AIMS on Applied Mathematics Vol. 12

IVAN KUPKA LEGACY
A Tour Through Controlled Dynamics

Bernard Bonnard
Monique Chyba
David Holcman
Emmanuel Trélat (Eds.)

American Institute of Mathematical Sciences

AMS 2020 Classifications: 93-xx Systems theory; control, 49-xx Calculus of variations and optimal control; optimization, 37-xx Dynamical systems and ergodic theory

ISBN-10: 1-60133-026-X; ISBN-13: 978-1-60133-026-0

aimsciences.org

Ivan Kupka (1937-2023)

Foreword

This volume is a tribute to Ivan Kupka who passed away at Easter 2023. It contains a collection of articles written by colleagues working on dynamical systems or control theory. Those colleagues either came across Ivan Kupka directly or indirectly through his articles. Personal remembrances as well as a description of his work can be found in a volume of the Gazette de Mathématiques that will appear in 2024 illustrating the journey of this exceptional mathematician that was as perceptive in his work than in his caring relations with other individuals. He leaves us all with the memory of a hard working and cultivated scientist (an erudite), gifted of an exceptional curiosity and fond of freedom. For those who had the chance to enter his circle he was also an loyal friend and always available.

Below is a description of some of his scientific contributions in control representing collaborative work.

- **End of 70s: Sufficient conditions for controllability for triplets of vector fields $\{A, \pm B\}$.** In collaboration with V. Jurdjevic. In this work the authors introduce in a vert throughout way allowed operations on a triple to expand it while preserving its controllability (constructing the Lie saturate). The most important technical point being the use of semi-simple Lie agebra classification to deduce sufficient conditions for controllability for invariants systems on semi-simple Lie groups (the triple being an invariant field). This work belongs to classical mathematics: transitive action of a semi group. It was then used for the case of semi-direct products.
- **From the 80s: Second-order optimality conditions for singular trajectories and generic properties.** In collaboration with B. Bonnard during a decade especially to study the role of singular trajectories in optimal control. In particular second-order optimality conditions were derived for singular trajectories of affine single-input systems. These conditions were implemented numerically using an algorithm differing from the initial calculation which uses a semi-normal form to evaluate the intrinsic second-order derivative of the end-point map and a (Fourier) suited representation of the control. In parallel a lot of time and wrong proofs were drafted before two generic properties for singular trajectories of affine single-input systems: minimal order and strict abnormality.
  Sub-Riemannian Geometry. Work in SR-geometry on the so-called Martinet case to study the role of abnormal trajectories and computation of spheres of small radius. This was done in the context of two PhD dissertations, the one of M. Chyba and then of E. Trélat. This work has its

origin in a paper never published on the use of elliptic functions for explicit calculations on a model stable enough inspired by the work of R. Montgomery.

- **Mid 80s: Classification of bang-bang extremals and Fuller phenomenon.** This work contains two important contributions. The first one is formal calculation of normal forms for discontinuous Hamiltonian systems and classification of generic bang-bang extremals. The second contribution rising from this activity which is extremely technical was the proof by Ivan Kupka of the ubiquity of the Fuller phenomenon. B. Bonnard: "I remember listening for a long time about his subsequent trials of proofs and the complete proof was only published as a confidential paper [1]".
- **End of 90s and beginning of 21st century.** The technical expertise of Ivan Kupka as well as his relentless work shines in the master class he taught for many years in Toronto at the end of the 90s and then in Paris, class which provides a panorama of the theory of dynamical systems of the 20th century including in particular KAM theory and the journey to ergodic theory and chaos. B. Bonnard: "I took notes and kept copies of his classes in Toronto which I unfortunately gave to him when he came back to Paris to teach this course. Im my opinion it should have be finalized as a book."
- **From 1998 onwards.** He collaborated with David Holcman, who describes this collaboration in his article in the Gazette de la Société Mathématique de France. In brief, this collaboration ranged from the study of linear PDEs on Riemannian manifolds, via the analysis of the singular perturbation of the first eigenvalue of the Laplacian, to the construction of Lyapunov functions on manifolds for dynamical systems [Morse-Smale type] for the convergence [when the diffusion term tends to zero] of this first eigenvalue to Topological Pressure. Other studies concerned conformal transformations of the triangle to study the meeting times of two Brownian particles on an interval; the calculation of loop formation times for Rouse-type polymers using the analysis of the Fokker-Planck equation in high dimensions; the analysis of passage times using Chavel-Feldman theory; and the calculation of the time of the first spermatozoon to find a small ovum in geometries with singularities (to model the shape of the uterus). Ivan had already shown an early interest in stochastic processes, which he had taught (notably at Grenoble).

Cet ouvrage est un hommage à Ivan Kupka disparu à Pâques 2023. Il contient un ensemble d'articles rédigés par des collègues travaillant en systèmes dynamiques ou en théorie du contrôle. Ceux-ci ont soit rencontré directement Ivan Kupka soit indirectement à travers ses articles. Des souvenirs personnels et une description de ses travaux pourront être lus dans un volume de la Gazette des Mathématiques qui paraîtra en 2024 et qui illustre la trajectoire de ce mathématicien exceptionnel qui fut tout aussi clairvoyant dans ses rapports bienveillants avec les individus. Il nous laisse à tous la mémoire d'un scientifique très travailleur et cultivé (un savant donc), doté d'une curiosité exceptionnelle et un grand amoureux de la liberté. Pour ceux qui ont eu la chance de le côtoyer il était aussi un ami fidèle et toujours disponible.

Ci-dessous une description d'une partie de sa contribution scientifique en contrôle représentant un travail partiellement en collaboration ou '' à proximité".

- **Fin des années 70: Conditions suffisantes de contrôlabilité pour des triplets de champs de vecteur** $\{A, \pm B\}$**.** *En collaboration avec V. Jurdjevic. Dans ce travail les auteurs définissent de façon très complète les opérations permises sur le triplet pour l'élargir tout en préservant sa contrôlabilité (construction du saturé de Lie). Le point technique le plus important étant l'utilisation de la classification des algèbres de Lie semi-simples pour en déduire des conditions suffisantes de contrôlabilité pour les systèmes invariants sur les groupes de Lie semi-*

simples (le triplet étant des champs invariants). Ce travail s'inscrivant bien dans le cadre des mathématiques classiques : action transitive d'un semi-groupe. On a ensuite utilisé ces travaux pour le cas des produits semi-directs.

- **A partir des années 80: Conditions d'optimalité du second-ordre pour les trajectoires singulières et propriétés génériques.** *En collaboration avec B. Bonnard pendant une dizaine d'années notamment pour étudier le rôle des trajectoires singulières en contrôle optimal. Ont été obtenues en particulier des conditions d'optimalité du second-ordre pour les trajectoires singulières pour les systèmes affines et mono-entrée. Ces conditions ont été implémentées numériquement avec un algorithme qui diffère du calcul initial qui utilise une forme semi-normale pour évaluer la dérivée seconde intrinsèque de l'application extrémité et une représentation adaptée (de Fourier) du contrôle. En parallèle il y a eu beaucoup de temps et de preuves fausses avant d'établir deux propriétés génériques des trajectoires singulières pour les systèmes affines mono-entrée : ordre minimal et stricte anormalité.*
  Géométrie sous-Riemannienne. Travail en géométrie sous-Riemannienne sur le cas dit Martinet pour étudier le rôle des trajectoires anormales et calculer la sphère de petit rayon. Ceci dans le contexte de la thèse de M. Chyba, puis celle d'E. Trélat. Ce travail a son origine dans un article jamais publié que Kupka qualifiait de taupinal sur l'utilisation des fonctions elliptiques pour des calculs explicites sur un modèle assez stable inspiré des travaux de R. Montgomery.
- **Milieu des années 80 : Classifications des extrémales bang-bang et phénomène de Fuller.** *Ce travail contient deux contributions importantes. La première est le calcul de formes normales pour les systèmes hamiltoniens discontinus et la classification des extrémales bang-bang génériques. Le second point culminant de cette activité très techniques a été la preuve par Ivan Kupka de l'ubiquité du phénomène de Fuller. B. Bonnard: "J'ai souvenir d'avoir écouté pendant longtemps ses tentatives successives de preuve et sa preuve complète constitue un chapitre assez confidentiel* [1] *."*
- **Fin des années 90 et debut années 2000.** *La maîtrise technique d'Ivan Kupka et son travail acharné culminent aussi dans son cours de master qu'il a enseigné plusieurs années à Toronto dès la fin des années 90, puis à Paris et qui fait un panorama complet de la théorie des systèmes dynamiques du 20-ième siècle incluant en particulier la théorie de KAM et le chemin vers la théorie ergodique et le chaos. B. Bonnard: "J'avais pris des notes et gardé les photocopies de ses cours de Toronto que je lui ai malheureusement données quand il est revenu enseigner ce cours à Paris. A mon avis cela aurait justifié d'être finalisé dans un livre."*
- **A partir de 1998.** *Collaboration avec David Holcman et celui-ci décrit cette collaboration dans son article de la Gazette de la Société Mathématique de France. En bref, cette collaboration allait de l'étude des EDP linéaires sur les variétés Riemanniennes, en passant par l'analyse de la perturbation singulière de la première valeure propre du Laplacien, mais aussi la construction de fonctions de Lyapunov sur les variétés pour des systèmes dynamiques [type Morse-Smale] pour la convergence [quand le terme de diffusion tend vers zéro] de cette première valeur propre vers la Pression Topologique. D'autres études concernaient les transformations conformes du triangle pour étudier les temps de rencontre de deux particules Browniennes sur un intervalle; le calcul du temps de la formation des boucles pour des polymères de type Rouse en utilisant l'analyse de l'équation de Fokker-Planck en grandes dimensions, l'analyse des temps de passage grâce à la théorie de Chavel-Feldman ou encore le calcul du temps du premier spermatozoïde*

[1] I. Kupka, "The ubiquity of Fuller's phenomenon," in: Nonlinear Controllability and Optimal Control, Monograph Textbook, Pure Appl. Math., No. 133 (ed. by H. Sussman), Dekker, New York (1990), pp. 313–350.

pour trouver un petit ovule dans des géométries avec des singularités (pour modéliser la forme de l'utérus). Très tôt déjà Ivan, avait commencé à s'intéresser aux processus stochastiques qu'il avait enseignés (notamment à Grenoble).

Bernard Bonnard, Monique Chyba,
David Holcman, Emmanuel Trélat

Thank you to all authors for their help to finalize this volume.

A Tour Through Ivan's Life

Contents

1

Optimality of Regular Extremals

Ivan Kupka

Portail des Mathématiques Jussieu-Chevaleret, Department of Mathematics, Paris, France

The first contribution to this volume are unpublished notes from Ivan Kupka himself to Bernard Bonnard about optimality of regular extremals. These notes are partially used in Chapter 1 of the book "Optimal Control with Applications in Space and Quantum Dynamics" by B. Bonnard and D. Sugny. AIMS Series on Applied Mathematics, Vol. 5 (2012).

I – OPTIMALITÉ DES EXTREMALES ORDINAIRES

X = espace d'état | $F: X \times U \to TX$ dynamique | $c: X \times U \to \mathbb{R}$ coût
U = " des entrée

$\pi: T^*X \to X$ projection canonique. $\sigma \in \Lambda^1 T^*X$ la forme canonique : $\sigma(w) = \langle T\pi(w), \pi_{T^*X}(w) \rangle$ $\pi_{T^*X}: TT^*X \to T^*X$ projection canonique du fibré tangent de T^*X

$\omega = -d\sigma$ est la structure symplectique sur T^*X

Lemme bien connu :

une 0 : Soit (M, ω_M) une variété lisse symplectique ; $V \subset M$ une hypersurface lisse. Si Λ est une sous variété Lagrangienne de M contenue dans V alors pour tout $m \in \Lambda$

$$T_m\Lambda \supset \ker(\omega | T_m V)$$

$\ker \omega | T_m V = \{ w \in T_m V \mid$ pour tout vecteur $v \in T_m V$ $\omega(w, v) = 0 \}$. Il est clair que $\ker(\omega | T_m V)$ est de dimension 1 : Notons que si $V = \{f = 0\}$ avec 0 valeur régulière de f, $\ker(\omega | T_m V) = \mathbb{R} df(m)$.

$H_\lambda : T^*X \times U \longrightarrow \mathbb{R}$ $H_\lambda(\xi, u) = \langle F(\pi(\xi), u), \xi \rangle - \lambda c(\pi(\xi),$

12

Un critère élémentaire d'optimalité :

Proposition 0 : Soit $(\bar z, \bar u) : \bar J = [\bar\alpha, \bar\beta] \longrightarrow T^*X \times U$ une extrémale ordinaire (pour H_1). Supposons qu'il existe un voisinage ouvert W de $\bar x(\bar J)$, $\bar x = \pi \circ \bar z$, et deux fonctions lisses $S : W \to \mathbb{R}$, $\hat u : W \to U$ telles que : (i) $\bar z(t) = dS(\bar x(t)) \in T^*_{\bar x(t)} X$ pour tout $t \in \bar J$

$\bar u(t) = \hat u(\bar x(t)) \in U$ " " " " "

(ii) pour tout $(x,u) \in W \times U$: $H_1(dS(x), u) \leq H_1(dS(x), \hat u(x))$,

(iii) : il existe une constante h telle que :

$H_1(dS(x), \hat u(x)) = h$ pour tout $x \in W$

Alors la trajectoire $(\bar x, \bar u) : \bar J \longrightarrow X \times U$ $\bar x = \pi \circ \bar z$ est optimale parmi toutes les trajectoires $(x,u) : [\alpha, \beta] \to W \times U$ du système telles que : 1) $x(\alpha) = \bar x(\bar\alpha)$, $x(\beta) = \bar x(\bar\beta)$

2) Uniquement dans le cas $c \neq 0$, $\beta - \alpha = \bar\beta - \bar\alpha$.

Preuve : posons $\mathcal{C}(x,u) = \int_\alpha^\beta c(x(t), u(t))\, dt = \text{coût}$

de la trajectoire $(x,u):[\alpha,\beta] \longrightarrow W \times U$. Par definition de H_1 on a que $\pi(dS(x)) = x$ pour tout $x \in X$:

$$c(x(t),u(t)) = \langle F(x(t),u(t),dS(x(t))\rangle - H_1(dS(x(t)),u(t))$$

$$c(x(t),u(t)) = dS(x(t)).\frac{dx(t)}{dt} - H_1(dS(x(t)),u(t)) = \frac{d}{dt}\left[S(x(t))\right] - H_1(dS(x(t)),u(t))$$

$$\mathcal{C}(x,u) = \int_\alpha^\beta c(x(t),u(t))dt = S(x(\beta)) - S(x(\alpha)) - \int_\alpha^\beta H_1(dS(x(t)),u(t))\,dt$$

$$\mathcal{C}(\bar{x},\bar{u}) = S(\bar{x}(\bar{\beta})) - S(\bar{x}(\bar{\alpha})) - \int_\alpha^\beta H_1(dS(\bar{x}(t)),\bar{u}(t))\,dt$$

Puisque $S(x(\beta)) = S(\bar{x}(\bar{\beta}))$, $S(x(\alpha)) = S(\bar{x}(\bar{\alpha}))$.

$$\mathcal{C}(x,u) - \mathcal{C}(\bar{x},\bar{u}) = \int_{\bar{\alpha}}^{\bar{\beta}} H_1(dS(\bar{x}(t),\bar{u}(t))dt - \int_\alpha^\beta H_1(dS(x(t),u(t))\,dt$$

Mais d'après (i) $\bar{u}(t) = \hat{u}(x(t))$ et d'après (iii) $H_1(dS(\bar{x}(t)),\hat{u}(\bar{x}(t))) = h$ et d'après (ii) et (iii) $H_1(dS(x(t)),u(t)) \leq h$.

$H_1(dS(x(t)),\hat{u}(x(t))) = h$.

Donc: $\mathcal{C}(x,u) - \mathcal{C}(\bar{x},\bar{u}) \geq h\left[\bar{\beta} - \bar{\alpha} + \alpha - \beta\right]$.

Si $h = 0$ $\mathcal{C}(x,u) \geq \mathcal{C}(\bar{x},\bar{u})$.

