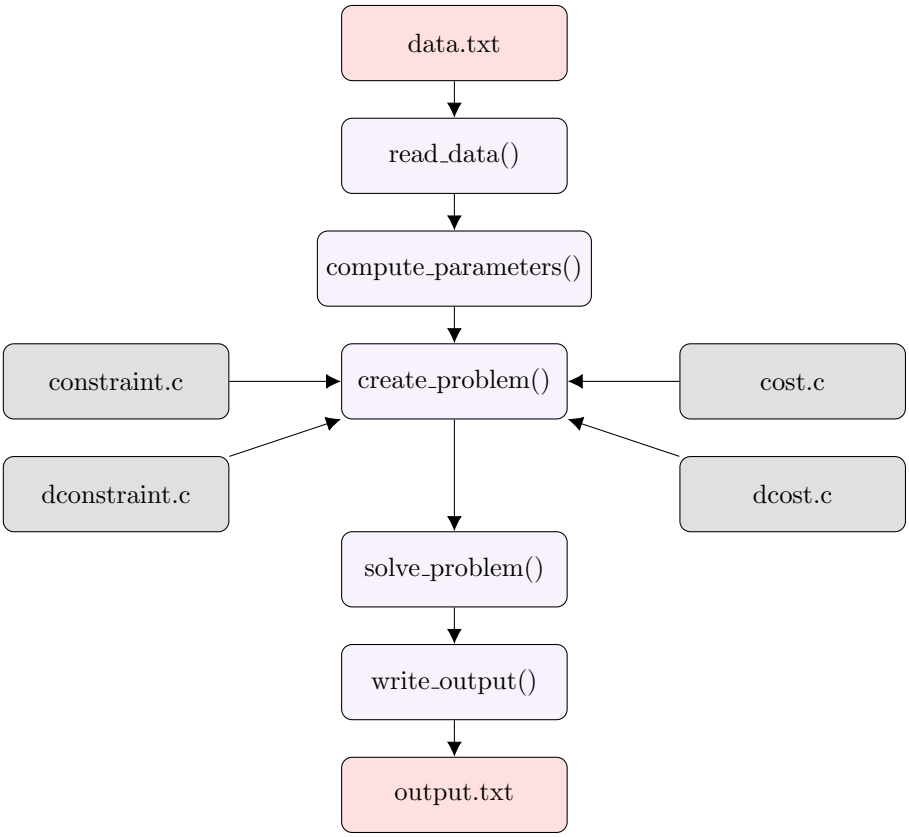



Figure A.1 – Description of the software to optimize a complete ballistic phase. The previously mentionned data that has to be provided by the user is given in the file `data.txt`.

separation. In other words, we chose times $\tau_k$ for $k \in [\![1, 6]\!]$, and enforce the following constraints:

$$p(\tau_k) = q(\tau_k) = r(\tau_k) = 0\,, \quad \forall k \in [\![1, 6]\!]\,.$$

Altogether, there are 9 intermediate constraints in this problem.

On Figure A.3, we plot the trajectories for the 6 components of the state of the launcher. We mark each separation with a blue diamond. Let us emphasize again that for each separation, some components may be left free (here, $\varphi$), imposing to treat the ballistic phase as a whole, and not as three diferent problems.

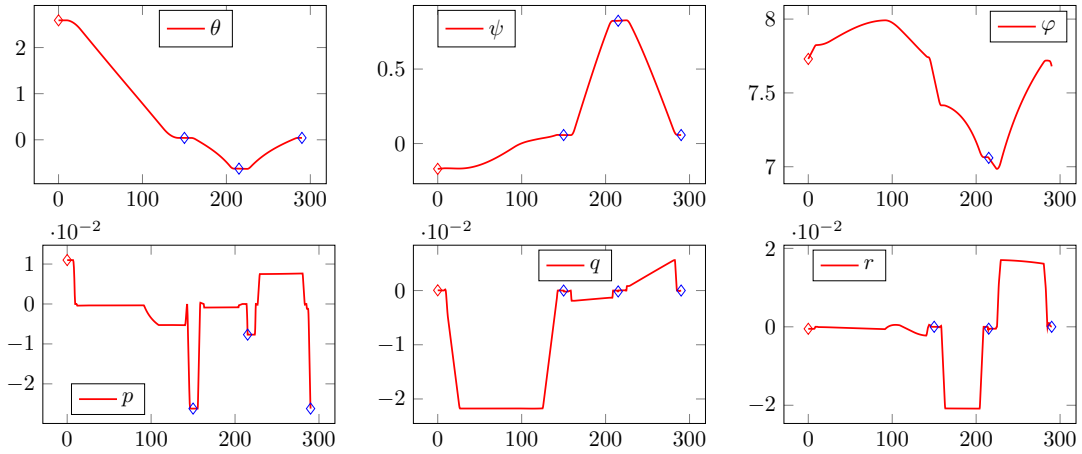



Figure A.2 – Trajectory for the optimization of a whole ballistic phase, starting from ◇. Each ◇ stands for the separation of a body.

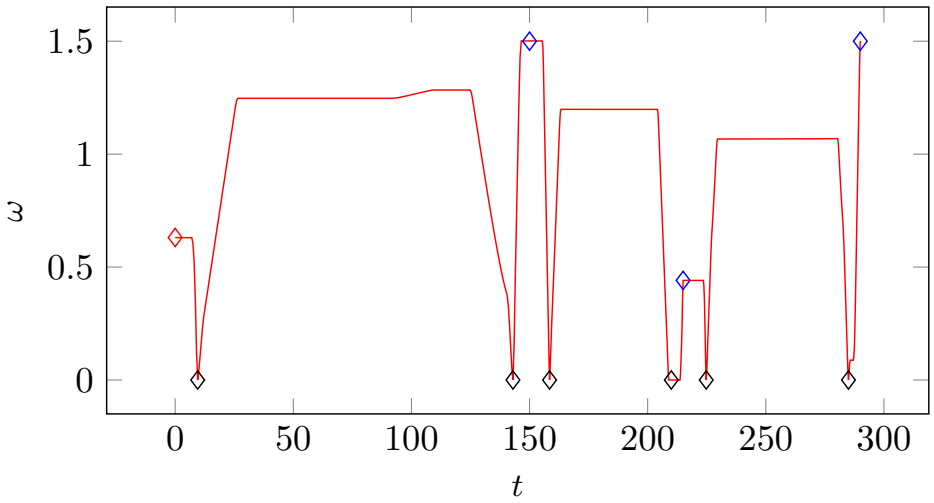



Figure A.3 – Angular velocity for the optimization of a whole ballistic phase, starting from ◇. Each ◇ stands for the separation of a body. Each ◇ corresponds to the control of the angular velocity to 0.