Si $h \neq 0$ mais $\beta - \alpha = \bar{\beta} - \bar{\alpha}$ de nouveau

$$\mathcal{C}(x,u) \geq \mathcal{C}(\bar{x},\bar{u}).$$

4

Définissons $\mathcal{H} : T^*X \to \overline{\mathbb{R}} = [-\infty, +\infty]$ par : $\mathcal{H}(z) = \sup_{u \in U} H_1(z, u)$. Notons que $\mathcal{H}$ est constante sur toute extremale ordinaire

Corollaire : Soit $(\bar{z}, \bar{u}) : \bar{J} = [\bar{\alpha}, \bar{\beta}] \to T^*X \times U$ une extremale ordinaire Les hypothèses de la Proposition 0 seront vérifiées s'il existe une sous variété Lagrangienne de T^*X et une fonction lisse $u_\Sigma : \Sigma \to U$ telles que : (i) $\bar{z}(t) \in \Sigma$ et $\bar{u}(t) = u_\Sigma(\bar{z}(t))$ pour tout $t \in \bar{J}$ (ii) $\Sigma \subset \{\mathcal{H} = h\}$ où h est la valeur de la fonction constante $\mathcal{H} \circ \bar{z} : \bar{J} \to \overline{\mathbb{R}}$ (iii) Pour tout $z \in \Sigma$ et tout $u \in U$, $H_1(z, u) \leq H_1(z, u_\Sigma(z)) = \mathcal{H}(z) = h$ (iv) La restriction de π à Σ est un difféomorphisme de Σ sur un ouvert W de X tel que $H_1(W; \mathbb{R}) = 0$

Preuve : soit $\varphi : W \to \Sigma$ l'inverse de la restriction $\pi|\Sigma$. Puisque Σ est Lagrangienne $d\varphi^*(\sigma) = -\varphi^*(d\sigma) = -\varphi^*(\omega|_\Sigma) = 0$. Comme $H^1(W; \mathbb{R}) = 0$

5

$\varphi^*(\sigma)$ est exacte : il existe une fonction lisse $S : W \to \mathbb{R}$ telle que $dS = \varphi^* \sigma$. Soit $z \in \Sigma$, $\pi(z) = \xi \in W$. Par définition de σ, pour tout $w \in T_z T^*X$, $\sigma(w) = \langle T\pi(w), z \rangle$. Prenons alors $v \in T_\xi X$ quelconque. Alors $T\varphi(v) \in T_z T\Sigma \subset T_z T^*X$ et $\sigma(T\varphi(v)) = \langle T\pi \circ T\varphi(v), z \rangle = \langle v, z \rangle$. Mais $\sigma(T\varphi(v)) = \varphi^*(\sigma)(v) = dS(\xi).v$. Donc pour tout $v \in T_\xi X$, $dS(\xi).v = \langle v, z \rangle$ i.e. $z = dS(\xi)$. Ceci montre que $\Sigma = \{ dS(x) \mid x \in W \}$. Prenons alors $\hat{u} : W \to \mathbb{R}$ comme $\hat{u} = u_\Sigma \circ \varphi$. Alors toutes les hypothèses de la Proposition 0 sont vérifiées : (i) de la Prop. résulte de (i) du Cor. et de la définition $\hat{u} = u_\Sigma \circ \varphi$ de $\hat{u}$ (ii) de la Prop. résulte de (iii) du Cor. et de la propriété $\Sigma = dS(W)$. (iii) de la Prop. résulte de (ii) du Cor.

Notons qu'on n'a pas supposé que Σ est tangent au champs de vecteurs hamiltonien $\vec{H}$ de H.

6

si ce champs existe. Mais le Lemme suivant montre que c'est automatique.

Lemme 1 : Soit Λ une sous-variété Lagrangienne de T^*X sur laquelle $\mathcal{H}$ est constante et qui possede un voisinage où $\mathcal{H}$ est lisse. Alors pour tout $z \in \Lambda$, $\overrightarrow{\mathcal{H}}(z)$ appartient à $T_z\Lambda$.

C'est une conséquence immédiate du Lemme 0

Maintenant le théorème principal.

THEOREME Soit $z_0 \in T^*X - 0_X$. Supposons qu'il existe un voisinage ouvert N de z_0 dans T^*X et une fonction lisse $u_N : N \to U$, tels que : (i) pour tout $z \in N$, $H_1(z, u_N(z)) = \mathcal{H}(z)$ et $\frac{\partial H_1}{\partial u}(z, u_N(z)) = 0$ (ii) $F(\pi(z_0), u_N(z_0)) \neq 0$

Alors il existe un voisinage ouvert Ω de z_0 dans T^*X, de projection $\mathcal{O} = \pi(\Omega)$ sur X telle que $H^1(\mathcal{O}; \mathbb{R}) = 0$, (Notons que $\mathcal{O}$ est un ouvert de X) ayant la propriété suivante :

17

$(\bar{z}, \bar{u}): \bar{J} = [\bar{\alpha}, \bar{\beta}] \longrightarrow \Omega \times U$ est une extremale ordinaire du système telle que $\bar{u}(t) = u_N(\bar{z}(t))$ pour presque tout $t \in \bar{J}$, la projection $(\bar{x}, \bar{u}): \bar{J} \longrightarrow O \times U$, $\bar{x} = \pi \circ \bar{z}$ est optimale parmi toutes les trajectoires (x, u): $[\alpha, \beta] \to O \times U$ du système telles que: 1) $x(\alpha) = \bar{x}(\bar{\alpha})$, $x(\beta) = \bar{x}(\bar{\beta})$ et 2) uniquement si $\mathcal{H} \circ \bar{z} = 0$ (rappelons que c'est une constante), $\beta - \alpha = \bar{\beta} - \bar{\alpha}$

Preuve: Pour demontrer ce theorème il suffit de montrer que Ω est feuilleté par un feuilletage $\mathcal{F}$ ayant les propriétés suivantes: chaque feuille $\mathcal{L}$ de $\mathcal{F}$ est une sousvariété Lagrangienne de $\dot{T}^*X$ contenue dans un ensemble de niveau de $\mathcal{H}$ et la restriction de π à $\mathcal{L}$ est un difféomorphisme de $\mathcal{L}$ sur O. En effet soit $(\bar{z}, \bar{u}): \bar{J}: [\bar{\alpha}, \bar{\beta}] \to \Omega \times U$ une extremale ordinaire du système telle que $\bar{u}(t) = u_N(\bar{z}(t))$ pour presque tout $t \in \bar{J}$.

8

Soit $\mathcal{L}_0$ la feuille de $\mathcal{F}$ contenant $\bar{\gamma}(\bar{\alpha})$. D'après le Lemme 1 $\mathcal{L}_0$ est tangente à $\overline{\mathcal{H}}$. Donc elle contient toute la courbe $\bar{\gamma}(\bar{s})$. En prenant $\Sigma = \mathcal{L}_0$, $u_\Sigma = u_N|_{\mathcal{L}_0}$, $W = 0$ on voit que toutes les hypothèses du Corollaire sont vérifiées et on obtient le théorème. Il faut donc construire F.

Notons d'abord que $\mathcal{H}$ est lisse dans N : si $\gamma \in N$, $\mathcal{H}(\gamma) = H_1(\gamma, u_N(\gamma))$. Prenons un système de coordonnées $x^1, \dots, x^d : D \to \mathbb{R}$ de X telles que : $D \subset \pi(N)$, $x^i(x_0) = 0$, $1 \le i \le d$, $x_0 = \pi(\gamma_0)$ et $dx^i\big(F(x_0, u_N(\gamma_0))\big) = \begin{cases} 1 & i = d \\ 0 & 1 \le i < d \end{cases}$ $d = \dim X$.

Ce système s'étend en un système de coordonnées symplectiques $\hat{x}^1, \dots, \hat{x}^d, \hat{f}_1, \dots, \hat{f}_d : \overline{\pi}^{-1}(D) \to \mathbb{R}$ de T^*X : $\hat{x}^i = x^i \circ \pi$. $\hat{f}_i(\alpha) = \alpha\left(\frac{\partial}{\partial x^i}\right)$ pour tout $\alpha \in T^*D = \overline{\pi}^{-1}(D)$. $\hat{f}_1, \dots, \hat{f}_d$ est le corepère dual du repère $\frac{\partial}{\partial x^1}, \dots, \frac{\partial}{\partial x^d}$

19

sur D. Comme $\frac{\partial \mathcal{H}}{\partial \hat{p}_i}(z) = \frac{\partial H_1}{\partial \hat{p}_i}(z, u_N(z)) = dx^i(F(\pi(z), u_N(z))$

pour tout $z \in N$, tout i $1 \leq i \leq d$:

$$\frac{\partial \mathcal{H}}{\partial \hat{p}_i}(z_0) = 0, \quad 1 \leq i \leq d-1, \quad \frac{\partial \mathcal{H}}{\partial \hat{p}_d}(z_0) = 1$$

(relativement compact)

On peut alors trouver un voisinage ouvert Z de z_0 dans l'hypersurface lisse $\bar{\pi}(D) \cap \{x^d = 0\}$ tel que les restriction à Z des fonctions $\hat{x}^1, \ldots, \hat{x}^{d-1}, \hat{p}_1, \ldots, \hat{p}_{d-1}, \mathcal{H}$ forment un système de coordonnées sur Z : en effet en z_0 : $d\hat{x}^1 \wedge \ldots \wedge d\hat{x}^{d-1} \wedge d\hat{p}_1 \wedge \ldots \wedge d\hat{p}_{d-1} \wedge d\mathcal{H} =$

$$d\hat{x}^1 \wedge \ldots \wedge d\hat{x}^{d-1} \wedge d\hat{p}_1 \wedge \ldots \wedge d\hat{p}_{d-1} \wedge \left(dp_d + \frac{\partial \mathcal{H}}{\partial x^d}(z_0)\, dx^d\right) \neq 0.$$

Puisque Z est relativement compact, il existe un $\varepsilon_0 > 0$ et une application lisse $\Phi : [-\varepsilon_0, \varepsilon_0] \times Z \longrightarrow N$ telle que: (i) $\Phi(0, z) = z$ pour $t, z \longrightarrow \Phi(t, z)$

tout $z \in Z$ (ii) $\frac{\partial \Phi}{\partial t} = \vec{\mathcal{H}} \circ \Phi$ (iii) L'application $\Phi : [-\varepsilon_0, \varepsilon_0] \times Z \longrightarrow \mathcal{D} = \bigcup_{|t| \leq \varepsilon_0} \Phi(t, Z)$ est un difféomorphisme. Définissons alors des fonctio

$\xi^1, \dots, \xi^d, \eta_1, \dots, \eta_d : \mathcal{D} \to \mathbb{R}$ comme suit : $\xi^i(\Phi(t,z)) = \hat{x}^i(z)$

$\eta_i(\Phi(t,z)) = \hat{h}_i(z)$ pour $1 \le i \le d-1$, $\xi^d(\Phi(t,z)) = t$

$\eta_d(\Phi(t,z)) = \mathcal{H}(z) = \mathcal{H}(\Phi(t,z))$ (i.e. $\eta_d = \mathcal{H} \mid \mathcal{D}$)

Il est clair que $\xi^1, \dots, \xi^d, \eta_1, \dots, \eta_d : \mathcal{D} \to \mathbb{R}$ forment un système de coordonnées sur $\mathcal{D}$. J'affirme que ce système est symplectique. Pour voir ceci notons d'abord que $\mathcal{D}$ est feuilleté par les hypersurfaces $\Phi(t, Z)$.

Soient $f, g : \pi^{-1}(D) \to \mathbb{R}$ deux fonctions lisses telles que $\{f, x^d\} = \{g, x^d\} = 0$. Leurs hamiltoniens $\vec{f}, \vec{g}$ sont tangents à Z le long de Z. Définissons les fonctions $f_1, g_1 : \mathcal{D} \to \mathbb{R}$ par $f_1(\Phi(t,z)) = f(z)$, $g_1(\Phi(t,z)) = g(z)$. Alors les hamiltoniens $\vec{f}_1, \vec{g}_1$ sont tangents aux feuilles $\Phi(t, Z)$. En effet $\vec{f}_1(\Phi(t,z)) = T\Phi_t(\vec{f}(z))$, $\vec{g}_1(\Phi(t,z)) = T\Phi_t(\vec{g}(z))$ pour tous les $t \in (-\varepsilon, \varepsilon)$, tous $z \in$ où $\Phi_t : Z \to \Phi(t,z)$ est le difféomorphisme $z \mapsto \Phi(t,z)$. Utilisant le fait que Φ_t est symplecti

11

on voit que: $\{f_1, g_1\}(\Xi(t,z)) = \omega\left(\vec{f}_1(\Xi(t,z)), \vec{g}_1(\Xi(t,z))\right) =$

$\omega\left(T\Xi_t(\vec{f}(z)), T\Xi_t(\vec{g}(z))\right) = \omega\left(\vec{f}(z), \vec{g}(z)\right) = \{f, g\}(z)$

pour tous les $(t;z) \in [-\varepsilon_0, \varepsilon_0] \times Z$.

En prenant pour f, g des fonctions parmi $\hat{x}^1, \ldots, \hat{x}^{d-1}$

$\hat{p}_1, \ldots, \hat{p}_{d-1}$ on voit que pour tous les i, j $1 \le i, j \le d-1$

$\{\xi^i, \xi^j\} = \{\hat{x}^i, \hat{x}^j\} = 0$, $\{\eta_i, \eta_j\} = \{\hat{p}_i, \hat{p}_j\} = 0$

$\{\xi^i, \eta_j\} = \{\hat{x}^i, \hat{p}_j\} = \delta^i_j$.

Finalement puisque ξ^d est constante sur les

feuilles $\Xi(t,z)$ et que les $\vec{\xi}^i$, $\vec{\eta}_j$, $1 \le i, j \le d-1$

sont tangents à ces feuilles, $\{\xi^d, \xi^i\} = d\xi^d(\vec{\xi}^i) = 0$

$\{\xi^d, \eta_i\} = d\xi^d(\vec{\eta}_i) = 0$ si $1 \le i \le d-1$.

$\{\xi^i, \eta_d\}(\Xi) = d\xi^i(\vec{\mathcal{H}})(\Xi) = d\xi^i\left(\frac{\partial \Xi}{\partial t}\right) = \frac{\partial}{\partial t}\left(\xi^i(\Xi)\right) = \frac{\partial}{\partial t}\hat{x}^i = 0$

De même $\{\eta_i, \eta_d\} = 0$, $1 \le i \le d-1$.

Et: $\{\xi^d, \eta_d\}(\Xi) = d\xi^d(\vec{\mathcal{H}})(\Xi) = \frac{\partial}{\partial t}\left(\xi^d(\Xi)\right) = \frac{\partial}{\partial t} t = 1$

On considère alors le feuilletage

$\mathcal{F}'$ de $\mathcal{D}$ défini par les composantes connexes

(12

des variétés de niveau $\{\eta_1 = c_1, \eta_2 = c_2, \dots, \eta_d = c_d\}$

Les feuilles de $\mathcal{F}'$ sont Lagrangiennes car $\{\eta_i, \eta_j\} = 0$, $1 \le i, j \le d$, et contenues dans les ensembles de niveau de $\mathcal{H}$ car $\eta_d = \mathcal{H}$. Si $\mathcal{L}'$ est une feuille et si $z \in \mathcal{L}' \cap \{x^d = 0\} = \mathcal{L}' \cap \{\xi^d = 0\}$, $T_z\mathcal{L}' = \bigcap_{i=1}^{d} \{d\eta_i|_z = 0\}$

$= \bigcap_{i=1}^{d-1} \{d\hat{F}_i|_z = 0\} \cap \{d\mathcal{H}|_z = 0\}$. Donc $T\pi : T_z\mathcal{L}' \to T_x X$, $x = \pi(z)$, est un isomorphisme. Ceci montre qu'il existe un voisinage ouvert $\mathcal{O}$ de z_0 tel que pour toute feuille $\mathcal{L}'$ de $\mathcal{F}'$, toute composante connexe de $\bar{\pi}^{-1}(\mathcal{O}) \cap \mathcal{L}'$ est appliquée difféomorphiquement sur $\mathcal{O}$ (Rappelons que Z est relativement compact). On prend pour Ω l'ensemble $\bar{\pi}^{-1}(\mathcal{O}) \cap \mathcal{D}$ et pour $\mathcal{F}$ le feuilletage induit sur Ω par $\mathcal{F}'$.

2

Optimal Control of the Lotka-Volterra Equations with Applications

Bernard Bonnard[1] and Jérémy Rouot[2]

[1] Institut Mathématique de Bourgogne, 9 rue Alain Savary, 21000 Dijon, France and Inria Sophia Antipolis. bernard.bonnard@u-bourgogne.fr
[2] Univ Brest, UMR CNRS 6205, Laboratoire de Mathématiques de Bretagne Atlantique, 6 avenue Le Gorgeu 29200 Brest jeremy.rouot@univ-brest.fr

Hommage personnel de Bernard Bonnard dans ce volume.

Plutôt que des discussions philosophiques et des hommages qu'Ivan appréciait peu j'ai préféré inclure dans ce volume deux articles écrits en collaboration avec des jeunes collègues et qui relèvent de son héritage.

Le premier co-écrit avec Jérémy Rouot utilise le modèle de dynamique des populations dû à Lotka et Volterra pour traiter le problème de réduction de l'infection d'un microbiote. Dans une première partie le problème est analysé avec les techniques de contrôle optimal géométrique. La seconde est motivée par une approche qui s'inscrit bien dans le cadre très récent inspiré de ma participation au congrès AMS d'Hawaii en 2018 où les mathématiques discrètes se sont positionnées en force en lien avec les progrès de la technologie du numérique. Nous proposons une analyse du problème dans le cadre des contrôles digitaux qui prennent bien en compte les contraintes logistiques médicales et une étude avec des techniques de commande optimale prédictive (d'optimisation donc). Je pense qu'Ivan aurait bien aimé cette impertinence par rapport aux mathématiques ''académiques ''. En tout cas j'ai voulu faire preuve de pragmatisme pour traiter ce type de problème. J'espère aussi rejoindre le point de vue des biologistes où le modèle ne sert qu'à calculer avec des opérations dans le cadre de l'algèbre linéaire les équilibres et leur stabilité. En l'occurrence pour le microbiote intestinal avec onze variables d'état jusqu'à $2^{11} = 2048$ *possibilités. La technique de calcul à horizon glissant conduit par ailleurs à déterminer le contrôle optimal en boucle fermée et présente beaucoup de souplesse pour modifier le modèle selon les domaines et le critère d'optimisation en incluant toutes les contraintes notamment digitales sur le contrôle.*

Le second co-écrit avec Olivier Cots, Yannick Privat et Emmanuel Trélat concerne le problème de Zermelo traité dans le cadre Hamiltonien du contrôle géométrique et appliqué à des problèmes de physique (contrôle quantique, transfert orbital et micro-magnétisme). Le contexte géométrique consiste à utiliser le groupe feedback pour classifier les systèmes et leurs trajectoires extrémales et construire des formes normales pour estimer l'ensemble des états accessibles dans le cas continu. Cette classification fait un premier tri entre le cas intégrable et non intégrable. Dans le cas intégrable une partie numérique complète l'analyse géométrique et permet une description complète du lieu conjugué et de coupure, illustrant bien à mon avis l'intérêt de combiner les deux approches. Dans le cas non intégrable la dynamique des géodésiques est très complexe car c'est une dynamique 3D

mais se traite en combinant les méthodes théoriques et numériques. Une étape complémentaire étant d'utiliser un modèle discrétisé pour estimer l'ensemble des états accessibles et sa frontière.

Summary. In this article, the Lotka–Volterra model is analyzed to reduce the infection of a complex microbiote. The problem is set as an optimal control problem, where controls are associated to antibiotic or probiotic agents, or transplantations and bactericides. Candidates as minimizers are selected using the Maximum Principle and the closed loop optimal solution is discussed. In particular a $2d$–model is constructed with four parameters to compute the optimal synthesis using homotopies on the parameters. It is extended to the $3d$–case to provide a geometric frame to direct and indirect numerical schemes.

2.1 Introduction and $2d$-Geometric Analysis

The Lotka-Volterra equations is a model to study biological species interactions and comes from a generalization of the prey-predator model, see [21]. In this memoir the problem is already set in the control frame since the model aims to explain the evolution of two fishing species in relation with diminution of the fishing activity during the first World War.

The system is written as the 2d-dynamics:

$$\frac{\mathrm{d}N_1}{\mathrm{d}t} = N_1(\lambda_1 + \mu_1 N_2), \quad \frac{\mathrm{d}N_2}{\mathrm{d}t} = N_2(\lambda_2 + \mu_2 N_1) \tag{2.1}$$

where N_1, N_2 are the two species, $N_1, N_2 \geq 0$ and $\lambda_1, \lambda_2, \mu_1, \mu_2$ are real parameters. In the prey predator model $\lambda_1 > 0, \lambda_2 < 0, \mu_1 < 0, \mu_2 > 0$.

The system is conservative and can be integrated using the first integral:

$$\mu_2 N_1 + \lambda_2 \ln N_1 - (\mu_1 N_2 + \lambda_1 \ln N_2) = \text{constant}.$$

In the prey predator model, the evolution of each species in the quadrant $N_1, N_2 > 0$ is periodic and there exist a single persistent equilibrium: $\Omega = (K_1, K_2)$. Moreover K_1, K_2 represents the averaged population of each species on a period T

$$\langle N_i \rangle = \frac{1}{T} \int_0^T N_i(t)\mathrm{d}t = K_i, \ i = 1, 2.$$

The effect of the fishing activity is to replace:

$$\lambda_1 \to \lambda_1 - \alpha\lambda, \qquad \lambda_2 \to \lambda_2 - \beta\lambda,$$

where α, β are the modes of destruction of each species and $\lambda(t)$ is the control intensity.

Constant controls lead to shift the persistent equilibrium and hence to shift the averaged populations.

More generally the model leads to consider two vector fields (X, Y) defined by (2.1) with different parameters and to introduce the control system:

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = u(t)X(x(t)) + (1 - u(t))Y(x(t)),$$

$x = (N_1, N_2)$ and $u \in [0, 1]$.

The Lotka–Volterra equations can more generally described the interaction of n-species $x = (x_1, \ldots, x_n)^\intercal$, $x_i \geq 0$, and is given by the dynamics:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = (\mathrm{diag}x)(Ax + r), \tag{2.2}$$

where $\mathrm{diag}x$ is the diagonal matrix with entries $(x_1, \ldots, x_n)$, $A = (a_{ij})$ is the matrix of interaction coefficients and $r = (r_1, \ldots, r_n)^\intercal$ is the vector of individual growth of the species. Recently based on the model of [20] of the intestinal microbiote with $n = 11$ species, Jones et al. [13] analyzed the problem of reducing C. difficile infection (a pathogenic agent) using either antibiotic or fecal transplantation.

Denoting by $X(x) = (\mathrm{diag}x)(Ax + r)$ the n-dimensional dynamics ($n = 11$) with parameters given in [20], the control system writes as:

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = X(x(t)) + u(t)Y(x(t)) + \sum_{i=1}^{k} \lambda_i \delta(t - t_i)Y'(x)), \tag{2.3}$$

where $Y(x) = (\mathrm{diag}x)\epsilon$, $\epsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ is the sensitivity vector to the antibiotic of the species and $u(t)$ is a piecewise constant mapping. The second control action is associated to jumps $x(t_i) \to x(t_i) + \lambda v$ in the state, and $Y'(x) = v$, corresponding to ratio of each species in the transplantation.

Denoting by x_1 the C. difficile population, the optimal control problem can be set as a Mayer problem: $\min x_1(t_f)$ where t_f is the number of days of the treatment or in a dual form: reach in minimum time t_f a specific level d of infection that is: $x_1(t_f) = d$.

The optimal control problem can be posed in the general frame of mixing permanent controls associated to antibiotic treatment or sampled-data controls associated to transplantations.

In both case the optimal control problem can be analyzed with an indirect scheme based on the Maximum Principle [17] in the permanent case or an adaptation in the sampled-data control case, or by a direct numerical optimization scheme.

In this article, the starting point is to analyze the effect of an antibiotic or probiotic treatment restricting to a control system of the form:

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = X(x(t)) + u(t)Y(x(t)),$$

with $x(t) \in \mathbb{R}^n$, the set of admissible controls $\mathscr{U}$ being the set of measurable mappings valued in $]-1, +1[$ (for convenience we assume $u = -1$ being associated to no treatment, $u = +1$ to maximum dosing regimen). We consider the problem of steering $x(0) = x_0$ to a terminal manifold N of codimension one, e.g.: $x_1 = d$, in minimum time. We mainly focus our study the the $2d$–case.

Our analysis is based on a series of recent articles [4, 15, 5] to classify the closed loop optimal solutions in a neighborhood of the terminal manifold, using semi-normal forms for the triple (X, Y, N), under generic assumptions. They can be globalized in the frame of polynomic systems using homotopies on the parameters.

It is completed in the $3d$–case by direct and indirect numerical schemes to provide robust optimal controls taking into account the combination of various treatments and medical logistical constraints [5, 8, 18] using preliminary geometric analysis.

2.2 The Maximum Principle in the Permanent Case and the Classification of the Extremals

2.2.1 Maximum Principle

Denote $F(x,u) = p \cdot (X(x) + uY(x))$ and $H = p \cdot F(x,u)$ the Hamiltonian lift defining the pseudo-Hamiltonian, $p \in \mathbb{R}^n \setminus \{0\}$ being the adjoint vector. If $(x(.), u(.))$ is optimal on $[0, t_f]$ then there exists $(z(.), u(.))$, $z = (x, p)$ such that a.e. :

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t}(t) &= \frac{\partial H}{\partial x}(x(t), p(t), u(t)), \\ \frac{\mathrm{d}p}{\mathrm{d}t}(t) &= -\frac{\partial H}{\partial p}(x(t), p(t), u(t)). \end{aligned} \tag{2.4}$$

Moreover the optimal control satisfies a.e. the maximization condition

$$H(z(t), u(t)) = \max_{|v| \leqq 1} H((z(t)), v) = M(z(t)), \tag{2.5}$$

where $M((z(t)) \geq 0$ is constant.

At the final time the transversality condition is satisfied:

$$p(t_f) \perp T^*_{x(t_f)} N. \tag{2.6}$$

Definition 1. *An extremal (z, u) is a solution of (2.4)-(2.5) on $[0, t_f]$. It is called a BC–extremal if the transversality condition (2.6) is satisfied. An extremal is called regular if a.e. $u(t) = \mathrm{sign} H_Y(z(t))$ and singular if $H_Y(z(t)) = 0$ identically. A regular extremal is called bang-bang (BB) if the the number of switches is finite. An extremal (x, p, u) is called strict if $p(.)$ is unique up to a factor.*

2.2.2 Small time classification of regular extremals near the switching surface.

One needs the following see [14] for the details.

Let $t \to z(t)$ be a regular extremal on $[0, t_f]$ and we denote by $t \to \Phi(z(t)) = H_Y(z(t))$ the switching function and let Φ_ε the switching function along a bang arc extremal with $u = \varepsilon = \pm 1$ constant. We denote respectively by σ_+, σ_-, bang arcs with $u = \pm 1$ and σ_s a singular arc, while $\sigma_1\sigma_2$ denotes a σ_1 arc followed by an σ_2 (where each arc of the sequence can be empty). We denote by Σ the switching surface $H_Y(z) = 0$ and Σ' the subset $H_Y(z) = \{H_Y, H_X\}(z) = 0$. The Lie bracket of two vector fields Z_1, Z_2 being computed with the convention $[Z_1, Z_2](x) = \frac{\partial Z_1}{\partial x}(x) Z_2(x) - \frac{\partial Z_2}{\partial x}(x) Z_1(x)$. If $H_i(z) = p \cdot Z_i(x)$ the Poisson bracket is $\{H_1, H_2\} = \mathrm{d}H_1(\boldsymbol{H}_2) = p \cdot [Z_1, Z_2](x)$, where $\boldsymbol{H}_2 := (\nabla_p H_2, -\nabla_x H_2)$ is the Hamiltonian vector field.

Deriving twice the switching function $\Phi(t)$ one gets:

$$\begin{aligned} \frac{\mathrm{d}\Phi}{\mathrm{d}t}(t) &= \{H_Y, H_X\}(z(t)), \\ \frac{\mathrm{d}^2\Phi}{\mathrm{d}t^2}(t) &= \{\{H_Y, H_X\}, H_X\}(z(t)) + u(t)\{\{H_Y, H_X\}, H_Y\}(z(t)). \end{aligned} \tag{2.7}$$

Let t be a switching time so that $\Phi(t) = 0$ and assume that at $z(t)$ the surface Σ' is regular.

Proposition 1. *Assume that the switching time t is ordinary that is: $\Phi(t) = 0$ and $\frac{\mathrm{d}\Phi}{\mathrm{d}t}(t)$ is non zero. Then near $z(t)$ every extremal projects onto $\sigma_+\sigma_-$ if $\frac{\mathrm{d}\Phi}{\mathrm{d}t}(t) > 0$ or $\sigma_-\sigma_+$ if $\frac{\mathrm{d}\Phi}{\mathrm{d}t}(t) < 0$.*

Proposition 2. *Assume that at the switching time t, the switching function $\Phi_\varepsilon(t)$ for $u = \varepsilon = \pm 1$ is such that $\frac{\mathrm{d}\Phi_\varepsilon}{\mathrm{d}t}(t) = 0$ and both $\frac{\mathrm{d}^2\Phi_\varepsilon}{\mathrm{d}t^2}(t) \neq 0$ where the second order derivative is given by (2.7). Then $z(t)$ is called a fold point and we have:*

- *In the parabolic case: $\frac{\mathrm{d}^2\Phi_+}{\mathrm{d}t^2}(t) \cdot \frac{\mathrm{d}^2\Phi_-}{\mathrm{d}t^2}(t) > 0$, each extremal near $z(t)$ projects onto $\sigma_\pm\sigma_\pm\sigma_\pm$.*
- *In the hyperbolic case: $\frac{\mathrm{d}^2\Phi_+}{\mathrm{d}t^2}(t) > 0$, $\frac{\mathrm{d}^2\Phi_-}{\mathrm{d}t^2}(t) < 0$ it projects onto $\sigma_\pm\sigma_s\sigma_\pm$.*
- *In the elliptic case $\frac{d^2\Phi_+}{\mathrm{d}t^2}(t) < 0$, $\frac{\mathrm{d}^2\Phi_-}{\mathrm{d}t^2}(t) > 0$, every extremal is bang-bang but the number of switches is not uniformly bounded.*

2.2.3 Computations of the singular extremals with minimal order

The computations is standard, see [3]. Derive twice with respect to time $H_Y(z(t)) = 0$ one gets

$$\begin{aligned} &H_Y(z(t)) = \{H_Y, H_X\}(z(t)) = 0, \\ &\{\{H_Y, H_X\}, H_X\}(z(t)) + u_s(t)\{\{H_Y, H_X\}, H_Y\}(z(t)) = 0. \end{aligned} \tag{2.8}$$

Assume the generalized Legendre-Clebsch condition $\{\{H_Y,H_X\},H_Y\}(z(t)) \neq 0$ holds for every t then from equation (2.8), $u_s(t) = u_s(z(t))$ is the dynamic feedback:

$$u_s(z) = -\frac{\{\{H_Y, H_X\}, H_X\}(z)}{\{\{H_Y, H_X\}, H_Y\}(z)}$$

and plugging such u_s in the pseudo-Hamiltonian defines the true Hamiltonian:

$$H_s(z) = H_X(z) + u_s(z)H_Y(z).$$

Hence we deduce:

Proposition 3. *Singular extremals with minimal order $\{\{H_Y, H_X\}, H_Y\}(z) \neq 0$ are solutions of the Hamiltonian dynamics $\boldsymbol{H}_s(z)$ restricted to the invariant surface Σ': $H_Y(z) = \{H_Y, H_X\}(z) = 0$.*

Definition 2. *Assume that we are in the strict case. Since the true Hamiltonian is constant then the singular trajectories projections of singular extremals of minimal order are stratified according to the following:*

- *Hyperbolic case: $H_X(z).\{\{H_Y, H_X\}, H_Y\}(z) > 0$,*
- *Elliptic case: $H_X(z).\{\{H_Y, H_X\}, H_Y\}(z) < 0$,*
- *Abnormal or exceptional case: $H_X(z) = 0$.*

2.2.4 Construction of the optimal synhesis in a neighborhood of N

Take a point x_0 which can be identified to 0. Assume that at such point the surface N is regular. We denote by $N^\perp$ the Hamiltonian lift: $\{z = (x, p); x \in N, p = n(x)\}$ where n is the normal to N at x.We shall assume that the cone of limit directions $\{X \pm Y\}$ is strict and one can suppose it is contained in an half-space, so that n can be chosen assuming $n(x) \cdot X(x) > 0$.

If $n(x) \cdot Y(x) \neq 0$,then every extremal near N is determined by the transversality condition: $u = +1$ if $n \cdot Y > 0$ and $u = -1$ otherwise. Switches can occur only near points such that Y is tangent to N, that is $n \cdot Y(x) = 0$.

The regular synthesis [10] amounts to compute in a neighborhood U of x_0, in the domain $n \cdot X(x) < 0$ the following strata:

- The switching locus W restricting to ordinary switches with strata W_+, W_- corresponding respectively to $\sigma_-\sigma_+$ or $\sigma_+\sigma_-$, and associated to optimal policies only.
- The set Σ_s filled by optimal $BC-$ singular arcs.
- The cut locus C defined as follows. Every optimal arc $\sigma(t)$ is integrated backwards in time, that is $\sigma(t)$ is defined on $[t_f, 0]$ so that $t_f < 0$ and $\sigma(0) \in N$. The cut locus is the closure of the set of points $z(t_c)$, $t_f < t_c < 0$ so that $z(t)$ is not optimal beyond the time t_c. It contains the separating locus formed by the set of points where there exits two distinct minimizers reaching N.

The contribution of the series of papers [4, 15, 5] describes the time minimal syntheses for all cases of codimension ≤ 2 in the jet spaces of the triples (X, Y, N) at $x_0 = 0$. We shall present the main application, restricting to the $2d$–case for the controlled Lotka-Volterra model, to describe geometrically the main features of the time minimal syntheses.

2.3 The Geometric Determination of the Time Minimal Syntheses for the Lotka-Volterra Model – $2d$-Case

2.3.1 Determination of the collinearity locus in relation with forced permanent equilibria

Plugging $u = \pm 1$ leads to forced equilibria with constant dosing regimen associated to no treatment with $u = -1$ and maximal dosing regimen with $u = +1$.

Hence in the n-dimensional case we introduce the collinearity locus as the one-dimensional variety defines as projection on the state space of the set:

$$\{(x_e, \lambda) \in \mathbb{R}^{n+1}; \lambda = -u_e, X(x_e) = \lambda Y(x_e)\}.$$

The constant control u_e is such that (x_e, u_e) is a forced equilibrium and it has to be feasible that is $|u_e| \leq 1$.

Following Volterra [21] one can choose for each dynamics $(\mathrm{diag}x)(Ax + r)$ dimensionless coordinates so that up to translation the dynamics takes the form $-\mathrm{diag}(x + 1)A^*x$, where the persistent equilibrium is identified to 0 and the spectrum of the linearized dynamics is given by $-\sigma(A^*)$ with $\sigma(A^*) = \{\lambda_1, \ldots, \lambda_n\}$ where each λ_i denotes an eigenvalue, with generalized eigenspace E_{λ_i}.

In the 2d-case the computation of the collinearity locus is simple and is the determinantal set

$$\mathscr{C} = \det(X(x), Y(x)) = 0.$$

Straightforward computations define a segment L_1 when restricting to the persistent quadrant: $x_1, x_2 > 0$. Furthermore a subsegment L_1' is defined due to the control restriction $|u_e| \leq 1$.

Each point of this segment determines a forced equilibrium with a corresponding spectra.

Example 1. Consider the conservative case described by (2.1) with parameters $(\lambda_1, \lambda_2, \mu_1, \mu_2)$ and $\Omega = (K_1, K_2)$ be the persistent equilibrium. The dynamics can be set in normalized coordinates introducing $n_i = \frac{N_i}{K_i}$ and $n_i \to n_i - 1$ so that it takes the form: $-\mathrm{diag}(x+1)A^*x$. Choosing Ω in the quadrant $N_i > 0$ imposes constraints: $\lambda_1\mu_1 > 0$ and $\lambda_2\mu_2 < 0$. One can choose the ratio $\lambda = \lambda_2/\mu_2$ as an homotopy parameter and consider the one-dimensional dynamics $\lambda \to (\mathrm{diag}x)(A(\lambda)x + r(\lambda))$ where λ can be restricted to a segment.