On Figure A.3, we display the angular velocity of the launcher over time. The constraints under the form $\omega(\tau_k) = 0$ ($k \in [\![1, 6]\!]$) are marked with a black diamond. In view of future applications, it is important to mention that once the ballistic phase has been optimized leaving some components free at each intermediate constraint, each part of the mission (for instance between two successive separations) can be used separately.

Finally, we also display on Figure A.4 the corresponding controls. The large number of switching times confirm *a posteriori* the choice to use an direct method, as it is a source of numerical difficulty in the context of indirect methods. We point out that at some point, the controls $u_{11}$ and $u_{12}$ do not reach their maximal value. It is a sign of the lower numerical accuracy of direct methods. Note also the presence of a singular arc on the control $u_8$.

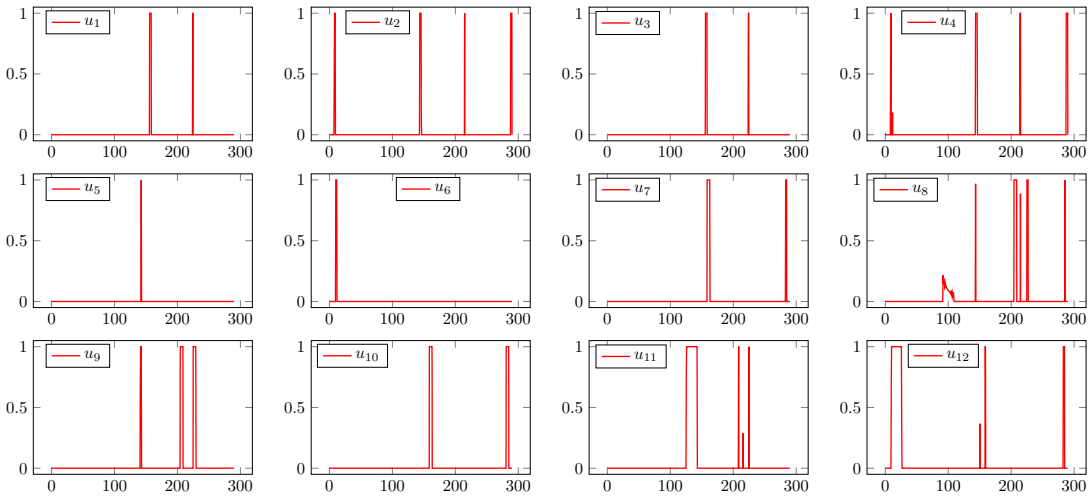



Figure A.4 – Controls for the optimization of a whole ballistic phase.

Appendix B

# Liouville's theorem

In this appendix, we are going to focus on differential systems $\dot{q}(t) = X(q(t))$ whose flow preserves the Lebesgue measure. Liouville's theorem states that a sufficient condition is that the vector field is divergence-free. Such a vector field is also sometimes named a solenoidal vector field. We are first going to give a proof in the linear case, and then in the nonlinear case.

**Linear case.** Let $\dot{v}(t) = A(t)v(t)$ be a linear differential equation. We denote $R(t, t_0)$ the linear mapping such that

$$R(t,t_0) : \left\{ \begin{array}{ccc} \mathbb{R}^n & \longrightarrow & \mathbb{R}^n, \\ v_0 & \mapsto & v(t,t_0,v_0), \end{array} \right.$$

where $v(t, t_0, v_0)$ stands for the solution to the differential equation with initial condition $v(t_0, t_0, v_0) = v_0$.

Let us denote $(v_1(t), \ldots, v_n(t))$ the columns of $R(t, t_0)$, and consider the application

$$T : \left\{ \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R}, \\ t & \mapsto & \det R(t,t_0) = \det(v_1(t), \ldots, v_n(t)). \end{array} \right.$$

The determinant is multilinear with respect to the columns, and we get the following expression for the derivative of $T$:

$$\begin{aligned} T'(t) &= \sum_{j=1}^{n} \det(v_1(t), \ldots, v_j'(t), \ldots, v_n(t)) \\ &= \sum_{j=1}^{n} \det(v_1(t), \ldots, A(t)v_j(t), \ldots, v_n(t)). \end{aligned}$$

The application $(v_1, \ldots, v_n) \mapsto \sum_{j=1}^{n} \det(v_1, \ldots, A(t)v_j, \ldots, v_n)$ also is an alternating multilinear function with respect to the columns. Therefore, it is proportional to the determinant: there exists a constant $K(t)$ such that for all $(v_1, \ldots, v_n)$,

$$\sum_{j=1}^{n} \det(v_1, \ldots, A(t)v_j, \ldots, v_n) = K(t) \det(v_1, \ldots, v_n).$$

Taking in the previous relation $(v_1, \dots, v_n) = (e_1, \dots, e_n)$ where $(e_1, \dots, e_n)$ stands for the canonical basis of $\mathbb{R}^n$, we get that $K(t) = \operatorname{Tr} A(t)$. We have shown that $T(\cdot)$ satisfies the differential equation

$$\begin{cases} T'(t) &= \operatorname{Tr}(A(t))T(t), \\ T(t_0) &= 1, \end{cases}$$

and the expression for the determinant of $R(t, t_0)$ follows

$$\det R(t, t_0) = \exp\left(\int_{t_0}^{t} A(s)ds\right).$$

We can now deduce Liouville's theorem, in the linear case.

**Proposition B.1 (Liouville's theorem - linear case).** – *Assume that for all $t$, $\operatorname{Tr} A(t) = 0$. Then the determinant of the matrix $R(t, t_0, x_0)$ is equal to 1, and the flow preserves the Lebesgue measure.*

**Nonlinear case.**

**Proposition B.2 (Liouville's theorem - nonlinear case).** – *Let $\exp(tX)$ be the flow of a nonlinear differential equation $\dot{y}(t) = X(y(t))$ such that the field $X$ is divergence-free,*

$$\nabla \cdot X(y) = \operatorname{Tr}(dX(y)) = 0.$$

*Then the flow preserves the Lebesgue measure.*

*Proof.* Let $y \in \mathbb{R}^n$. We start by computing the time derivative of the flow $\exp(tX)(y)$. By the very definition of the flow, we get that

$$\frac{d}{dt} \exp(tX)(y) = X(\exp(tX)(y)).$$

Differentiating this equation with respect to $y$, and switching the order of the derivatives, we get that

$$\frac{d}{dt} d\exp(tX)(y) = dX(\exp(tX)(y)) d\exp(tX)(y).$$

Besides, it holds that $d\exp(0X)(y) = Id$. Therefore, we can apply the result in the linear case to the mapping $t \mapsto d\exp(tX)(y)$ and the linearized system

$$\dot{v}(t) = A(t)v(t),$$

where $A(t) = dX(\exp(tX)(y))$.