2.3.2 Determination of the singular locus

In the $2d$–case, using $H_Y(z) = \{H_X, H_Y\}(z) = 0$, the singular locus is the determinantal set $\mathscr{S}$ defined by:

$$\det(Y(x), [Y, X](x)) = 0.$$

In the persistent space they formed a line passing through the origin.

For some parameters value, the collinear and singular loci intersects at a single point denoted O. The main point of this section will be to discuss the construction of the time minimal synthesis in a neighborhood of O, illustrating the applications of the concepts and techniques from [4, 15, 5]. This will lead to identify four parameters to construct the global syntheses by homotopy. The geometric schematic picture is represented on Fig.2.1 where we have reported symbolically on the extremities of the collinear locus the two cases studied by Volterra [21], illustrating clearly the global issues.

In the $2d$-case, much information about the global synthesis can be deduced using the clock form one-form ω defined outside the collinearity locus by the relations:

$$\omega(X) = p \cdot X(x) = 1, \quad \omega(Y) = p \cdot Y(x) = 0.$$

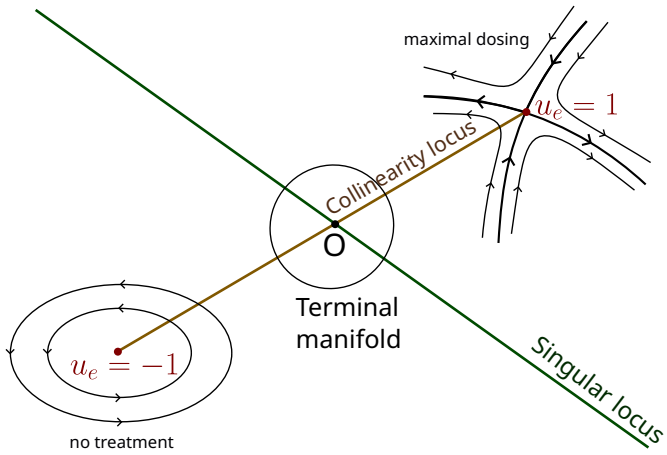


Fig. 2.1: Schematic representation of a case study: end–points of the collinear locus and intersection of the singular and collinear locus.

Green's theorem allows to deduces optimality status of $\sigma_+\sigma_-$ vs $\sigma_-\sigma_+$, in different domains, observing that $\mathrm{d}\omega$ vanishes precisely on the singular locus.

Since Lie brackets have complicated values, the use of a semi-normal form for the actions of local changes of coordinates and feedbacks $u \to -u$ aims to simplify the computations.

In particular, such a construction will be useful to deduce the time minimal synthesis in a neighborhood of 0 and identify the homotopy parameters to construct the global synthesis.

Construction of the semi-normal form

First of all, one can choose coordinates such that $O = (0, 0)$ and Y is identified to the vector field $Y = \frac{\partial}{\partial x_2}$ (this amounts mainly to choose ln-coordinates), furthermore the singular direction can be identified to the axis (Ox_1).

Expanding X in the jet space at $O = (0, 0)$, this leads to analyze the control system:

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = -\lambda x_1 + \alpha x_2^2, \frac{\mathrm{d}x_2}{\mathrm{d}t} = (u - u_e),$$

with $u_e \in]-1, +1[$, $\mid u \mid \leq 1$ and $\alpha > 0$.

Properties of the system

Computing Lie brackets in those coordinates shows relevant simplifications:

- $X(x) = (-\lambda x_1 + \alpha x_2^2)\frac{\partial}{\partial x_1} - u_e \frac{\partial}{\partial x_2}$,
- $Y(x) = \frac{\partial}{\partial x_2}$,
- $[Y, X](x) = -2\alpha x_2 \frac{\partial}{\partial x_1}$,
- $[[Y, X], Y](x) = -2\alpha \frac{\partial}{\partial x_1}$.

Hence the singular line is given by: $x_2 = 0$ and restricting to this line one has:

$$X(x_1) = -\lambda x_1 \frac{\partial}{\partial x_1}, \quad [[Y, X], Y](x_1) = -2\alpha \frac{\partial}{\partial x_1}.$$

Therefore for the restriction one has:

$$[[Y, X], Y](x_1) = \frac{2\alpha}{\lambda} X(x_1).$$

Then we have:

- The origin is an abnormal singular arc reduced to a point and the subarc of the line $x_2 = O$ is hyperbolic in $x_1 > 0$ and the subarc is elliptic if $x_1 < 0$.
- The singular control along the line $x_2 = 0$ is given by: $u = u_e$ and is constant and strictly admissible if $u_e \in]-1, +1[$.
- The collinear set is given by the parabola: $x_1 = \frac{\alpha x_2^2}{\lambda}$.
- The clock form is: $\omega = \frac{\mathrm{d}x_1}{(-\lambda x_1 + \alpha x_2^2)}$.

Moreover for every constant control $u = \varepsilon,\ \varepsilon = \pm 1$, the extremal system can be integrated.

One can construct a case study taking as terminal manifold N a circle centered at O=$(0, 0)$, with radius d intersecting the singular line at $(\pm d, 0)$. The time minimal synthesis outside the disk and near the two points $(\pm d, 0)$ can be directly deduced from the classification of [5], thanks to the curvature of the terminal manifold in the chosen normal coordinates. It is represented on Fig.2.2 and we have:

- *Top:* $(-d,0)$ lifts into a fold elliptic point. The singular line is time maximizing. The optimal policy is $\sigma_+\sigma_-$ or $\sigma_-\sigma_+$ using the clock form and we have represented the two strata of the switching locus: $W = W_- \cup W_+$ and there exists a cut locus C. The three curves of the stratification are ramifying at $(-d,0)$.
- *Bottom:* $(d,0)$ lifts into an hyperbolic fold point and the time minimal synthesis is of the form: $\sigma_-\sigma_s$ or $\sigma_+\sigma_s$.

To construct the complete synthesis one must glue the two cases along the exterior of the circle and fill the interior of the disk.

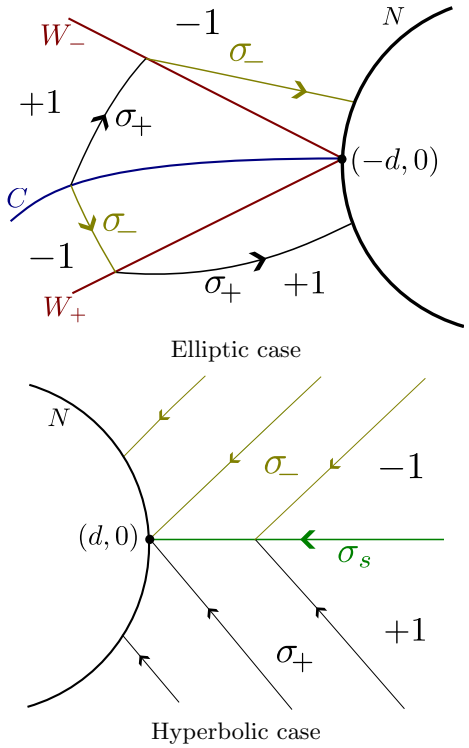


Fig. 2.2: $2d$–syntheses near $(\pm d,0)$ outside the disk.

To simplify the computations, we have assume that $u_e = 0$. The synthesis is represented on Fig.2.3.

Note that the singular line prolongated onto a cut locus terminating at $(d,0)$. In the non symmetric case $u_e \neq 0$, the cut locus persists but is not coinciding with this segment.

In this synthesis we assume that the two points $(\pm d,0)$ lift into fold points. But clearly we can obtain more general cases unfolding the syntheses with a parameter w by taking the system

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = -\lambda x_1 + w x_2 + \alpha x_2^2, \frac{\mathrm{d}x_2}{\mathrm{d}t} = (u - u_e),$$

where w is a constant.

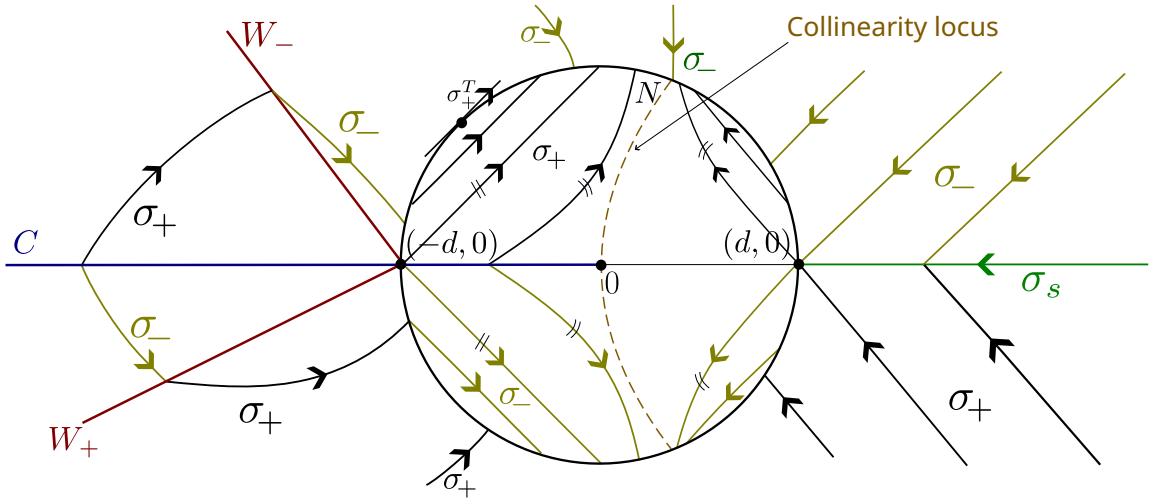


Fig. 2.3: Gluing hyperbolic and elliptic case with N being a circle; the symmetric case $u_e = 0$.

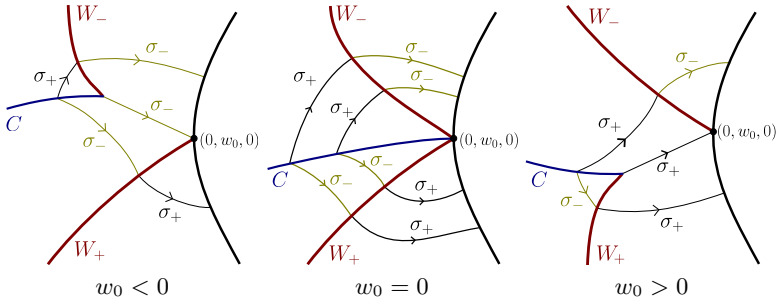


Fig. 2.4: Unfolding with parameter w_0 in the elliptic case.

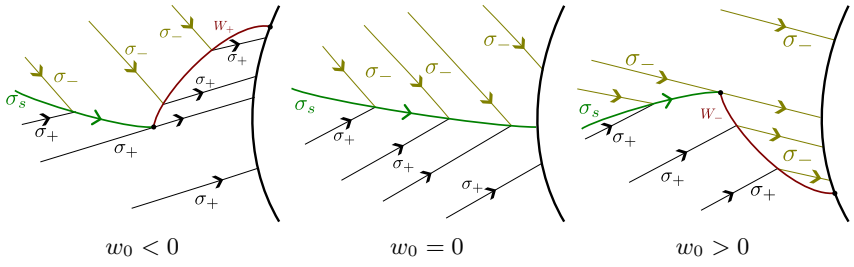


Fig. 2.5: Unfolding with parameter w_0 in the hyperbolic case.

This leads to unfold the synthesis as represented on figs. 2.4-2.5. Note that the sign of w is not relevant in the pictures since one can change u into $-u$ in the computations.

The switching locus W can be evaluated expanding the switching function, where the expansions are described in [5] and are in any case of order at most 2.

2.3.3 Computations on the $2d$–model

In this section we present direct computations on the $2d$-model vs the use of the semi-normal form. To simplify the notations we note (x, y) the $2d$-coordinates so that one has:

$$X = (x(r_1 + a_{11}x + a_{12}y), y(r_2 + a_{21}x + a_{22}y))^\intercal,$$
$$Y = (x\varepsilon_1, y\varepsilon_2)^\intercal.$$

Using ln–coordinates it takes the form:

$$X = ((r_1 + a_{11}e^x + a_{12}e^y), (r_2 + a_{21}e^x + a_{22}e^y))^\intercal,$$
$$Y = (\varepsilon_1, \varepsilon_2)^T.$$

Lie brackets are invariant and can computed in such coordinates which simplify the calculations since the vector field Y becomes constant.

Moreover one can impose in the class two geometric normalizations to clarify the analysis.

Normalizations

- One can suppose that the persistent equilibrium is $\Omega = (1, 1)$.
- One can assume that the persistent singular locus is the line: $y = x$.

This leads respectively to:

$$r_1 = -(a_{11} + a_{12)}, r_2 = -(a_{21} + a_{22}), \tag{2.9}$$

and

$$\varepsilon_1(\varepsilon_2 a_{11} - \varepsilon_1 a_{21}) = \varepsilon_2(\varepsilon_1 a_{22} - \varepsilon_2 a_{12}). \tag{2.10}$$

Lie brackets are given by:

$$[X, Y] = (x(\varepsilon_1 a_{11}x + \varepsilon_2 a_{12}y), y(\varepsilon_1 a_{21}x + \varepsilon_2 a_{22}y))^\intercal,$$
$$[[Y, X], Y] = (-x(\varepsilon_1^2 a_{11}x + \varepsilon_2^2 a_{12}y), -y(\varepsilon_1^2 a_{21}x + \varepsilon_2^2 a_{22}y))^\intercal,$$

the Lie bracket $[[X, Y], X]$ is more complex and takes in ln–coordinates the form:

$[[Y, X], X] = ((\varepsilon_1 a_{11}e^x(r_1 + a_{12}e^y) + \varepsilon_2 a_{12}e^y(r_2 + a_{21}e^x) - a_{11}e^x\varepsilon_2 a_{12}e^y - a_{12}e^y\varepsilon_1 a_{21}e^x), (\varepsilon_1 a_{21}e^x(r_1 + a_{12}e^y) + \varepsilon_2 a_{22}e^y(r_2 + a_{21}e^x) - a_{21}e^x\varepsilon_2 a_{12}e^y - a_{22}e^y\varepsilon_1 a_{21}e^x))^\intercal$.

One introduces the following determinants:

$$D = \det(Y, [[Y, X], Y]),$$
$$D' = \det(Y, [[Y, X], X]), \qquad D'' = \det(Y, X).$$

The generalized Legendre-Clebsch condition holds if along the singular line $y = x$,

$$D = xy[\varepsilon_1^2 x(\varepsilon_2 a_{11} - \varepsilon_1 a_{21}) + \varepsilon_2^2 y(\varepsilon_2 a_{12} - \varepsilon_1 a_{22})]$$

is non zero.

This gives restricting to $y = x$,

$$\frac{D}{xy} = xC,$$
$$C = \varepsilon_1^2(\varepsilon_2 a_{11} - \varepsilon_1 a_{21}) + \varepsilon_2^2(\varepsilon_2 a_{12} - \varepsilon_1 a_{22}) \neq 0.$$

Using the normalization condition (2.10) we get the condition

$$(\varepsilon_1\varepsilon_2 - \varepsilon_2^2)(\varepsilon_1 a_{22} - \varepsilon_2 a_{12}) \neq 0.$$

The singular control along the singular line $y = x$ is given by:

$$u_s = -\frac{D'_{|y=x}}{D_{|y=x}}.$$

Computing D' restricted to $y = x$ leads to introduce the coefficients:
$A = \varepsilon_1\varepsilon_2 a_{11} r_1 + \varepsilon_2^2 a_{12} r_2 - \varepsilon_1^2 a_{21} r_1 - \varepsilon_1\varepsilon_2 a_{22} r_2$,
$B = \varepsilon_1\varepsilon_2 a_{11} a_{12} + \varepsilon_2^2 a_{12} a_{21} - \varepsilon_2^2 a_{11} a_{12} - \varepsilon_1\varepsilon_2 a_{12} a_{21} - \varepsilon_1^2 a_{21} a_{22} - \varepsilon_1\varepsilon_2 a_{21} a_{22} + \varepsilon_1\varepsilon_2 a_{21} a_{12} + \varepsilon_1^2 a_{22} a_{21}$.
Hence the first component (projecting on the $x-axis$) of $-u_s Y$ restricting to the singular line $y = x$ takes the form

$$-\frac{(A + Bx)}{C}\varepsilon_1 x.$$

It has to vanishes at $x = 1$, so that $B = -A$. The derivative at $x = 1$ is $\frac{-\varepsilon_1(A+2B)}{C} = \frac{\varepsilon_1 A}{C}$.

Similarly at $\Omega = (1, 1)$, X has to vanishes, which corresponds to (2.9) and the derivative at $x = 1$ is $-r_1$.

Hence the dynamics along the singular line at $x = 1$ is regular if

$$-r_1 + \varepsilon_1 \frac{A}{C} \neq 0. \tag{2.11}$$

Note that we can reverse the orientation on the singular line changing in the same category X into $-X$.

In particular one deduces the following:

Theorem 1. *Under regularity conditions previously described, the singular flow along the singular line belongs to the one dimensional Lotka–Volterra form:* $\frac{\mathrm{d}x}{\mathrm{d}t} = x(r + ax)$ *and at the persistent equilibrium point the eigenvalue of the linearized dynamics is given by* $-r_1 + \varepsilon_1 \frac{A}{C}$.

2.3.4 Conclusion

Our study shows the main features to compute time minimal syntheses in different neighborhood of the origin and with different terminal manifolds. The main singularity is the interaction between the collinearity and the singular loci. We have introduced a semi-normal form with four homotopy parameters describing the main features of the geometric construction. Different cases can be analyzed gluing different syntheses. In particular the detailed computations of Section 2.3.3 show the role of the singular locus to extend the synthesis for large times.

2.4 From 2*d*–Case to 3*d*–Case and Numerical Simulations

2.4.1 The geometric frame

In this section we consider a $3d$ controlled Lotka–Volterra dynamics of the form

$$\frac{\mathrm{d}x}{\mathrm{d}t}(t) = X(x(t)) + \sum_{i=1}^{2} u_i(t) Y_i(x(t)),\ x = (x_1, x_2, x_3)^\intercal \in K \coloneqq \mathbb{R}^3_+, \tag{2.12}$$

where x_1 is the infected population and $u = (u_1, u_2),\ 0 \leq u_1 \leq 1,\ 0 \leq u_2 \leq 1 + \varepsilon,\ \varepsilon > 0$.
In this control system,

- X stands for the non controlled Lotka-Volterra dynamics given by $X = \mathrm{diag}x(Ax + r)$,
- $Y_1 = \mathrm{diag}x\ \epsilon$, ϵ is the constant sensitivity vector associated to a probiotics $\epsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3)^\intercal$, $\varepsilon_i \geq 0,\ i = 1, 2, 3$.
- $Y_2 = \mathrm{diag}x\ \epsilon$', ϵ' is the constant sensitivity vector associated to an antibiotic $\epsilon' = (\varepsilon'_1, \varepsilon'_2, \varepsilon'_3)^\intercal$, $\varepsilon'_i \leq 0,\ i = 1, 2, 3$.

Our aim is to reduce the x_i-population using the following protocol:

- Prior to infection use the probiotic to reinforce the microbiote.
- Having detected a given level of infection, reach in minimum time a forced equilibrium associated to a given level of infection. Moreover this level has to be stabilisable.

2.4.2 A rough classification of Lotka–Volterra dynamics

Definition 3. *We restrict to the case $r_i > 0,\ i = 1, 2, 3$ so that the origin O is a repeller and there exist axial equilibria $e_1 = (1, 0, 0)^\intercal$, $e_2 = (0, 1, 0)^\intercal$ and $e_3 = (0, 0, 1)^\intercal$. The system is called totally competitive if for all $1 \leq i, j \leq 3,\ a_{ij} < 0$. Additional equilibria may exist in the cone K and are denoted respectively : interior equilibrium Ω and $E_{jk},\ j < k$ related to extinction of species $i \neq j, k$.*

One has the following result from [1].

Proposition 4. *In the totally competitive case, there exists an unique Lipschitz invariant manifold Π that attracts $K \setminus \{0\}$ and every trajectory in $K \setminus \{0\}$ is asymptotic to one in Π. The manifold Π is homeomorphic to the closed unit simplex $S = \{x \geq 0,\ x_1 + x_2 + x_3 = 1\}$ under radial projection. Moreover the boundary of the basin of repulsion of the origin coincides with Π.*

Proposition 5. *The system obtained by adding a probiotic vector fields to a totally competitive system is also totally competitive. The same holds for antibiotic provided $|u_2|$ is small enough.*

2.4.3 Construction of the carrying simplex

The following is crucial for numeric computation of Π. We denote by $\varphi_t,\ t \geq 0$ the positive semi–flow generated by the Lotka–Volterra dynamics and M_t is the image by φ_t of the triangle S whose vertex are the three axial equilibria.

Proposition 6. *In the totally competitive case the sequence of surfaces (with corners) M_t converges uniformly to Π as t tends to $+\infty$.*

Remark 1. The interesting point is in a more general context to evaluate the boundary of the basin of repulsion of the origin.

2.4.4 Collinear set and singular dynamics

Consider the single–input dynamics

$$\frac{\mathrm{d}x}{\mathrm{d}t}(t) = X(x(t)) + u(t)Y(x(t)), \quad 0 \leq u \leq 1, \tag{2.13}$$

where Y is one of the vector field Y_1, Y_2 associated to a probiotic or an antibiotic agent.

Collinear set

It is defined by $C = \{x_e : \ \exists u_e \text{ constant such that } X(x_e) + u_e Y(x_e) = 0\}$ and x_e is a forced equilibrium. The control u_e is said feasible if $0 \leq u_e \leq 1$.

Singular dynamics

In the $3d$ case, the singular control can be computed as a feedback [3]. This is given by the following proposition.

Proposition 7. *Consider the single–input control system* (2.13) *in* $\mathbb{R}^3$. *Introduce the following determinants :*

- $D = \det(Y, [Y, X], [[Y, X], Y])$,
- $D' = \det(Y, [Y, X], [[Y, X], X])$,
- $D'' = \det(Y, [Y, X], X)$.

Then the singular control with minimal order is given by the feedback:

$$u_s(x) = -\frac{D'(x)}{D(x)},$$

so that the singular dynamics is defined by :

$$\frac{\mathrm{d}x}{\mathrm{d}t} = X_s(x) = X(x) - \frac{D'(x)}{D(x)}\, Y(x).$$

The sets $D'' = 0$, $DD'' > 0$ and $DD'' < 0$ are foliated respectively by exceptional, hyperbolic and elliptic arcs and are invariant for the integral curves of the vector field $X_s(x)$.

Computations

One has the following expressions of D, D', D'' in the original coordinates :

$$\begin{aligned} D(x)/x_1x_2x_3 = \\ &\left(\varepsilon_1^2 x_1 a_{21} + \varepsilon_1\left(\varepsilon_2\left(x_2a_{22} - x_1a_{11}\right) + \varepsilon_3 x_3 a_{23}\right) - \varepsilon_2\left(\varepsilon_2 x_2 a_{12} + \varepsilon_3 x_3 a_{13}\right)\right) \\ &\left(\varepsilon_1^2 x_1 a_{31} + \varepsilon_2^2 x_2 a_{32} + \varepsilon_3^2 x_3 a_{33}\right) + \left(\varepsilon_1^2 x_1 a_{11} + \varepsilon_2^2 x_2 a_{12} + \varepsilon_3^2 x_3 a_{13}\right)\left(\varepsilon_2^2 x_2 a_{32}\right. \\ &\left. + \varepsilon_3\varepsilon_2\left(x_3a_{33} - x_2a_{22}\right) - \varepsilon_3^2 x_3 a_{23} + \varepsilon_1 x_1\left(\varepsilon_2 a_{31} - \varepsilon_3 a_{21}\right)\right) \\ &- \left(\varepsilon_1^2 x_1 a_{21} + \varepsilon_2^2 x_2 a_{22} + \varepsilon_3^2 x_3 a_{23}\right)\left(\varepsilon_1^2 x_1 a_{31} + \varepsilon_1\left(\varepsilon_2 x_2 a_{32} + \varepsilon_3\left(x_3 a_{33} - x_1 a_{11}\right)\right)\right. \\ &\left. - \varepsilon_3\left(\varepsilon_2 x_2 a_{12} + \varepsilon_3 x_3 a_{13}\right)\right), \end{aligned}$$

$$
\begin{aligned}
D'(x)/x_1x_2x_3 = \\
&\left(-\varepsilon_1^2 x_1\ a_{21}+\varepsilon_1\left(\varepsilon_2\left(x_1a_{11}-x_2a_{22}\right)-\varepsilon_3x_3a_{23}\right)+\varepsilon_2\left(\varepsilon_2x_2a_{12}+\varepsilon_3x_3a_{13}\right)\right)\\
&\left(\varepsilon_2x_2\left(x_1a_{12}a_{31}-a_{32}\left(x_1a_{21}+x_3\left(a_{23}-a_{33}\right)+r_2\right)\right)-\varepsilon_1x_1\left(r_1a_{31}+x_3\left(a_{13}-a_{33}\right)a_{31}\right.\right.\\
&\left.\left.+x_2\left(a_{12}a_{31}-a_{21}a_{32}\right)\right)+\varepsilon_3x_3\left(-r_3a_{33}+x_1a_{31}\left(a_{13}-a_{33}\right)+x_2a_{32}\left(a_{23}-a_{33}\right)\right)\right)\\
&+\left(\varepsilon_2^2\left(-x_2\right)a_{32}+\varepsilon_3\varepsilon_2\left(x_2a_{22}-x_3a_{33}\right)+\varepsilon_3^2x_3a_{23}+\varepsilon_1x_1\left(\varepsilon_3a_{21}-\varepsilon_2a_{31}\right)\right)\\
&\left(-\varepsilon_1x_1\left(r_1a_{11}+x_2a_{12}\left(a_{11}-a_{21}\right)+x_3a_{13}\left(a_{11}-a_{31}\right)\right)+\varepsilon_2x_2\left(x_3a_{13}a_{32}\right.\right.\\
&\left.\left.-a_{12}\left(x_1\left(a_{21}-a_{11}\right)+x_3a_{23}+r_2\right)\right)-\varepsilon_3x_3(a_{13}\left(x_1\left(a_{31}-a_{11}\right)+x_2a_{32}+r_3\right)\right.\\
&\left.-x_2a_{12}a_{23})\right)-\left(-\varepsilon_1^2x_1a_{31}+\varepsilon_1\left(\varepsilon_3\left(x_1a_{11}-x_3a_{33}\right)-\varepsilon_2x_2a_{32}\right)+\varepsilon_3(\varepsilon_2x_2a_{12}\right.\\
&\left.+\varepsilon_3x_3a_{13})\right)\left(\varepsilon_1x_1\left(x_3a_{23}a_{31}-a_{21}\left(x_3a_{13}+x_2\left(a_{12}-a_{22}\right)+r_1\right)\right)+\varepsilon_2x_2\right.\\
&\left.\left(-r_2a_{22}+x_1a_{21}\left(a_{12}-a_{22}\right)+x_3a_{23}\left(a_{32}-a_{22}\right)\right)+\varepsilon_3x_3(x_1a_{13}a_{21}-a_{23}(x_1a_{31}\right.\right.\\
&\left.\left.+x_2\left(a_{32}-a_{22}\right)+r_3))\right),
\end{aligned}
$$

$$
\begin{aligned}
D''(x)/x_1x_2x_3 = \\
&\left(-\varepsilon_1^2x_1a_{21}+\varepsilon_1\left(\varepsilon_2\left(x_1a_{11}-x_2a_{22}\right)-\varepsilon_3x_3a_{23}\right)+\varepsilon_2\left(\varepsilon_2x_2a_{12}+\varepsilon_3x_3a_{13}\right)\right)\\
&\left(x_1a_{31}+x_2a_{32}+x_3a_{33}+r_3\right)+\left(-\varepsilon_2^2x_2a_{32}+\varepsilon_3\varepsilon_2\left(x_2a_{22}-x_3a_{33}\right)+\varepsilon_3^2x_3a_{23}\right.\\
&\left.+\varepsilon_1x_1\left(\varepsilon_3a_{21}-\varepsilon_2a_{31}\right)\right)\left(x_1a_{11}+x_2a_{12}+x_3a_{13}+r_1\right)+\left(\varepsilon_1^2x_1a_{31}+\varepsilon_1\left(\varepsilon_2x_2a_{32}\right.\right.\\
&\left.\left.+\varepsilon_3\left(x_3a_{33}-x_1a_{11}\right)\right)-\varepsilon_3\left(\varepsilon_2x_2a_{12}+\varepsilon_3x_3a_{13}\right)\right)\left(x_1a_{21}+x_2a_{22}+x_3a_{23}+r_2\right).
\end{aligned}
$$

Remark 2. Similar computations hold in the bi-input case with $Y(x) = \text{span}\{Y_2(x), Y_2(x)\}$. Note that $Y(x)$ is integrable since $[Y_1, Y_2] = 0$. Such computations are necessary to identify the singular dynamics, which have to be avoided because of strong accessibility problems, see [8].

2.4.5 Computational path as a medical protocol

- Classify the eight equilibria computing the spectrum of the Jacobian matrix evaluated at these points.
- The objective function to minimize is $x \to x_1(T)$, which can be reformulated as a problem of reaching the surface $x_1(t_f) \leq x_1^{\min}$ in minimum time where $x_1^{\min}$ is a given threshold, representing the level from which the detection of the infection is possible.

We shall take into account various constraints on the system in the framework of sampled-data control :

- *Infection constraints.* The infected population has to be lower than a given threshold $x_1^{\max}$, representing the maximum level of infection.
- *Logistic constraints.* The therapy consists of delivering treatment on specific times intervals $[t_i, t_{i+1}]$, where the duration $t_{i+1} - t_i$ is bigger than an interpulse $t_{i+1} - t_i \geq I_m$ e.g. $I_m = 1$ day. The controls $u_i, i = 1, 2$ are constant on each interpulse and are bounded by some constants m_i.
- At final time T of the caring therapy (e.g. $T = 40$ days) it is required that the final point $x(T)$ is in a stability domain of a forced equilibrium point denoted x_{ef} associated to an admissible control.

- Additional L^2–constraints can be added to take into account the cost or the total amount of available drug e.g. antibiotic.

2.4.6 Numerical simulations

The previous geometric analysis leads to design direct numerical schemes or semi-direct scheme based on NMPC method (Nonlinear Model Predictive Control) [18].

We present preliminary results on the $3d$ Lotka Volterra system (2.12) with $X(x) = \mathrm{diag}x(r - Ax)$, $r = (1,1,1)^\intercal$, $A = \begin{pmatrix} 1 & \alpha & \beta \\ \beta & 1 & \alpha \\ \alpha & \beta & 1 \end{pmatrix}$ and in the case $\alpha + \beta > 2$, $\alpha < 1$. This implies that the carrying simplex is the plane $x + y + z = 3/(1 + \alpha + \beta)$ and the interior equilibrium exists and is an unstable focus (see Table 2.1).

Equilibria and stability

The spectrums of the Jacobians $\frac{\partial X}{\partial x}$ evaluated at the eight free equilbria are given in Table 2.1.

Free equilibria $X(x_e) = 0$	Spectrum spec $\left(\frac{\partial X}{\partial x}_{\mid x=x_e}\right)$
$(0,0,0)$	$\{1,1,1\}$
$(0,0,1)$, $(1,0,0)$, $(0,1,0)$	$\{-1, 1-\alpha, 1-\beta\}$
$(\frac{1}{\alpha+\beta+1}, \frac{1}{\alpha+\beta+1}, \frac{1}{\alpha+\beta+1})$	$\left\{-1, \frac{\alpha+\beta-2-i\sqrt{3}\lvert\alpha-\beta\rvert}{2(\alpha+\beta+1)}, \frac{\alpha+\beta-2+i\sqrt{3}\lvert\alpha-\beta\rvert}{2(\alpha+\beta+1)}\right\}$
$(0, \frac{\alpha-1}{\alpha\beta-1}, \frac{\beta-1}{\alpha\beta-1})$, $(\frac{\beta-1}{\alpha\beta-1}, 0, \frac{\alpha-1}{\alpha\beta-1})$ $(\frac{\alpha-1}{\alpha\beta-1}, \frac{\beta-1}{\alpha\beta-1}, 0)$	$\left\{\frac{(\alpha-1)(\beta-1)}{\alpha\beta-1}, -1, \frac{-\alpha^2+\alpha\beta+\alpha-\beta^2+\beta-1}{\alpha\beta-1}\right\}$

Table 2.1: Spectrum of the Jacobian matrix of X evaluated at the eight equilibria of the May and Leonard model.

Optimization problem

To reduce the x_1 population in the biological frame, where stability of the final point $x_1(T)$ is required we proceed as follows.

1. Prior to the infection, we assume that the patient is in a healthy stable state represented as a stable equilibrium point (x_{20}, x_{30}) of a two-dimensional Lotka–Volterra system.
2. Suppose the patient got infected by a pathogen agent at time $t = 0$. At time $t > 0$ its state is represented as a $3d$ vector $(x_1(t), x_2(t), x_2(t))$ with $x(0) = (x_{10}, x_{20}, x_{30})$, $x_{10} \gg x_1^{\min} > 0$. This vector is governed by a $3d$ Lotka-Volterra system.

3. From $x(0)$, we accelerate the evolution of the state to the variety Π. This is formulated as a minimum time control problem $3d$ Lotka–Volterra system using probiotics only.
4. Finally, we reach, in minimum time using antibiotics, an healthy region $N : x_1(T) \leq k\, x_1^{\min}$, where $k < 1$ is a scaling factor. To ensure that the final point is in a stable healthy region, the stability can be obtained by a pole placement method [11].

Direct method

We illustrate our previous four steps protocol by computing a trajectory $x_{ref}(.)$ using the Bocop software [5] with $\alpha = 0.2$, $\beta = 2$.

The point $(x_2(0), x_3(0)) = (0.1, 0.1)$ corresponds to a forced stable equilibrium of the $2d$ Lotka–Volterra system : it is a point on the collinearity set associated to the control $u_e = 0.75$ and for suitable values of $\varepsilon_1, \varepsilon_2$ and, where the eigenvalues of the Jacobian matrix

$$\frac{\partial}{\partial x}(X(x) + uY(x))_{|x=(x_2(0),x_3(0)),u=u_e}$$

have strictly negative real parts.

At time $t = 0$, a pathogen agent is measured with $x_1(0) = 0.1 =: x_1^{\min}$. Then in Phase 1 we accelerate the evolution of the state (x_1, x_2, x_3), governed by the $3d$ Lotka–Volterra system

$$\begin{pmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \end{pmatrix} = \mathrm{diag}x \left(\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 1 & \alpha & \beta \\ \beta & 1 & \alpha \\ \alpha & \beta & 1 \end{pmatrix} x(t) + u(t) \begin{pmatrix} 1.5 \\ 0.4 \\ 1 \end{pmatrix} \right),$$

towards the flat carrying simplex $\Pi : x + y + z = 1/(1 + \alpha + \beta)$ by using probiotics only.