We get the expression for the determinant of $d\exp(tX)(y)$:

$$\det(d\exp(tX)(y)) = \exp\left(\int_0^t \operatorname{Tr} A(s)ds\right).$$

By hypothesis, $\operatorname{Tr} A(s) = \operatorname{Tr}(dX(\exp(tX)(y))) = \nabla \cdot X(\exp(tX)(y)) = 0$ as the vector field is divergence free. We get that

$$\det(d\exp(tX)(y)) = 1.$$

It follows easily that the flow $\exp(tX)$ preserves the Lebesgue measure: Let $A$ be an open set

of finite measure. Performing a change of variables in the integral yields:

$$\begin{aligned} |\exp(tX)(A)| &= \int_{\exp(tX)(A)} dy \\ &= \int_A |\det(d\exp(tX)(y))|\, dy \\ &= |A|\,. \end{aligned}$$

□

Appendix C

# Linear Algebra

## C.1 Singular value decomposition and pseudoinverse

Let $A \in \mathcal{M}_{n,N}(\mathbb{R})$. The matrix $A^*A$ is hermitian, and its eigenvalues are real and nonnegative. Indeed, let $\lambda \in \mathbb{C}$ be an eigenvalue with an eigenvector $x$ : Then $\|Ax\|^2 = \langle Ax, Ax\rangle = \langle A^*Ax, x\rangle = \langle \lambda x, x\rangle = \lambda \|x\|^2$. The singular values of a matrix $A \in \mathcal{M}_{n,N}(\mathbb{R})$ are the square roots of the (real and nonnegative) eigenvalues of $A^*A$.

This enables us to define the singular value decomposition (SVD) of a matrix :

**Definition C.1.** – *Let $A \in \mathcal{M}_{n,N}(\mathbb{R})$ with $r$ positive singular values. Then, there exist $U \in \mathcal{M}_n(\mathbb{R})$ and $V \in \mathcal{M}_N(\mathbb{R})$, unitary matrices, and $\tilde{\Sigma} \in \mathcal{M}_{n,N}(\mathbb{R})$ a diagonal matrix of the form :*

$$\tilde{\Sigma} = \begin{pmatrix} \Sigma & 0_{n,N-r} \\ 0_{n-r,r} & 0_{n-r,N-r} \end{pmatrix}$$

*with $\Sigma \in \mathcal{M}_r(\mathbb{R})$ (with $r = \mathrm{rank}(A)$) whose diagonal entries are the positive singular values of $A$, such that $A = U\tilde{\Sigma}V^*$*

If $\|\cdot\|_2$ denotes the induced norm for matrices corresponding to the euclidian norm, we easily get from the singular-value definition that $\|A\|_2 = \sigma_{\max}$, where $\sigma_{\max}$ denotes the largest singular value of A. The singular-value decomposition of a matrix enables us to define the pseudoinverse of a matrix.

**Definition C.2.** – *Let $A \in \mathcal{M}_{n,N}(\mathbb{R})$, and $A = U\tilde{\Sigma}V^*$ its SVD decomposition. The pseudo-inverse of $A$ is defined by $A^\dagger = V\tilde{\Sigma}^\dagger U^* \in \mathcal{M}_{N,n}(\mathbb{R})$, with :*

$$\tilde{\Sigma}^\dagger = \begin{pmatrix} \Sigma^{-1} & 0_{r,n-r} \\ 0_{N-r,r} & 0_{N-r,r} \end{pmatrix} \in \mathcal{M}_{N,n}(\mathbb{R})$$

We recall then a few properties of the pseudoinverse, that are needed to understand how to solve a least-squares problem.

**Proposition C.1.** – *(i) $\left\|A^\dagger\right\|_2 = 1/\sigma_{\min}$ where $\sigma_{min}$ is the smallest positive singular value of $A$.*

*(ii)* *The operator* $AA^\dagger$ *is the orthogonal projection on* $\operatorname{range} A$*, and* $I - AA^\dagger$ *is the orthogonal projection on* $(\operatorname{range} A)^\perp$.

*(iii)* *The operator* $A^\dagger A$ *is the orthogonal projection on* $(\ker A)^\perp$*, and* $I - A^\dagger A$ *is the orthogonal projection on* $\ker A$.

*(iv)* $\ker A^\dagger = (\operatorname{range} A)^\perp$.

*(v)* $\operatorname{range} A^\dagger = (\ker A)^\perp$.

*Proof.* A straightforward computation yields

$$AA^\dagger = U \begin{pmatrix} I_r & 0_{r,n-r} \\ 0_{n-r,r} & 0_{n-r,n-r} \end{pmatrix} U^*,$$

which is the expression of an orthogonal projector. Besides, $\operatorname{range}(AA^\dagger) \subset \operatorname{range}(A)$, and from the equality of dimensions, we get the equality between $\operatorname{range}(AA^\dagger)$ and $\operatorname{range}(A)$. Thus, $AA^\dagger$ is the orthogonal projector on $\operatorname{range}(A)$, and it follows that $I - AA^\dagger$ is the orthogonal projector on $\operatorname{range}(A)^\perp$.
The proof for $A^\dagger A$ is similar : we get from a simple computation that it represents an orthogonal projector. Besides, $\ker A \subset \ker(A^\dagger A)$, and they have same dimension $(N - r)$, so $\ker(A^\dagger A) = \ker A$, and we get that $A^\dagger A$ is the orthogonal projector on $(\ker A)^\perp$.
We have that $\ker(A^\dagger) \subset \ker(AA^\dagger)$. But we have just shown that $AA^\dagger$ is the orthogonal projector on $\operatorname{range}(A)$, therefore, $\ker(AA^\dagger) = \operatorname{range}(A)^\perp$. Besides $\dim(\ker(A^\dagger)) = \dim(\operatorname{range}(A)^\perp) = n-r$, so we conclude by equality of the dimensions.
Finally, $(\ker A)^\perp = \operatorname{range}(A^\dagger A) \subset \operatorname{range} A^\dagger$, and we conclude again by equality of the dimensions : $\dim((\ker A)^\perp) = N - \dim(\ker A) = N - (N - r) = r = \operatorname{rank} A^\dagger$. □

## C.2 A least-squares problem

In this section, we consider the following least-squares problem : Given a matrix $A \in \mathcal{M}_{n,N}(\mathbb{R})$ and a vector $b \in \mathbb{R}^n$, find a solution of the optimization problem :

$$\min_{y \in \mathbb{R}^N} \|Ay - b\|$$

The result we want to emphasize here is the link between the pseudo inverse and the solution of the least-squares problem. Note that no assumption is made on $N$ and $n$. That is, the following result holds if the linear system is underdetermined $(n \geqslant N)$, square $(n = N)$ or overdetermined $(n \leqslant N)$.