When we are close enough to Π, Phase 2 is initiated and we minimize the time to reach the healthy region $x_1(T) < x_1^{\min}/2$, using antibiotics only with $Y(x) = \mathrm{diag}x\,(-0.5, -1.4, -1)^\intercal$. The trajectory resulting from these two phases is displayed in Fig.2.6.

NMPC tracking

A predictive controller model is constructed through the minimization of a cost function involving the state and control inputs over a finite time horizon, with consideration for constraints on both inputs and states. The feedback control input has a piecewise affine structure of the form $u_f = -K\, x + \Lambda$, which can be easily deployed in microcontrollers for fast processes.

First, we start by constructing a discrete-time approximation of the bi-input system (2.12) using a Tustin bilinear transformation with sampling period $\tau > 0$. Let $x_{ref}(\cdot)$ be a reference trajectory. Define the cost

$$J(x, u_f) \coloneqq \sum_{k=1}^{\eta} \|e(k)\|_2^2 + \|\Delta u_f(k)\|_2^2 + \kappa e(k)^\intercal \Delta u_f(k),$$

where η is the horizon, $e(k) = x_{ref}(k) - x(k)$ is the predicted error, $\Delta u_f(k) = u_f(k) - u_f(k-1)$ is the incremental feedback control input and $\kappa > 0$ is a parameter.
The feedback control is computed by minimizing the cost J subject to the discrete dynamic constraint obtained with the Tustin transformation and such that each control component is in $[0, 1]$.

Plugging the feedback control in the control dynamics (2.12) yields a closed loop system, robust with respect to perturbations and uncertainties on the state x.

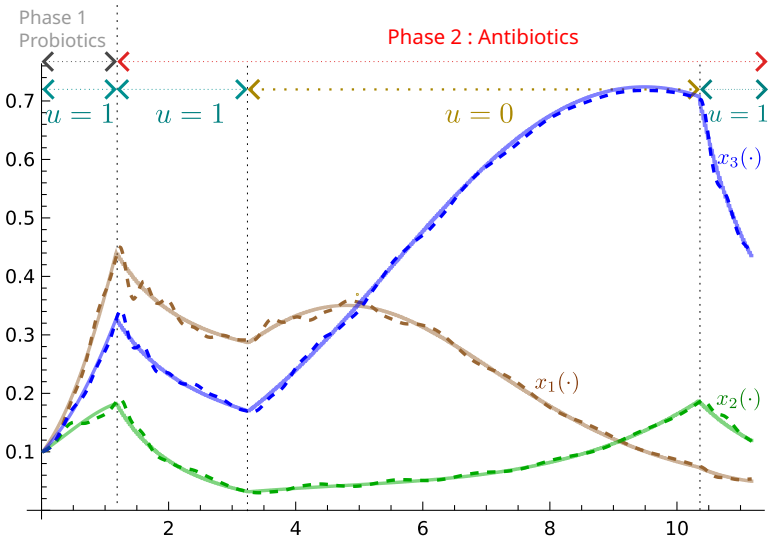


Fig. 2.6: *(Continuous curves)* Reference trajectory x_{ref} computed with a direct method on a $3d$ totally competitive Lotka–Volterra model following the four steps protocol given in section 2.4.5. *(Dashed curves)* Tracking trajectory obtained via a NMPC tracking method on the reference trajectory.

Numerical results.

Take the trajectory $x_{ref}(\cdot)$, computed with the direct method, as a reference trajectory. The feedback control is computed with the following NMPC parameters : $\tau = 0.005$, $\eta = 5$, $\kappa = 0.002$ and the parameters of the dynamics (2.12) are $\epsilon_1 = (1.5, 0.4, 1.0)^\intercal$, $\epsilon_2 = (0.5, -1.4, -1.0)^\intercal$, $\alpha = 0.2$ and $\beta = 2$. Time evolution of the trajectory of the resulting closed loop system is displayed in Figure 2.6 as dashed curves showing the ability to track the reference signal.
Note that we can accelerate the recovery by computing the feedback control with a predicted error of the form $e(k) = x_{ref}(k + p) - x(k)$ for some integer $p > 1$.

2.5 Conclusion

This article presents briefly a combination of geometric and numerical methods to analyze the problem of reduction of a complex microbiote by a pathogenic agent. It leads to robust optimal control schemes to quantify the effect of different medical protocols. Computations are presented for the $2d$-system and for $3d$-totally competitive Lotka–Volterra models.

References

1. S. Baigent, Geometry of carrying simplices of 3-species competitive Lotka-Volterra systems, *Nonlinearity*, **26**, (2013) 1001–1029.

2. Team Commands, Inria Saclay, BOCOP: an open source toolbox for optimal control, http://bocop.org, 2017.
3. B. Bonnard, M. Chyba, The role of singular trajectories in control theory, Springer Verlag, New York, 2003, 357 pages.
4. B. Bonnard, G. Launay, M. Pelletier, Classification générique de synthèses temps minimales avec cible de codimension un et applications, *Annales de l'I.H.P. Analyse non linéaire*, **14** no.1 (1997), 55–102.
5. B. Bonnard, J. Rouot, Towards Geometric Time Minimal Control without Legendre Condition and with Multiple Singular Extremals for Chemical Networks, *Advances in Nonlinear Biological Systems, Modeling and Optimal Control, AIMS on applied Maths*, **11** (2021), 1–34.
6. B. Bonnard, J. Rouot, Optimal Control of the Controlled Lotka-Volterra Equations with Applications - The Permanent Case, *SIAM J. Appl. Dyn.*, **22** no. 4 (2023), 2761–2791.
7. B. Bonnard, J. Rouot, Feedback Classification and Optimal Control with Applications to the Controlled Lotka-Volterra Model, Preprint 2023: hal-03861565.
8. B. Bonnard, J. Rouot, C. Silva, Geometric Optimal Control of the Generalized Lotka-Volterra Model of the Intestinal Microbiome, Accepted for publication in OCAM (2024) : hal-03861565.
9. V.G. Boltyanskii, Sufficient conditions for optimality and the justification of the dynamic programming method, *SIAM J. Control*, **4** (1966), 326–361.
10. P. Brunovský, Existence of regular synthesis for general control problems, *J. Differential Equations*, **38** no. 3 (1980), 317–343.
11. H. Hermes, On the Synthesis of a Stabilizing Feedback Control via Lie Algebraic Methods, *SIAM J. Control Optim.*, **18**, no. 4, (1980) 352–361.
12. J. Hofbauer, J. W.-H. So, Multiple limit cycles for three dimensional Lotka-Volterra equations, *Applied Math. Lett.*, **7**, no. 6, (1994) 65–70.
13. E.W. Jones, P. S. Clarcke, J. M. Carslon, Navigation of outcome in a generalized Lotka–Volterra model of the microbiome, *Advances in Nonlinear Biological Systems, Modeling and Optimal Control, AIMS on applied Maths*, **11** (2021), 97–117.
14. I. Kupka, Geometric theory of extremals in optimal control problems. I. The fold and Maxwell case, *Trans. Amer. Math. Soc.*, **299** no.1 (1987), 225–243.
15. G. Launay, M. Pelletier, The generic local structure of time-optimal synthesis with a target of codimension one in dimension greater than two, *Journal of Dynamical and Control Systems*, **3**, no. 165 (1997).
16. S. Nikitin, Piecewise-Constant Stabilization, *SIAM J Control Optim.*, **37**, no. 3, (1999) 911–933.
17. L.S. Pontryagin, V.G. Boltyanskii, R.V. Gamkrelidze, E.F. Mishchenko, The mathematical theory of optimal processes, *Oxford, Pergamon Press*, 1964, 362 pages.
18. J. Rawlings, D. Mayne, M. Diehl, Model Predictive Control: Theory, Computation, and Design, Santa Barbara, CA: Nob Hill, 2017, 819 pages.
19. H. Schättler, U. Ledzewicz, Optimal control for mathematical models of cancer therapies. An application of geometric methods, *Interdisciplinary Applied Mathematics*, **42**. Springer, New York, 2015, 496 pages.
20. R.R. Stein, V. Bucci, N.C. Toussaint, C.G. Buffie, G. Rätsch, E.G. Pamer, et al., Ecological modelling from time-series inference: insight into dynamics and stability of intestinal microbiota, *PLos Comp. Biology*, **9** no. 12 (2013).
21. V. Volterra, Leçons sur la théorie mathématique de la lutte pour la vie, Les Grands Classiques Gauthier-Villars. Éditions Jacques Gabay, Sceaux, 1990, 215 pages.

3

Zermelo Navigation on the Sphere with Revolution Metrics

Bernard Bonnard[1], Olivier Cots[2], Yannick Privat[34], and Emmanuel Trélat[5]

[1] Institut Mathématique de Bourgogne and INRIA Sophia Antipolis (bernard.bonnard@u-bourgogne.fr).
[2] Institut de Recherche en Informatique de Toulouse, UMR CNRS 5505, Université de Toulouse, INP-ENSEEIHT, France (olivier.cots@irit.fr).
[3] Université de Lorraine, CNRS, Institut Elie Cartan de Lorraine, Inria, BP 70239 54506 Vandœuvre-lès-Nancy Cedex, France (yannick.privat@univ-lorraine.fr).
[4] Institut Universitaire de France (IUF).
[5] Sorbonne Université, CNRS, Université Paris Cité, Inria, Laboratoire Jacques-Louis Lions (LJLL), F-75005 Paris, France (emmanuel.trelat@sorbonne-universite.fr).

This article is dedicated to I. Kupka.

Summary. In this article motivated by physical applications, the Zermelo navigation problem on the two-dimensional sphere with a revolution metric is analyzed within the framework of minimal time optimal control. The Pontryagin maximum principle is used to compute extremal curves and a neat geometric frame is introduced using the Carathéodory-Zermelo-Goh transformation. Assuming that the current is of revolution, the geodesics are sorted according to a Morse-Reeb classification. We then illustrate the relevance of this classification using various examples from physics: the Lindblad equation in quantum control, the averaged Kepler case in space mechanics and the Landau-Lifshitz equation in ferromagnetism.

Keywords.

Zermelo navigation problem; Minimal time geometric control; Morse-Reeb classification.

3.1 Introduction

A Zermelo navigation problem on the two-dimensional sphere M with a revolution metric is defined by a pair (g, F_0) where g is a metric of revolution on M and F_0 is a smooth vector field on M called the *current*. Using the control framework [9], the problem can be formulated as a *minimal time transfer problem* between two points $q_0, q_1 \in M$ for the single-input control-affine system

$$\frac{\mathrm{d}q}{\mathrm{d}t}(t) = F_0(q(t)) + \sum_{i=1}^{2} u_i(t)\, F_i(q(t)), \tag{3.1}$$

where the control $u = (u_1, u_2)$ is subject to the constraint $\|u\|^2 = u_1^2 + u_2^2 \leq 1$ and $q = (r, \theta)$ are the polar coordinates for the metric of revolution $g = \mathrm{d}r^2 + m^2(r)\,\mathrm{d}\theta^2$ with $m(r) > 0$ (see [2]). The two smooth vector fields

$$F_1 = \frac{\partial}{\partial r}, \quad F_2 = \frac{1}{m(r)}\frac{\partial}{\partial \theta}$$

form an orthonormal frame, and the current F_0 is

$$F_0(q) = \mu_1(q)\frac{\partial}{\partial r} + \mu_2(q)\frac{\partial}{\partial \theta}$$

where $\mu_1(q)$ is the *vertical* component and $\mu_2(q)$ is the *horizontal* component. The current is said to be of revolution if μ_1 and μ_2 do not depend on θ. The surface M is the (closure of) the union of two domains: the region of weak current where $\|F_0\|_g < 1$ and the region of strong current where $\|F_0\|_g > 1$.

The above problem is a generalization of the historical problem of the *quickest nautical path* introduced and studied by Carathéodory and Zermelo in [10, 21] where one can find a complete study in the case of a linear current, the metric being the Euclidean metric.

Borrowing the point of view of the historical problem, a neat geometric frame was introduced in [21], parametrizing the curves by the *heading angle* α of the ship, extending the control system to a single-input control-affine system

$$\frac{\mathrm{d}\tilde{q}}{\mathrm{d}t}(t) = X(\tilde{q}(t)) + v(t)\,Y(\tilde{q}(t))$$

where $\tilde{q} = (r, \theta, \alpha)$ and v is the time derivative of α. This transform, referred to as the *Carathéodory-Zermelo-Goh* transformation, leads to analyze the problem using iterated *Lie brackets* of the vector fields X and Y.

In this article we perform the analysis in the case of revolution. Thanks to *Clairaut condition*, the extremal dynamics can be integrated and studied using an extension of the *Morse-Reeb classification* for 2D Hamiltonian system [3]. Preliminary results where obtained in [5] in the case of an horizontal current and are here extended to the general case. Extremal curves are sorted by distinguishing r-periodic and r-aperiodic curves.

Another contribution of this article is to analyze three case studies. The first is the so-called *averaged Kepler case*, appearing also in space mechanics [3]. Geometrically it amounts to analyzing the effect of the curvature on the historical example. It is a case of revolution, with horizontal current only. The second case comes from quantum control and is related to the control of the Lindblad equation. We propose a simplified dynamics model corresponding to a case of revolution with vertical current. The final study, based on [11], concerns the Landau-Lifshitz model for ellipsoidal ferromagnetic samples. We propose an alternative frame to study the controllability problem of the magnetic moment.

The article is organized in two sections. In Section 3.2, we recall the Pontryagin maximum principle [16] and we present the geometric tools to analyze the extremals. The Carathéodory-Zermelo-Goh transformation is introduced in details to classify the extremals with respect to the induced action of the feedback group. In the case of revolution, the Morse-Reeb classification is introduced to refine the classification of the extremals. It amounts roughly to extending the Liouville-Mineur-Arnold theorem [2]. Extremals are either r-periodic or r-aperiodic curves, in relationship with weak and strong current domains. Section 3.3 provides the details of the analysis in three case studies.

3.2 Pontryagin Maximum Principle and Geometric Analysis of the Hamiltonian Dynamics

3.2.1 Pontryagin maximum principle

For $i \in \{1,2\}$, let F_i be a smooth vector field on M; we denote by $H_i(q,p) = \langle p, F_i(q)\rangle$ the Hamiltonian lift, in local coordinates $z = (q,p)$ the coordinates on T^*M with $p = (p_r, p_\theta)$ (adjoint vector). The *pseudo-Hamiltonian* is the cost-extended Hamiltonian defined by

$$H(z,u) = H_0(z) + \sum_{i=1}^{2} u_i\, H_i(z) + p^0$$

where $p^0 \in \mathbb{R}$ is the dual variable of the cost. We define the maximized Hamiltonian by

$$M(z) = \max_{\|u\|\leq 1} H(z,u).$$

According to the Pontryagin maximum principle [16], any minimal (or maximal) time trajectory, solution of (Σ_u) on $[0\,, t_f]$, must be the projection onto M of an extremal, that is a quadruple $(q(\cdot), p(\cdot), p^0, u(\cdot))$, with $(p(\cdot), p^0) \neq (0,0)$, satisfying

$$\frac{\mathrm{d}q}{\mathrm{d}t}(t) = \frac{\partial H}{\partial p}(z(t),u(t)), \quad \frac{\mathrm{d}p}{\mathrm{d}t}(t) = -\frac{\partial H}{\partial q}(z(t),u(t)), \tag{3.2}$$

and the maximization condition

$$H(z(t),u(t)) = M(z(t)) \tag{3.3}$$

for almost every $t \in [0\,, t_f]$. Moreover, we have $M(z(t)) = 0$ for every $t \in [0\,, t_f]$. Furthermore, if the trajectory is minimal time then $p^0 \leq 0$ and if the trajectory is maximal time then $p^0 \geq 0$.

The projection onto M of an extremal is called a *geodesic*. Then, the Pontryagin maximum principle says that any minimal time trajectory must be a geodesic. Recall anyway that this is only a necessary condition for optimality and that, conversely, a geodesic may fail to be minimal time. A geodesic is said to be *strict* if it has a unique extremal lift, up to scaling. An extremal is said to be *normal* if if $p^0 \neq 0$ and *abnormal* (or exceptional) if $p^0 = 0$. In the normal case, it is said to be *hyperbolic* if $p^0 < 0$ and *elliptic* if $p^0 > 0$.

In the present situation, it follows from the maximization condition that:

- extremal controls are given by $u_i(z) = H_i(z)/\|p\|_g$, for $i = 1,2$, where

$$\|p\|_g = \big(H_1^2(z) + H_2^2(z)\big)^{1/2} = \left(p_r^2 + \frac{p_\theta^2}{m^2(r)}\right)^{1/2};$$

- the maximized Hamiltonian is $M(z) = H_0(z) + \|p\|_g + p^0$;
- any extremal is solution of the Hamiltonian system

$$\frac{\mathrm{d}q}{\mathrm{d}t}(t) = \frac{\partial M}{\partial p}(z(t)), \qquad \frac{\mathrm{d}p}{\mathrm{d}t}(t) = -\frac{\partial M}{\partial q}(z(t)).$$

3.2.2 Carathéodory-Zermelo-Goh transformation and geodesics parameterization

In their seminal study, Carathéodory and Zermelo introduced the heading angle to parameterize the geodesics [10], which amounts to using the Goh transformation in optimal control. Since for the geodesics one has $\|u\| = 1$, one can set $u = (\cos\alpha, \sin\alpha)$, α being the heading angle of the ship. Let $\tilde{q} = (q, \alpha)$ be the extended state and set

$$X(\tilde{q}) = F_0(q) + \cos\alpha\, F_1(q) + \sin\alpha\, F_2(q), \quad Y(\tilde{q}) = \frac{\partial}{\partial\alpha}.$$

This leads to augment (Σ_u) to the single-input control-affine system:

$$\frac{\mathrm{d}\tilde{q}}{\mathrm{d}t}(t) = X(\tilde{q}(t)) + v(t)\, Y(\tilde{q}(t)) \tag{3.4}$$

and the derivative of the heading angle $v(t) = \alpha'(t) \in \mathbb{R}$ is called the *accessory control*. Denoting $\tilde{z} = (\tilde{q}, \tilde{p})$, $\tilde{p} = (p, p_\alpha)$, we define the extended pseudo-Hamiltonian by

$$\widetilde{H}(\tilde{z}, v) = \langle \tilde{p}, X(\tilde{q}) + v\, Y(\tilde{q})\rangle + p^0.$$

By [4, Chapter 6], in this representation, geodesic curves become *singular trajectories* of (3.4).

Recall that the Lie bracket of two vector fields U, V is defined by

$$[U, V](\tilde{q}) = \frac{\partial U}{\partial\tilde{q}}(\tilde{q})\, V(\tilde{q}) - \frac{\partial V}{\partial\tilde{q}}(\tilde{q})\, U(\tilde{q})$$

and is related to the Poisson bracket by $\{H_U, H_V\}(\tilde{z}) = \mathrm{d}H_U(\tilde{z}) \cdot \vec{H}_V(\tilde{z})$ by the relation

$$\{H_U, H_V\}(\tilde{z}) = \langle \tilde{p}, [U, V](\tilde{q})\rangle,$$

where H_U, H_V are the Hamiltonian lifts of U and V. It is easy to check that

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial\widetilde{H}}{\partial v}\bigg|_{(\tilde{q},\tilde{p},v)} = \langle \tilde{p}, [Y, X](\tilde{q})\rangle,$$
$$\frac{\partial}{\partial v}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\frac{\partial\widetilde{H}}{\partial v}\bigg|_{(\tilde{q},\tilde{p},v)} = \langle \tilde{p}, [[Y, X], Y](\tilde{q})\rangle.$$

Proposition 1. *Defining*

$$D = \det(Y, [Y, X], [[Y, X], Y]),$$
$$D' = \det(Y, [Y, X], [[Y, X], X]),$$
$$D'' = \det(Y, [Y, X], X),$$

any extremal control v is given by the feedback

$$v(t) = v_s(\tilde{q}(t)) = -\frac{D'(\tilde{q}(t))}{D(q(t))} \tag{3.5}$$

and the geodesics are solutions of

$$\frac{\mathrm{d}\tilde{q}}{\mathrm{d}t}(t) = X(\tilde{q}(t)) + v_s(\tilde{q}(t))\, Y(\tilde{q}(t)) = X_s(\tilde{q}(t)). \tag{3.6}$$

Moreover:

- *hyperbolic geodesics are in the region where* $DD'' > 0$*;*
- *elliptic geodesics are in the region where* $DD'' < 0$*;*
- *abnormal (or exceptional) geodesics are in the region where* $D'' = 0$*.*

Proof. We refer to [4, Sec. 3.4]. A singular control-trajectory pair $(\tilde{q}, v)$ satisfies

$$H_Y(\tilde{z}) = \{H_Y, H_X\}(\tilde{z}) = 0$$
$$\{\{H_Y, H_X\}, H_X\}(\tilde{z}) + v\,\{\{H_Y, H_X\}, H_Y\}(\tilde{z}) = 0$$

and this leads to

$$0 = \langle \tilde{p}, Y(\tilde{q})\rangle = p_\alpha,$$
$$0 = \langle \tilde{p}, [Y, X](\tilde{q})\rangle,$$
$$0 = \langle \tilde{p}, [[Y, X], X](\tilde{q}) + v\,[[Y, X], Y](\tilde{q})\rangle.$$

Hence, since $\tilde{p} \in \mathbb{R}^3 \setminus \{0\}$, $\tilde{p}$ can be eliminated. Moreover, every geodesic is strict and $D(\tilde{q})$ is never vanishing. Hence, the geodesic control $v(\cdot)$ is given by (3.5). The geodesic classification follows.

3.2.3 Feedback pseudo-group G_f and singularity analysis

Given a pair (X, Y) of vector fields, the set of triples (φ, α, β), where φ is a local diffeomorphism and where $u = \alpha(x) + \beta(x)\,u'$, with $\beta \neq 0$, is a feedback, acts on (X, Y). This action induces the pseudo-feedback group G_f.

Theorem 1 ([4]). *Let* λ_s *be the mapping which yields for each pair* (X, Y) *the dynamics* (3.6)*. Then* λ_s *is a covariant mapping, i.e., the following diagram is commutative:*

$$\begin{array}{ccc} (X, Y) & \xrightarrow{\lambda_s} & X_s \\ {\scriptstyle G_f}\big\downarrow & & \big\downarrow{\scriptstyle G_f} \\ (X', Y') & \xrightarrow{\lambda_s} & X'_s \end{array}$$

Proof. The proof follows from straightforward computations on the determinants D, D'.

Definition 1. *We define the collinear set by* $\mathscr{C} = \{q \mid \exists \alpha,\ F_0(q) + \cos\alpha\, F_1(q) + \sin\alpha\, F_2(q) = 0\}$.

Proposition 2.

1. *The geodesic curves are immersed curves outside of the collinear set.*
2. *Only abnormal geodesics can be non-immersed curves when meeting the collinear set.*

Proof. This comes from the relation between the set $\{\|F_0\|_g = 1\}$ and the collinear set. Indeed take $q_0 \in \{\|F_0\|_g = 1\}$, then there exists α_0 such that:

$$F_0(q_0) + \cos\alpha_0 F_1(q_0) + \sin\alpha_0 F_2(q_0) = 0$$

so that for the dynamics $\dot{q} = 0$ when meeting the collinear set. If $q(t)$ is a geodesic, one has $p_\alpha = 0$ and the Hamiltonian vanishes. It is constant along any geodesic, hence the geodesic is abnormal.

Singularity analysis.

A remarkable property of the geodesics already observed in the historical example (see [10]) is the existence of a cusp singularity for the abnormal curves when meeting the set $\{\|F_0\|_g = 1\}$. This serves as a model to construct a normal form to analyze in a general framework this situation [6].

Theorem 2. *Let $q_1 \in \{\|F_0\|_g = 1\}$. Let σ be a geodesic such that $q_1 = \sigma(0)$ is not an immersion at $t = 0$ and α_1 be the heading at $t = 0$. Then the geodesic σ has an abnormal extremal lift and if $\alpha(\cdot)$ is the heading angle we have only two situations:*

1. *If $\dot\alpha(0) \neq 0$ and $\{\|F_0\|_g = 1\}$ is regular at q_1, then σ has a semi-cubical cusp at q_1.*
2. *If $\dot\alpha(0) = 0$, then $\tilde q_1 = (q_1, \alpha_1)$ is a singular point of the dynamics (3.6) and the spectrum of the linearized dynamics is a feedback invariant.*

Proof. The complete proof is provided in [6] but we indicate the main idea of the proof. The problem is local in a neighborhood of q_1 and we can choose coordinates (x, y) in which $q_1 = (0, 0)$, $F_0 = -\partial/\partial x$ and $g = a(x, y)\left(\mathrm{d}x^2 + \mathrm{d}y^2\right)$ (*isothermal form*). Moreover, F_0 and F_1 have opposite directions. It then suffices to expand $F_0 = b(x, y)\partial/\partial x + c(x, y)\partial/\partial y$ and g at $(0, 0)$ to evaluate $\{\|F_0\|_g = 1\}$, D, D' and D''.

The following theorem describes the optimality properties of the geodesics in a *conic neighborhood* of the small-time reference abnormal arc σ_a.

Theorem 3. *Assume that the reference abnormal arc has a semi-cubical cusp at $t = 0$, then for t small enough:*

1. *The abnormal arc is minimal time from $\sigma_a(t) = q_0$, $t < 0$, until $\sigma_a(0) = q_1$.*
2. *Hyperbolic geodesics starting from q_0 in a conic neighborhood of the abnormal arc are self-intersecting and are minimal time up until their second intersection with the abnormal arc, this point being excluded.*
3. *Elliptic geodesics starting from q_0 in a conic neighborhood of the reference abnormal are maximal time and are confined in the weak current domain $\{\|F_0\|_g < 1\}$.*

The behaviors of the geodesics described in Theorem 3 are represented on Figure 3.1.

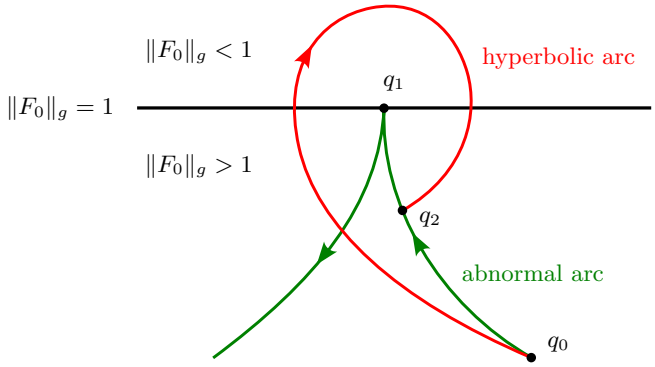


Fig. 3.1: Cusp singularity and self-intersecting arcs in a neighborhood of $\|F_0\|_g = 1$.

Remark 1. In particular, within the framework of Theorem 3, it follows that the minimal control time function is not continuous near the abnormal arc. This implies a loss of local controllability along this arc.

3.2.4 Optimality analysis

Let $\tilde{\sigma}$ be a reference geodesic defined on $[0, t_f]$, $\tilde{\sigma}(t) = (q(t), \alpha(t))$, $\tilde{\sigma}(0) = (q_0, \alpha_0)$ with q_0 being a fixed initial point. The *first conjugate time* along $\tilde{\sigma}$ is the first time t_{1c} at which $\tilde{\sigma}$ ceases to be minimizing, compared with geodesic curves $\tilde{q}$ such that $\tilde{q}(0) = (q_0, \tilde{\alpha}_0)$, $|\alpha_0 - \tilde{\alpha}_0|$ small enough, that is in a conic neighborhood of the reference geodesic. Fixing q_0, the set of first conjugate points is called the *conjugate locus* $C(q_0)$. The *cut time* t_c is the first time at which σ ceases to be (globally) optimal. The set of cut points is called the *cut locus* $\Sigma(q_0)$. Fixing q_0 and q_1 on M, we denote by $T(q_0, q_1)$ the *minimal time value function*, that is, $T(q_0, q_1) = \min t_f$ among all trajectories $q(\cdot)$ such that $q(0) = q_0$ and $q(t_f) = q_1$. The problem is said to be *geodesically complete* if for all $q_0, q_1 \in M$ there exists a minimal time geodesic joining q_0 to q_1.

Proposition 3.

1. *Cusp points correspond to conjugate points along abnormal geodesics.*
2. *In a neighborhood of a cusp point q_1 the time transfer from the point q_0 to q_2 (see Figure 3.1) is larger along the hyperbolic arc than along the abnormal arc.*

Proof. The first assertion comes from Theorem 3. The second assertion is obtained by straightforward computations (see [6] for details).

3.2.5 Liouville-Mineur-Arnold theorem and classification of geodesics in the Riemannian case

We recall the standard Liouville-Mineur-Arnold theorem which is crucial to understand the Hamiltonian dynamics (see [2]).

Theorem 4. *Let (M, ω) be a 4-dimensional symplectic manifold. Let H and G be two smooth functions such that $\{H, G\} = 0$, $\overrightarrow{H}$, $\overrightarrow{G}$ are complete, and H, G are functionally independent. Consider the level surfaces $T_\xi = \{H = \xi_1,\ G = \xi_2\}$ for any $\xi = (\xi_1, \xi_2)$. If T_ξ is connected and compact, then:*

1. *each T_ξ is diffeomorphic to a 2-dimensional torus T^2 called a* Liouville torus*;*
2. *the Liouville foliation is locally trivial and there exist symplectic coordinates (I, φ) called* action-angle variables *in which the dynamics of $\overrightarrow{H}$ become*

$$\frac{\mathrm{d}I_k}{\mathrm{d}t} = 0, \quad \frac{\mathrm{d}\varphi_k}{\mathrm{d}t} = \alpha_k(I), \quad k = 1, 2,$$

and the motion is quasi-periodic.

Application to the Riemannian case on the 2-sphere of revolution $\mathbf{M} = \mathbf{S^2}$.

Consider the family of metrics on S^2 given by $g_\lambda = \mathrm{d}r^2 + m_\lambda^2(r)\,\mathrm{d}\theta$ with

$$m_\lambda^2(r) = \frac{\sin^2 r}{1 - \lambda \sin^2 r}$$

where $\lambda \in [0\,,1)$ is an homotopic parameter, $\lambda = 0$ corresponds to the *round sphere* and $\lambda = 1$ is the *Grushin case*, which is singular at the equator $r = \pi/2$. They were introduced in [3]. The case $\lambda = 4/5$ corresponds to the *averaged Kepler case*.

In the Riemannian case, minimizing the length is equivalent to minimize the energy so that from the Pontryagin maximum principle we infer the following result.

Proposition 4. *Geodesics are solutions of the Hamiltonian dynamics given by the Hamiltonian function*

$$H = \frac{1}{2}\left(H_1^2 + H_2^2\right),$$

with $H_i = \langle p, F_i\rangle$, *for* $i = 1, 2$ *and* $F_1 = \dfrac{\partial}{\partial r}$, $F_2 = \dfrac{1}{m(r)}\dfrac{\partial}{\partial\theta}$. *By homogeneity, one can parametrize by arc length:* $H = 1/2$, *so that for the geodesics the* r*-dynamics is solution of*

$$\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2 = 1 - V(r, p_\theta) \tag{3.7}$$

where $V(r, p_\theta) = 1 - \frac{p_\theta^2}{m^2(r)}$ *is the* potential, *and* p_θ *is constant (Clairaut relation). The* θ*-dynamics satisfies*

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = \frac{p_\theta}{m^2(r)}. \tag{3.8}$$

The metric is reflectionally symmetric with respect to the equator $r = \pi/2$ ($m(r) = m(\pi - r)$) and every geodesic intersects the equator so that the dynamics can be integrated with $q(0) = (\pi/2, 0)$.

Proposition 5. *One can assume that* $p_\theta \in [0\,, m(\pi/2)]$. *Geodesics are given by:*

- *the equator solution* $r = \pi/2$ *for* $p_\theta = m(\pi/2)$;
- *the meridian solution for* $p_\theta = 0$;
- *geodesics which are quasi-periodic.*

Proof. To integrate, one can substitute r by $\pi/2 - r$, so that the equator is identified to $r = 0$, while $m(r)$ is substituted by $m(r) = \cos^2 r/(1 - \lambda\cos^2 r)$. Starting from the equator with $0 < p_\theta < 1/m(r)$, using the ascending branch of (3.7), r oscillates periodically between $-r^+ \leq r \leq r^+$ where r^+ is the positive root of $V(r, p_\theta) = 1$. This leads to r-periodic geodesics.

The second step is to integrate by quadrature the equation (3.8). Altogether, this gives quasi-periodic solutions which are either periodic or dense in a 2-dimensional torus.

Hence, the Riemannian case associated to the family of metrics g_λ fits in the geodesic frame of the Liouville-Mineur-Arnold theorem, provided that the homogeneity $H(\lambda p) = \lambda^2 H(p)$ is taken into account. This opens the way to analyze the Zermelo navigation problem in the case of revolution, based on the mechanical framework, which we do next.

3.2.6 Classification of the geodesics for Zermelo navigation problems on the two-sphere for revolution metrics

Motivated by the applications, we restrict our study to metrics $m_\lambda(r) = \sin^2 r/(1-\lambda\sin^2 r)$ where $\lambda \in [0\,,1]$. For $\lambda = 1$ this corresponds to the singular Grushin case. The current takes the form

$$F_0(q) = \mu_1(r)\frac{\partial}{\partial r} + \mu_2(r)\frac{\partial}{\partial \theta}$$

and the maximized Hamiltonian is

$$M = p_r\,\mu_1(r) + p_\theta\,\mu_2(r) + \|p\|_g + p^0 \tag{3.9}$$

with $\|p\|_g = \sqrt{p_r^2 + p_\theta^2/m^2(r)}$. Moreover, one has $M = 0$ and the hyperbolic, elliptic and abnormal cases correspond respectively to $p^0 < 0$, $p^0 > 0$ and $p^0 = 0$. Using the Pontryagin maximum principle, we get the following result.

Proposition 6. *The geodesics dynamics are the solutions of*

$$\left.\begin{aligned} \frac{\mathrm{d}r}{\mathrm{d}t} &= \mu_1(r) + \frac{p_r}{\|p\|_g} \\ \frac{\mathrm{d}p_r}{\mathrm{d}t} &= -p_r\,\mu_1'(r) - p_\theta\,\mu_2'(r) - \frac{p_\theta^2}{\|p\|_g}\frac{m'(r)}{m^3(r)} \end{aligned}\right\} \tag{3.10}$$

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = \mu_2(r) + \frac{p_\theta}{\|p\|_g}\frac{1}{m^2(r)} \tag{3.11}$$

and $p_\theta = \text{constant}$.

Definition 2. *Fixing* p_θ*, the Hamiltonian dynamics* (3.10) *associated to* M*, restricted to the* (r, p_r)*-space, is called the* Morse-Reeb dynamics.