**Proposition C.2.** – *The vector* $x_b = A^\dagger b$ *is a solution of the least-squares problem. Moreover, in the case where the problem has several solutions, it is the one with minimal norm (for the euclidian norm): let* $x \neq x_b$ *such that* $\|Ax_b - b\|_2 = \|Ax - b\|_2$*, then* $\|x_b\|_2 \leqslant \|x\|_2$

*Proof.* We have $Ax - b = A(x - x_b) + (AA^\dagger - I)b$, with $A(x - x_b) \in \operatorname{range} A$ and $(AA^\dagger - I) \in (\operatorname{range} A)^\perp$ (see proposition C.1). Thus, for all $x \in \mathbb{R}^N$, $\|Ax - b\|^2 = \|A(x - x_b)\|^2 + \|(AA^\dagger - I)b\|^2 = \|A(x - x_b)\|^2 + \|Ax_b - b\|$. And we get that $\|Ax_b - b\|^2 \leqslant \|Ax - b\|^2$, so $x_b$ is a solution of the least-squares problem. If the previous inequality is an equality for some $x$, we get that $\|A(x - x_b)\| = 0$, and thus $A(x - x_b) = 0$, i.e $x - x_b \in \ker A$. Let $z = x - x_b = x - A^\dagger b$. We get that (see proposition C.1)

$$x = \underbrace{z}_{\in \ker A} + \underbrace{A^\dagger b}_{\in (\ker A)^\perp}$$

So, $\|x\|^2 = \|z\|^2 + \left\|A^\dagger b\right\|^2 = \|z\|^2 + \|x_b\|^2$, and $\|x_b\|^2 \leqslant \|x\|^2$. So $x_b$ is indeed the solution of minimal euclidian norm. □

Let $A \in \mathcal{M}_{n,N}(\mathbb{R})$, with $n \leqslant N$, a matrix of maximal rank, i.e $n$ (or, in other words, $A$ is surjective). Thus, $A$ has exactly $n$ positive singular values, which we will denote $\sigma_1 \geqslant \cdots \geqslant \sigma_n > 0$. Given $b \in \mathbb{R}^n$, we know from the surjectivity of $A$ that there is a solution to the equation $Ax = b$. Moreover, according to proposition C.2, we know that $A^\dagger b$ is the solution of minimal norm (for the euclidian norm). And we get the following estimate (see Proposition C.1) :

$$\begin{aligned}\|x_b\|_2 &= \left\|A^\dagger b\right\|_2 \\ &\leqslant \left\|A^\dagger\right\|_2 \cdot \|b\|_2 \\ &\leqslant \frac{1}{\sigma_n} \cdot \|b\|_2\end{aligned}$$

## C.3 Condition number of a matrix

**Case of a square matrix.** In this section, we give some details on the condition number of a matrix. Let $A \in \mathcal{M}_n(\mathbb{R})$. Let $x$ be the solution of some linear system

$$Ax = b.$$

Let $\delta b$ be a perturbation of the right hand term, and $x + \delta x$ be the solution of

$$A(x + \delta x) = b + \delta b,$$

that is, $\delta x$ is a solution of the equation $A\delta x = \delta b$. It follows that

$$\|\delta x\|_2 \leqslant \left\|A^{-1}\right\|_2 \|\delta b\|_2 .$$

If one wishes to have a relative error estimate,

$$\begin{aligned}\frac{\|\delta x\|_2}{\|x\|_2} &\leqslant \left\|A^{-1}\right\|_2 \frac{\|\delta b\|_2}{\|x\|} \\ &\leqslant \left\|A^{-1}\right\|_2 \cdot \|A\|_2 \frac{\|\delta b\|_2}{\|b\|}.\end{aligned}$$

The relative error we make on the solution when having an error on the right-hand term is controlled by the quantity $\left\|A^{-1}\right\|_2 \cdot \|A\|_2$ : this is the condition number of the matrix $A$ with respect to the euclidian norm, denoted by $\mathrm{cond}_2(A)$. Note that this quantity is always greater than 1, as

$$\begin{aligned}\left\|A^{-1}\right\|_2 \cdot \|A\|_2 &\geqslant \left\|A^{-1}A\right\|_2 \\ &\geqslant \|I_n\|_2 \\ &= 1\end{aligned}$$

The condition number of a matrix has a very elegant geometric interpretation : from the SVD decomposition, it can be derived that the image of the unit sphere in $\mathbb{R}^n$ by an inversible matrix is an ellipsoid, whose semi-major axis is $\sigma_{\max}$ and whose semi-minor axis is $\sigma_{\min}$, as displayed

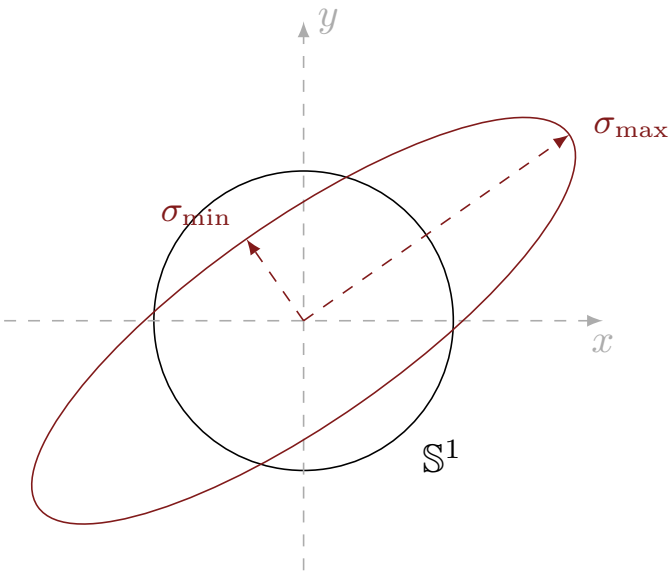



Figure C.1 – In dimension 2, image by an inverible matrix of the unit sphere $\mathbb{S}^1$.

on Figure C.1 in dimension 2. Thus, the condition number of a matrix measures how close the ellipsoid is to a sphere.

**Case of a non square matrix.** In Chapter 4, we give an algorithm to control a system even in presence of perturbations, based on the computation of the solution of a linear system

$$dE \cdot \delta\mathcal{T} = \delta x,$$

where $dE$ is a matrix of dimension $n \times N$, with $N$ possibly larger than $n$. If the matrix $dE$ is of full rank $n$, with the singular values $\sigma_1 \geqslant \cdots \geqslant \sigma_n > 0$, then the image of the unit sphere of $\mathbb{R}^N$ by the matrix $dE$ is, up to a rotation,

$$dE(\mathbb{S}^{N-1}) = \left\{ x = (x_1, \ldots, x_n) \mid \left(\frac{x_1}{\sigma_1}\right)^2 + \ldots + \left(\frac{x_n}{\sigma_n}\right)^2 \right\}.$$

# Bibliographie


[ADDJ91] C. Abdallah, D. M. Dawson, P. Dorato, and M. Jamshidi. Survey of robust control for rigid robots. *IEEE Control Systems*, 11(2) :24–30, 1991.

[AE14] Usman Ali and Magnus Egerstedt. Optimal control of switched dynamical systems under dwell time constraints. In *53rd IEEE Conference on Decision and Control, CDC 2014, Los Angeles, CA, USA, December 15-17, 2014*, pages 4673–4678, 2014.

[AG90] Eugene L. Allgower and Kurt Georg. *Numerical Continuation Methods : An Introduction*. Springer-Verlag New York, Inc., New York, NY, USA, 1990.

[AK02] G. Allaire and S.M. Kaber. *Algèbre linéaire numérique*. Mathématiques pour le 2e cycle. Ellipses, 2002.

[AM71] Brian D. O. Anderson and John B. Moore. *Linear optimal control*. Prentice-Hall Inc., Englewood Cliffs N.J., 1971.

[AN06] P. Apkarian and D. Noll. Nonsmooth $h_\infty$ synthesis. *IEEE Transactions on Automatic Control*, 51(1) :71–86, 2006.

[ANTT04] Pierre Apkarian, Dominikus Noll, Jean-Baptiste Thevenet, and Hoang Duong Tuan. A Spectral Quadratic-SDP Method with Applications to Fixed-Order $H_2$ and $H_\infty$ Synthesis. *European Journal of Control*, 10(6) :527–538, 2004.

[Ari16] Arianespace. *Ariane 5 User's Manual*, Oct. 2016.

[AS04] A.A. Agrachev and Y. Sachkov. *Control Theory from the Geometric Viewpoint*. Control theory and optimization. Springer, 2004.

[BBM98] M. S. Branicky, V. S. 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, Jan 1998.

[BC03a] B. Bonnard and M. Chyba. *Singular Trajectories and their Role in Control Theory*. Mathématiques et Applications. Springer Berlin Heidelberg, 2003.

[BC03b] Bernard Bonnard and Monique Chyba. *Singular trajectories and their role in control theory*, volume 40 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 2003.

[BCL15] Olivier Bonnefon, Jérôme Coville, and Guillaume Legendre. Concentration phenomenon in some non-local equation. *Preprint arXiv :1510.01971*, 2015.

[BD05] Sorin C. Bengea and Raymond A. DeCarlo. Optimal control of switching systems. *Automatica*, 41(1) :11–27, 2005.

[Beu65a] Frederick J. Beutler. The operator theory of the pseudo-inverse. I. Bounded operators. *J. Math. Anal. Appl.*, 10 :451–470, 1965.

[Beu65b] Frederick J. Beutler. The operator theory of the pseudo-inverse. II. Unbounded operators with arbitrary range. *J. Math. Anal. Appl.*, 10 :471–493, 1965.

[BFLT03] 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, Apr 2003.

[BFT05] B. Bonnard, L. Faubourg, and E. Trélat. Optimal control of the atmospheric arc of a space shuttle and numerical simulations with multiple-shooting method. *Math. Models Methods Appl. Sci.*, 15(1) :109–140, 2005.

[BFT06] B. Bonnard, L. Faubourg, and E. Trélat. *Mécanique céleste et contrôle des véhicules spatiaux*. Mathématiques et Applications. Springer Berlin Heidelberg, 2006.

[BGFB94] S. Boyd, L.E. Ghaoui, E. Feron, and V. Balakrishnan. *Linear Matrix Inequalities in System and Control Theory*. Studies in Applied Mathematics. Society for Industrial and Applied Mathematics, 1994.

[BH75] A.E. Bryson and Y.C. Ho. *Applied Optimal Control : Optimization, Estimation and Control*. Halsted Press book'. Taylor & Francis, 1975.

[BNPvS93] Dr Roland Bulirsch, Dipl Math Edda Nerz, Priv-Doz Dr Hans Josef Pesch, and Dipl Math Oskar von Stryk. Combining direct and indirect methods in optimal control : Range maximization of a hang glider. In *Optimal control*, pages 273–288. Springer, 1993.

[Car17] Cécile Carrère. Optimization of an in vitro chemotherapy to avoid resistant tumours. *Journal of Theoretical Biology*, 413 :24–33, Jan 2017.

[CBB92] MIS Costa, JL Boldrini, and RC Bassanezi. Optimal chemical control of populations developing drug resistance. *Mathematical Medicine and Biology*, 9(3) :215–226, 1992.

[CCG12] J.-B. Caillau, O. Cots, and J. Gergaud. Differential continuation for regular optimal control problems. *Optimization Methods and Software*, 27(2) :177–196, 2012.

[CDG12] J.-B. Caillau, B. Daoud, and J. Gergaud. Minimum fuel control of the planar circular restricted three-body problem. *Celestial Mechanics and Dynamical Astronomy*, 114(1) :137–150, Oct 2012.

[Cea11] Nalin A. Chaturvedi and et al. Rigid-body attitude control – using rotation matrices for continuous, singularity-free control laws, 2011.

[CFPT13] Marco Caponigro, Massimo Fornasier, Benedetto Piccoli, and Emmanuel Trélat. Sparse stabilization and optimal control of the Cucker-Smale model. *Math. Control Relat. Fields*, 3(4) :447–466, 2013.

[CGN03] J.B. Caillau, J. Gergaud, and J. Noailles. 3D Geosynchronous Transfer of a Satellite : Continuation on the Thrust. *Journal of Optimization Theory and Applications*, 118(3) :541–565, Sep 2003.

[CHS+09] M. Chyba, T. Haberkorn, S.B. Singh, R.N. Smith, and S.K. Choi. Increasing underwater vehicle autonomy by reducing energy consumption. *Ocean Engineering*, 36(1) :62–73, 2009. Autonomous Underwater Vehicles.

[CHSC08] M. Chyba, T. Haberkorn, R.N. Smith, and S.K. Choi. Design and implementation of time efficient trajectories for autonomous underwater vehicles. *Ocean Engineering*, 35(1) :63–76, 2008.

[CHT12] Max Cerf, Thomas Haberkorn, and Emmanuel Trélat. Continuation from a flat to a round earth model in the coplanar orbit transfer problem. *Optimal Control Applications and Methods*, 33(6) :654–675, 2012.

[CHT17] Maxime Chupin, Thomas Haberkorn, and Emmanuel Trélat. Low-Thrust Lyapunov to Lyapunov and Halo to Halo with $L^2$-Minimization . *ESAIM : Mathematical Modelling and Numerical Analysis*, 51(3) :965–996, 2017.

[Cla90] F.H. Clarke. *Optimization and Nonsmooth Analysis*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1990.

[CLC16] Rebecca H. Chisholm, Tommaso Lorenzi, and Jean Clairambault. Cell population heterogeneity and evolution towards drug resistance in cancer : Biological and mathematical assessment, theoretical treatment optimisation. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1860(11) :2627–2645, Nov 2016.

[CLL16] Rebecca H Chisholm, Tommaso Lorenzi, and Alexander Lorz. Effects of an advection term in nonlocal lotka–volterra equations. *Communications in Mathematical Sciences*, 14(4) :1181–1188, 2016.

[Cor92] Jean Michel Coron. Global asymptotic stabilization for controllable systems without drift. *Mathematics of Control, Signals and Systems*, 5(3) :295–312, Sep 1992.

[Cov13] Jerome Coville. Convergence to equilibrium for positive solutions of some mutation-selection model. *Preprint arXiv :1308.6471*, 2013.

[CR13] Xavier Cabré and Jean-Michel Roquejoffre. The influence of fractional diffusion in fisher-kpp equations. *Communications in Mathematical Physics*, 320(3) :679–722, 2013.

[D+04] Odo Diekmann et al. A beginner's guide to adaptive dynamics. *Banach Center Publications*, 63 :47–86, 2004.

[D.10] Curtis Howard D. *Orbital mechanics for engineering students*, volume Elsevier aerospace engineering series. Butterworth-Heinemann, 2010.

[DGKF89] J. C. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis. State-space solutions to standard $h_2$ and $h_\infty$ control problems. *IEEE Transactions on Automatic Control*, 34(8) :831–847, 1989.

[DJMP05] Odo Diekmann, Pierre-Emanuel Jabin, Stéphane Mischler, and Benoît Perthame. The dynamics of adaptation : an illuminating example and a Hamilton-Jacobi approach. *Theoretical Population Biology*, 67(4) :257–271, 2005.

[DK08] 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.

[DK11] A. V. Dmitruk and A. M. Kaganovich. Maximum principle for optimal control problems with intermediate constraints. *Computational Mathematics and Modeling*, 22(2) :180–215, Apr 2011.

[dNDL13] B. d'Andréa Novel and M. De Lara. *Control Theory for Engineers : A Primer*. Environmental Science and Engineering / Environmental Engineering. Springer Berlin Heidelberg, 2013.

[DS81] J. Doyle and G. Stein. Multivariable feedback design : Concepts for a classical/modern synthesis. *IEEE Transactions on Automatic Control*, 26(1) :4–16, 1981.

[FGK93] R. Fourer, D.M. Gay, and B.W. Kernighan. *AMPL : A Modeling Language for Mathematical Programming*. Scientific Press, 1993.

[FGK02] Robert Fourer, David M Gay, and Brian W Kernighan. A modeling language for mathematical programming. *Duxbury Press*, 36(5) :519–554, 2002.

[Ful63] A. T. Fuller. Study of an Optimum Non-linear Control System. *Journal of Electronics and Control*, 15(1) :63–71, 1963.

[GA94] Pascal Gahinet and Pierre Apkarian. A linear matrix inequality approach to $h_\infty$ control. *International Journal of Robust and Nonlinear Control*, 4(4) :421–448, 1994.

[Gah92] P. Gahinet. A convex parametrization of $h_\infty$ suboptimal controllers. In *[1992] Proceedings of the 31st IEEE Conference on Decision and Control*, pages 937–942 vol.1, 1992.

[GFL96] Jian-Hua Ge, P.M. Frank, and Ching-Fang Lin. Robust $H_\infty$ state feedback control for linear systems with state delay and parameter uncertainty. *Automatica*, 32(8) :1183–1185, 1996.

[GH06] Gergaud, Joseph and Haberkorn, Thomas. Homotopy method for minimum consumption orbit transfer problem. *ESAIM : COCV*, 12(2) :294–310, 2006.

[GLGL14] James Greene, Orit Lavi, Michael M Gottesman, and Doron Levy. The impact of cell density and mutations in a model of multidrug resistance in solid tumors. *Bulletin of mathematical biology*, 76(3) :627–653, 2014.

[GP05a] Mauro Garavello and Benedetto Piccoli. Hybrid necessary principle. 43 :1867–1887, 01 2005.

[GP05b] Mauro Garavello and Benedetto Piccoli. Hybrid Necessary Principle. *SIAM Journal on Control and Optimization*, 43(5) :1867–1887, 2005.

[GVL13] Gene H. Golub and Charles F. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013.

[HMG04] Haberkorn T., Martinon P., and Gergaud J. Low Thrust Minimum-Fuel Orbital Transfer : A Homotopic Approach. *Journal of Guidance, Control, and Dynamics*, 27(6) :1046–1060, 2004. doi : 10.2514/1.4022.

[HNW08] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I : Nonstiff Problems*. Springer Series in Computational Mathematics. Springer Berlin Heidelberg, 2008.

[KC99] K. C. Koh and H. S. Cho. A smooth path tracking algorithm for wheeled mobile robots with dynamic constraints. *J. Intell. Robotics Syst.*, 24(4) :367–385, 1999.

[Kha92] Hassan K. Khalil. *Nonlinear systems*. Macmillan Publishing Company, New York, 1992.

[KS72] Huibert Kwakernaak and Raphael Sivan. *Linear optimal control systems*. Wiley-Interscience [John Wiley & Sons], New York-London-Sydney, 1972.

[KS89] Arthur J. Krener and Heinz Schättler. The structure of small-time reachable sets in low dimensions. *SIAM J. Control Optim.*, 27(1) :120–147, 1989.

[KS06] Marek Kimmel and Andrzej Świerniak. Control theory approach to cancer chemotherapy : Benefiting from phase dependence and overcoming drug resistance. In Avner Friedman, editor, *Tutorials in Mathematical Biosciences III*, volume 1872 of *Lecture Notes in Mathematics*, pages 185–221. Springer Berlin / Heidelberg, 2006.

[KT99] M. Krstic and P. Tsiotras. Inverse optimal stabilization of a rigid spacecraft. *IEEE Transactions on Automatic Control*, 44(5) :1042–1049, 1999.

[Kup87] I. Kupka. Geometric theory of extremals in optimal control problems. I. The fold and Maxwell case. *Trans. Amer. Math. Soc.*, 299(1) :225–243, 1987.

[LCDH15] Tommaso Lorenzi, Rebecca H Chisholm, Laurent Desvillettes, and Barry D Hughes. Dissecting the dynamics of epigenetic changes in phenotype-structured populations exposed to fluctuating environments. *Journal of theoretical biology*, 386 :166–176, 2015.

[Lin07] F. Lin. *Robust Control Design : An Optimal Control Approach*. RSP. Wiley, 2007.

[LLC+13] Alexander Lorz, Tommaso Lorenzi, Jean Clairambault, Alexandre Escargueil, and Benoît Perthame. Effects of space structure and combination therapies on phenotypic heterogeneity and drug resistance in solid tumors. *arXiv preprint arXiv :1312.6237*, 2013.

[LLC+15] Alexander Lorz, Tommaso Lorenzi, Jean Clairambault, Alexandre Escargueil, and Benoît Perthame. Modeling the effects of space structure and combination therapies on phenotypic heterogeneity and drug resistance in solid tumors. *Bulletin of mathematical biology*, 77(1) :1–22, 2015.

[LLH+13] Alexander Lorz, Tommaso Lorenzi, Michael E Hochberg, Jean Clairambault, and Benoît Perthame. Populational adaptive evolution, chemotherapeutic resistance and multiple anti-cancer therapies. *ESAIM : Mathematical Modelling and Numerical Analysis*, 47(02) :377–399, 2013.

[LLY95] H. Li, X. Li, and J. Yong. *Optimal Control Theory for Infinite Dimensional Systems*, chapter 4 - Necessary conditions for optimal control. Systems & control : foundations & applications. Birkhäuser, 1995.

[LM67a] E.B. Lee and L. Markus. *Foundations of optimal control theory*, chapter 4 - The maximal principle and the existence of optimal controllers for nonlinear processes. SIAM series in applied mathematics. Wiley, 1967.

[LM67b] E.B. Lee and L. Markus. *Foundations of optimal control theory.* SIAM series in applied mathematics. Wiley, 1967.

[LMM14] Helene Leman, Sylvie Meleard, and Sepideh Mirrahimi. Influence of a spatial structure on the long time behavior of a competitive lotka-volterra type system. 20, 01 2014.

[LS06] Urszula Ledzewicz and Heinz Schättler. Drug resistance in cancer chemotherapy as an optimal control problem. *Discrete and Continuous Dynamical Systems Series B*, 6(1) :129, 2006.

[LS14] Urszula Ledzewicz and Heinz Schättler. On optimal chemotherapy for heterogeneous tumors. *Journal of Biological Systems*, 22(02) :177–197, 2014.

[LW92] Liu Qiang and Wie Bong. Robust time-optimal control of uncertain flexible spacecraft. *Journal of Guidance, Control, and Dynamics*, 15(3) :597–604, 1992. doi : 10.2514/3.20880.

[LY12] Xungjing Li and Jiongmin Yong. *Optimal control theory for infinite dimensional systems.* Springer Science & Business Media, 2012.

[MBKK05] H. Maurer, C. Büskens, J.-H. R. Kim, and C. Y. Kaya. Optimization methods for the verification of second order sufficient conditions for bang–bang controls. *Optimal Control Applications and Methods*, 26(3) :129–156, 2005.

[MG92] D. McFarlane and K. Glover. A loop-shaping design procedure using h infinity synthesis. *IEEE Transactions on Automatic Control*, 37(6) :759–769, 1992.

[MO04] Helmut Maurer and Nikolai P. Osmolovskii. Second Order Sufficient Conditions for Time-Optimal Bang-Bang Control. *SIAM Journal on Control and Optimization*, 42(6) :2239–2263, 2004.

[OS92] R. Outbib and G. Sallet. Stabilizability of the angular velocity of a rigid body revisited. *Systems & Control Letters*, 18(2) :93–98, 1992.

[PBGM62] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko. *The mathematical theory of optimal processes.* Translated from the Russian by K. N.

Trirogoff ; edited by L. W. Neustadt. Interscience Publishers John Wiley & Sons, Inc. New York-London, 1962.

[PCLT17] Camille Pouchol, Jean Clairambault, Alexander Lorz, and Emmanuel Trélat. Asymptotic analysis and optimal control of an integro-differential system modelling healthy and cancer cells exposed to chemotherapy. *Journal de Mathématiques Pures et Appliquées*, 2017.

[Per06] Benoît Perthame. *Transport equations in biology.* Springer Science & Business Media, 2006.

[Pes94] Hans Josef Pesch. A practical guide to the solution of real-life optimal control problems. *Control and cybernetics*, 23(1) :2, 1994.

[Pic99] B. Piccoli. Necessary conditions for hybrid optimization. In *Proceedings of the 38th IEEE Conference on Decision and Control (Cat. No.99CH36304)*, volume 1, pages 410–415 vol.1, 1999.

[Poi90] H. Poincaré. *Sur le probleme des trois corps et les équations de la dynamique.* Acta Mathematica. F. & G. Beijer, 1890.

[PT17] Camille Pouchol and Emmanuel Trélat. Global stability with selection in integro-differential lotka-volterra systems modelling trait-structured populations. *arXiv preprint arXiv :1702.06187*, 2017.

[SC05] M. S. Shaikh and P. E. Caines. Optimality zone algorithms for hybrid systems computation and control : From exponential to linear complexity. In *Proceedings of the 44th IEEE Conference on Decision and Control*, pages 1403–1408, 2005.

[SC07] 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, Sept 2007.

[Sch88] Heinz Schättler. On the local structure of time-optimal bang-bang trajectories in $\mathbf{R}^3$. *SIAM J. Control Optim.*, 26(1) :186–204, 1988.

[SJ72] Héctor J Sussmann and Velimir Jurdjevic. Controllability of nonlinear systems. *Journal of Differential Equations*, 12(1) :95–116, 1972.

[SL15] Heinz Schättler and Urszula Ledzewicz. *Optimal Control for Mathematical Models of Cancer Therapies.* Springer New York, 2015.

[ST10a] C. J. Silva and Emmanuel Trélat. Smooth Regularization of Bang-Bang Optimal Control Problems. *IEEE Trans. Automat. Contr.*, 55(11) :2488–2499, 2010.

[ST10b] Cristiana Silva and Emmanuel Trélat. Smooth regularization of bang-bang optimal control problems. *IEEE Trans. Automat. Control*, 55(11) :2488–2499, 2010.

[Sus99] Héctor J. Sussmann. *A nonsmooth hybrid maximum principle*, pages 325–354. Springer London, London, 1999.

[Sus00] H. J. Sussmann. Set-valued differentials and the hybrid maximum principle. In *Proceedings of the 39th IEEE Conference on Decision and Control (Cat. No.00CH37187)*, volume 1, pages 558–563 vol.1, 2000.

[SV94] T Singh and S.R. Vadali. Robust time-optimal control - Frequency domain approach. *Journal of Guidance, Control, and Dynamics*, 17(2) :346–353, 1994. doi : 10.2514/3.21204.

[Tré00] E. Trélat. Some Properties of the Value Function and Its Level Sets for Affine Control Systems with Quadratic Cost. *Journal of Dynamical and Control Systems*, 6(4) :511–541, Oct 2000.

[Tré05a] Emmanuel Trélat. *Contrôle optimal*, chapter 6 - Contrôle Optimal. Mathématiques Concrètes. [Concrete Mathematics]. Vuibert, Paris, 2005. Théorie & applications. [Theory and applications].

[Tré05b] Emmanuel Trélat. *Contrôle optimal*. Mathématiques Concrètes. [Concrete Mathematics]. Vuibert, Paris, 2005. Théorie & applications. [Theory and applications].

[Tré12] E. Trélat. Optimal control and applications to aerospace : Some results and challenges. *Journal of Optimization Theory and Applications*, 154(3) :713–758, 2012.

[TSL09] Haihua Tan, Shaolong Shu, and Feng Lin. An optimal control approach to robust tracking of linear systems. *International Journal of Control*, 82(3) :525–540, 2009.

[vSB92] O. von Stryk and R. Bulirsch. Direct and indirect methods for trajectory optimization. *Annals of Operations Research*, 37(1) :357–373, Dec 1992.

[War12] Yorai Wardi. Optimal control of switched-mode dynamical systems. *{IFAC} Proceedings Volumes*, 45(29) :4 – 8, 2012. 11th {IFAC} Workshop on Discrete Event Systems.

[WB06a] Andreas Wächter and Lorenz T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1) :25–57, Mar 2006.

[WB06b] Andreas Wächter and Lorenz T Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical programming*, 106(1) :25–57, 2006.

[Win63] T.G Windeknecht. Optimal stabilization of rigid body attitude. *Journal of Mathematical Analysis and Applications*, 6(2) :325–335, 1963.

[WSL93] Wie Bong, Sinha Ravi, and Liu Qiang. Robust time-optimal control of uncertain structural dynamic systems. *Journal of Guidance, Control, and Dynamics*, 16(5) :980–983, 1993. doi : 10.2514/3.21114.

[XA00] Xuping Xu and P. J. Antsaklis. Optimal control of switched systems : new results and open problems. In *Proceedings of the 2000 American Control Conference. ACC (IEEE Cat. No.00CH36334)*, volume 4, pages 2683–2687 vol.4, 2000.

[XdSC92] L. Xie and E. de Souza Carlos. Robust h infin; control for linear systems with norm-bounded time-varying uncertainty. *IEEE Transactions on Automatic Control*, 37(8) :1188–1191, 1992.

[XSCZ06] Shengyuan Xu, Peng Shi, Yuming Chu, and Yun Zou. Robust stochastic stabilization and control of uncertain neutral stochastic time-delay systems. *Journal of Mathematical Analysis and Applications*, 314(1) :1–16, 2006.

[YL00] K. H. You and E. B. Lee. Robust, near time-optimal control of nonlinear second order systems with model uncertainty. In *Proceedings of the 2000. IEEE International Conference on Control Applications. Conference Proceedings (Cat. No.00CH37162)*, pages 232–236, 2000.

[ZA15] Feng Zhu and Panos J. Antsaklis. Optimal control of hybrid switched systems : A brief survey. *Discrete Event Dynamic Systems*, 25(3) :345–364, 2015.

[Zam81] G. Zames. Feedback and optimal sensitivity : Model reference transformations, multiplicative seminorms, and approximate inverses. *IEEE Transactions on Automatic Control*, 26(2) :301–320, 1981.

[ZDG96] K. Zhou, J.C. Doyle, and K. Glover. *Robust and Optimal Control*. Feher/Prentice Hall Digital an. Prentice Hall, 1996.

[ZTC16a] Jiamin Zhu, Emmanuel Trélat, and Max Cerf. Minimum time control of the rocket attitude reorientation associated with orbit dynamics. *SIAM J. Control Optim.*, 54(1) :391–422, 2016.

[ZTC16b] Jiamin Zhu, Emmanuel Trélat, and Max Cerf. Planar tilting maneuver of a spacecraft : singular arcs in the minimum time problem and chattering. *Discrete and Continuous Dynamical Systems - Series B*, 16(4) :1347–1388, 2016.

## Contrôle optimal et robuste de l'attitude d'un lanceur
**Aspects Théoriques et Numériques**


### Résumé

L'objectif premier de cette thèse est d'étudier certains aspects du contrôle d'attitude d'un corps rigide, afin d'optimiser la trajectoire d'un lanceur au cours de sa phase balistique. Nous y développons un cadre mathématique permettant de formuler ce problème comme un problème de contrôle optimal avec des contraintes intermédiaires sur l'état. En parallèle de l'étude théorique de ce problème, nous avons mené l'implémentation d'un logiciel d'optimisation basé sur la combinaison d'une méthode directe et d'un algorithme de point intérieur, permettant à l'utilisateur de traiter une phase balistique quelconque. Nous entendons par là qu'il est possible de spécifier un nombre quelconque de contraintes intermédiaires, correspondant à un nombre quelconque de largages de charges utiles.
En outre, nous avons appliqué les méthodes dites indirectes, exploitant le principe du maximum de Pontryagin, à la résolution de ce problème de contrôle optimal. On cherche dans ce travail à trouver des trajectoires optimales du point de vue de la consommation en ergols, ce qui correspond à un coût $L^1$. Réputé difficile numériquement, ce critère peut être atteint grâce à une méthode de continuation, en se servant d'un coût $L^2$ comme intermédiaire de calcul et en déformant progressivement ce problème $L^2$. Nous verrons également d'autres exemples d'application des méthodes de continuation.
Enfin, nous présenterons également un algorithme de contrôle robuste, permettant de rejoindre un état cible à partir d'un état perturbé, en suivant une trajectoire de référence tout en conservant la structure bang-bang des contrôles. La robustesse d'un contrôle peut également être améliorée par l'ajout de variations aiguilles, et un critère qualifiant la robustesse d'une trajectoire à partir des valeurs singulières d'une certaine application entrée-sortie est déduit.



**Mots clés :** contrôle optimal, contrôle d'attitude, phase balistique, méthode de continuation, méthodes directes, méthodes indirectes, contrôle robuste, contrainte intermédiaire


---

## Optimal and robust attitude control of a launcher
**Theoretical and numerical aspects**


### Abstract

The first objective of this work is to study some aspects of the attitude control problem of a rigid body, in order to optimize the trajectory of a launcher during a ballistic flight. We state this problem in a general mathematical setting, as an optimal control problem with intermediate constraints on the state. Meanwhile, we also implement an optimization software that relies on the combination of a direct method and of an interior-point algorithm to optimize any given ballistic flight, with any number of intermediate constraints, corresponding to any number of satellite separations.
Besides, we applied the so-called indirect methods, exploiting Pontryagin maximum principle, to the resolution of this optimal control problem. In this work, optimal trajectories with respect to the consumption are looked after, which corresponds to a $L^1$ cost. Known to be numerically challenging, this criterion can be reached by performing a continuation procedure, starting from a $L^2$ cost, for which it is easier to provide a good initialization of the underlying optimization algorithm. We shall also study other examples of applications for continuation procedures.
Eventually, we will present a robust control algorithm, allowing to reach a target point from a perturbed initial point, following a nominal trajectory while preserving its bang-bang structure. The robustness of a control can be improved introducing needle-like variations, and a criterion to measure the robustness of a trajectory is designed, involving the singular value decomposition of some end-point mapping.



**Keywords:** optimal control, attitude control, ballistic phase, continuation method, direct methods, indirect methods, robust control, intermediate constraint


---


**Laboratoire Jacques Louis Lions**
4 Place Jussieu – 75005, PARIS – France