The main point of the study of the geodesics is to analyze the behaviors of the Morse-Reeb dynamics. To fix the geometric frame we recall next the Morse-Reeb classification of the orbits.

A recap of Reeb classification of 2d-Hamiltonian systems

In this section we present a brief recap of the construction in the 2D Hamiltonian case to deduce our construction, the presentation being based on references [1, 2, 15]. Without losing any generality, one can assume that the 2d-symplectic manifold is the cotangent space T^*M of a 1d-manifold M. Let $z = (p, q)$ be canonical (Darboux) coordinates. Let $H(p, q)$ be an Hamiltonian where $q \in M$ and $\alpha = p\,dq$ is the Liouville form on T^*M and the 2-form ω is the derivative $d\alpha$. We assume that $O = (0, 0)$ is an equilibrium point of the dynamics so that $\mathrm{D}H(O) = 0$. Expanding in the jet-space at O, we note H_2 the quadratic term of the Hamiltonian.

Thanks to Williamson [20], the computations of normal (Jordan) forms in the $2n$-case are reduced to the action of the symplectic group $\mathrm{Sp}(n, \mathbb{R})$. Note that from [13] each symplectomorphism is locally represented by a generating function. Among those, each diffeomorphism $Q = f(q)$ with $\partial f/\partial q$ invertible can be extended to a symplectic transformation with generating mapping

$$S(q, P) = f(q)^{\mathrm{T}} P$$

so that

$$p = \left(\frac{\partial f}{\partial q}(q)\right)^{\mathrm{T}} P, \quad Q = f(q).$$

The diffeomorphism is denoted φ and the induced symplectomorphism $\vec{\varphi}$. It is called a *Mathieu transformation.*

Note that in 2d-case $\mathrm{Sp}(1,\mathbb{R}) = \mathrm{Sl}(2,\mathbb{R})$ and the canonical form coincides with the volume form. From generic point of view we have two situations. In symplectic coordinates the quadratic Hamiltonian is given by:

- Elliptic case: $H_2(P,Q) = \frac{1}{2}\lambda(P^2+Q^2)$;
- Hyperbolic case: $H_2(P,Q) = \frac{1}{2}\lambda(PQ)$.

Lemma 1. *In the previous computations the only linear symplectic invariant is λ which corresponds respectively to the spectrum of $\overrightarrow{H}_2$ that is $\pm i\lambda$ in the elliptic case and $\pm\lambda$ in the hyperbolic case.*

The second step following [1] is to construct the Birkhoff normal form at order m where the polynomic term of the Taylor expansion of H is truncated at order $2m$ and writes

$$H_m = h(x),$$

where $h(x)$ is a polynomial of degree m depending on:

- Elliptic case: $x = (P^2+Q^2)$;
- Hyperbolic case: $x = PQ$.

This normal form is obtained using the Poincaré-Dulac method reducing the Hamiltonian by successive compositions of symplectomorphisms close to the identity and parametrized by their generating functions, see [13] for an algorithmic description of the method. This computation leads to compute a sequence of symplectic invariants in the jet space, generalizing the spectrum λ of the quadratic part.

Lemma 2. *Using the previous calculation, one gets a sequence of symplectic invariants which are the coefficients of the Taylor series of $h(x)$ at $x = 0$.*

Reeb classification.

The previous computation leads to introduce the Reeb classification. The Birkhoff normal leads to compute with an arbitrary accuracy the level sets of H which are in the coordinates (P,Q):

- Elliptic case: concentric circles;
- Hyperbolic case: they are identified to hyperbolas.

Let us introduce the orbits as level sets of H which form a one-dimensional foliation of the symplectic space identified to T^*M. Two points of the space are called equivalent if they belong to the same orbit and we denote by $\sim$ this equivalence relation. The Reeb space denoted $\mathscr{R}$ is the the topological quotient space $T^*M/\!\sim$. In this construction one can define a measure μ on the quotient space by projecting the canonical measure on T^*M on the quotient, using the canonical projection π.

Using this approach the singular points of the dynamics associated to $\overrightarrow{H}$ are the solutions of $\mathrm{D}H = 0$, that is the singular orbits. One can use [15] for a complete description of this construction to classify globally the level sets of the Hamiltonian H and the introduction of the Reeb graph to encode this construction.

Extension of the Morse-Reeb classification to the Zermelo case

Roughly spoken it amounts to classifiy the Hamiltonian dynamics from (3.9) restricting to Mathieu symplectomorphisms in the (r, p_r) space. It has to be adapted using the following obvious property.

Lemma 3. *The Hamiltonians M satisfies $M(\lambda p, \lambda p^0) = \lambda M(p, p^0)$ for $\lambda > 0$.*

Introduction of the potential.

From (3.10), one has:

$$p_r^2 + \frac{p_\theta^2}{m^2(r)} = \left(p^0 + p_r\,\mu_1(r) + p_\theta\,\mu_2(r)\right)^2. \tag{3.12}$$

and from the dynamics (3.10), we deduce:

$$\left(\frac{\mathrm{d}r}{\mathrm{d}t} - \mu_1(r)\right)^2 = 1 - \frac{p_\theta^2}{m^2(r)\left(p^0 + p_r\,\mu_1(r) + p_\theta\,\mu_2(r)\right)^2}. \tag{3.13}$$

In particular, eq. (3.13) generalizes the eq. (3.7) of the Riemannian case, in the case of a parallel current.

Proposition 7. *In the case of a parallel current: $\mu_1(r) = 0$, the r-dynamics is described by the mechanical system:*

$$\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2 = 1 - V(r, p_\theta) \tag{3.14}$$

where

$$V(r, p_\theta) = \frac{p_\theta^2}{m^2(r)\left(p^0 + p_\theta\,\mu_2(r)\right)^2}$$

is the potential.

Hence, in particular we have [5].

Proposition 8. *In the case of a parallel current, an equator $r = r^*$ constant solution of the geodesic dynamics corresponds to a singular point of the Morse-Reeb dynamics with $p_r^* = 0$. The pair (r^*, p_θ^*) is given by solving $V = 1$ and $\partial V/\partial r = 0$. The associated singularity is hyperbolic (resp., elliptic) if and only if $\frac{\partial^2 V}{\partial r^2} < 0$ (resp., $\frac{\partial^2 V}{\partial r^2} > 0$). A separatrix geodesic such that $r(t) \to r^*$ as $t \to \infty$ is necessarily associated to an hyperbolic equator (r^*, p_θ^*).*

Definition 3. *In the case of parallel current, on the two-sphere of revolution, the elliptic case splits into short r-periodic orbits contained in one hemisphere and long periodic orbits crossing the equatorial plane.*

The case of a general current.

If $\mu_1(r)$ is not identically zero, p_r occurs in the right-hand-side of equation (3.13) and hence the r-dynamics has to be analyzed in a more general framework. We proceed as follows. One can write (3.12) as a second order polynomial

$$P(p_r) = a\,p_r^2 + b\,p_r + c = 0 \tag{3.15}$$

with

$$a = 1 - \mu_1^2(r), \quad b = -2\mu_1(r)(p^0 + p_\theta \mu_2(r)), \quad c = \frac{p_\theta^2}{m^2(r)} - (p^0 + p_\theta \mu_2(r))^2.$$

The discriminant of the polynomial P is given by

$$\Delta = b^2 - 4ac = 4(\mu_1^2(r) - 1)\frac{p_\theta^2}{m^2(r)} + 4(p^0 + p_\theta \mu_2(r))^2.$$

The r-dynamics writes

$$\frac{\mathrm{d}r}{\mathrm{d}t} = \mu_1 + \frac{p_r}{\|p\|_g}.$$

Taking the square, one gets a second order equation:

$$P'(p_r) = a' p_r^2 + b' p_r + c' = 0 \tag{3.16}$$

with

$$a' = a = 1 - \mu_1^2, \quad b' = 0, \quad c' = -\frac{p_\theta^2}{m^2}\,\mu_1^2.$$

Hence, the Morse-Reeb classification amounts to analyze the orbits solution of (3.15) and the dynamics on each orbit is given by (3.10). In particular, one needs to solve $P = P' = 0$ and we introduce the following [19].

Definition 4. *The resultant $R(P, P')$ of the two polynomial is given by the determinant of the 4×4 matrix*

$$\begin{pmatrix} a & 0 & a & 0 \\ b & a & 0 & a \\ c & b & c' & 0 \\ 0 & c & 0 & c' \end{pmatrix}.$$

Computations.

We fix p_θ and we compute the roots of $R = 0$. Details are given next in the Lindblad case where practically, the *discrete symmetric group* has to be used to simplify the computations.

3.2.7 A case study with vertical current

Lindblad equation and simplified current.

The dynamics in the Euclidean coordinates $q = (x, y, z)$ are given by

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= -\Gamma x + u_2 z, \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= -\Gamma y - u_1 z, \\ \frac{\mathrm{d}z}{\mathrm{d}t} &= \gamma_- - \gamma_+ z + u_1 y - u_2 x. \end{aligned}$$

The set of parameters $\Lambda = (\Gamma, \gamma_-, \gamma_=)$ is such that: $\Gamma \geq \gamma_+/2 > 0$, $\gamma_+ \geq \gamma_-$ so that the *Bloch ball*: $|q| \leq 1$ is invariant for the dynamics. The parameter Γ is called the *dephasing rate*. We have $\gamma_+ = \gamma_{12} + \gamma_{21}$, $\gamma_- = \gamma_{12} - \gamma_{21}$, where γ_{12}, γ_{21} are the *population relaxation rates*.

The control is the *complex Rabi laser frequency*: $u = u_1 + iu_2$ and we assume that $|u| \leq 1$. Denoting by

$$G_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & +1 & 0 \end{pmatrix}, \quad G_2 = \begin{pmatrix} 0 & 0 & +1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix},$$

G_1 and G_2 correspond respectiveley to rotations around the axis Ox and Oy. The induced metric on the 2-sphere is the Grushin metric.

In order to simplify Lie brackets computations, the original system can be written as the control-affine system

$$\frac{\mathrm{d}q}{\mathrm{d}t} = (G_0 q + v_0) + u_1 G_1 q + u_2 G_2 q$$

with

$$G_0 = \begin{pmatrix} -\Gamma & 0 & 0 \\ 0 & -\Gamma & 0 \\ 0 & 0 & -\gamma_+ \end{pmatrix}, \quad v_0 = \begin{pmatrix} 0 \\ 0 \\ \gamma_- \end{pmatrix},$$

which corresponds to the action of the semi-direct product $\mathrm{Gl}(3,\mathbb{R}) \oplus_s \mathbb{R}^3 \subset \mathrm{Gl}(4,\mathbb{R})$ on the $\mathbb{R}^3$-space with coordinates q identified to the affine space $(1, q)$. The Lie bracket being given by

$$[(a, A), (b, B)] = (Ab - Ba, AB - BA).$$

Using spherical coordinates $x = \rho \sin r \cos\theta$, $y = \rho \sin r \sin\theta$, $z = \rho \cos r$, and using a control feedback preserving the Euclidean norm, the system writes:

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = \gamma_- \cos r - \rho(\gamma_+ \cos^2 r + \Gamma \sin^2 r), \tag{3.17}$$

$$\frac{\mathrm{d}r}{\mathrm{d}t} = -\frac{\gamma_- \sin r}{\rho} + \frac{\sin 2r}{2}(\gamma_+ - \Gamma) + v_2, \tag{3.18}$$

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = -\cot r \, v_1. \tag{3.19}$$

Susch a system is defined in the Bloch ball which is 3-dimensional while the control $v = (v_1, v_2)$ is valued in the 2-dimensional unit ball. Hence, it corresponds to sub-Finsler geometric problem. It will define a Zermelo type navigation problem in the so-called integrable case where γ_- is interpreted as a dissipation parameter setting $\gamma_- = 0$.

The analysis of the orbits fits in the previous section except that singularities occur on the equator since the metric g is singular, in particular the equator is not a solution. Note that we can take the homotopy parameter $\lambda < 1$, $\lambda \sim 1$.

Lemma 4. *If $\gamma_- = 0$, then, the current is vertical and is zero at the equator $r = \pi/2$ and is maximal in each hemisphere at $r = \pi/4$, $\pi/4 + \pi/2$. It can be compensated by a feedback provided $|\gamma_+ - \Gamma| < 2$, thus defining a sub-Finsler problem.*

If we set: $\tilde{r} = \ln \rho$, so that the first equation of (3.17) can be integrated by quadrature and becomes a cyclic variable for the dynamics. This gives in fine two cyclic variables $\tilde{r}$ and θ. Introducing the adjoint vector $p = (p_{\tilde{r}}, p_r, p_\theta)$ the leading maximized Hamiltonian writes

$$M = -(\gamma_+ \cos^2 r + \Gamma \sin^2 r)\, p_{\tilde{r}} + \frac{\sin 2r}{2}(\gamma_+ - \Gamma)\, p_r + \sqrt{p_r^2 + p_\theta^2 \cot^2 r} + p^0.$$

Since $p_{\tilde{r}}$ is constant, this gives a family like of Zermelo navigation problems on the 2-sphere of revolution associated to the Grushin metric given in the dual form by:

$$\|p\|_g = \sqrt{p_r^2 + p_\theta^2 \cot^2 r}.$$

The problem is a test bed case to develop in more details the Morse-Reeb classification, using specific symmetries.

Lemma 5.

– *One has the following symmetries.*

1. *Since the current is vertical ($\mu_2 = 0$), one has either $\frac{\mathrm{d}\theta}{\mathrm{d}t} \equiv 0$ if and only if $p_\theta = 0$ and if $p_\theta \neq 0$, $\frac{\mathrm{d}\theta}{\mathrm{d}t}$ is not vanishing. Hence, we can assume $p_\theta \geq 0$.*
2. *The Hamiltonian M is invariant for the central symmetry $(r, p_r) \mapsto (\pi - r, -p_r)$.*

In particular, one has:

Lemma 6. *Due to the central symmetry, r-periodic geodesics split into short periodic orbits contained in one hemisphere and long periodic orbits crossing the equator.*

Lemma 7. *Since μ_1 is not identically zero, the set of solutions $\{r \mid 1 - \mu_1^2(r) = 0\}$ forms* barriers *and with $\mu_1(r) = \sin 2r\,(\gamma_+ - \Gamma)/2$, this leads if $|\gamma_+ - \Gamma| > 2$ to two barriers in each hemisphere, in which the dynamics is trapped.*

Lemma 8. *Let $t \mapsto (r(t), p_r(t))$ be an orbit of the dynamics so that $|p_r(t)| \to +\infty$ as $t \to +\infty$. Then, the supporting orbit is not compact and moreover $r(t) \to r_0$ as $|t| \to +\infty$, where r_0 is a barrier.*

A simplified model for the complete analysis and numerical simulations

The study of the Lindblad case comes down to the analysis of the family of systems on the 2D sphere with revolution metric given by

- the vertical current: $F_0 = \delta \sin 2r \frac{\partial}{\partial r}$, where δ is a parameter.
- the metrics $m_\lambda^2(r) = \sin^2 r/(1 - \lambda \sin^2 r)$ where $\lambda \in [0\,, 1)$ is a homotopic parameter.

The current is zero at the poles and at the equator and $\|F_0\|_g$ is maximal at $r = \pi/4$, $\pi/4 + \pi/2$. The Finsler case corresponds to $|\delta| < 1$. In each hemisphere the two barriers coincide in the case $|\delta| = 1$ and the dynamics can be studied locally by expanding $\sin 2r$ at $r = \pi/4$ to describe the transition. For the case $|\delta| > 1$, the controllability property can be studied using the barriers since the current is pointing either toward the poles or toward the equator.

Computations of the singularities of the dynamics.

We have $M = p_r \mu(r) + \|p\|_g + p^0$, with $\|p\|_g = \sqrt{p_r^2 + p_\theta^2/m_\lambda^2(r)}$, $\mu(r) = \delta \sin 2r$ and $m_\lambda^2(r) = \sin^2 r/(1 - \lambda \sin^2 r)$. The dynamics reads

$$\frac{\mathrm{d}r}{\mathrm{d}t} = \frac{\partial M}{\partial p_r} = \mu(r) + \frac{p_r}{\|p\|_g}$$
$$\frac{\mathrm{d}p_r}{\mathrm{d}t} = -\frac{\partial M}{\partial r} = -p_r\, \mu'(r) + \frac{p_\theta^2}{m_\lambda^3(r)} \frac{m_\lambda'(r)}{\|p\|_g}.$$

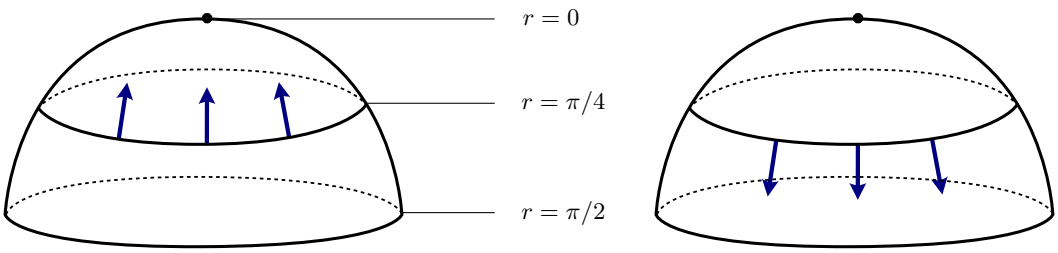


Fig. 3.2: Representation of the current in the North hemisphere. On the left, $\delta > 0$ while on the right $\delta < 0$.

Singularity analysis.

We distinguish between two cases:

- Case $p_r = 0$. We must solve $\mu(r) = 0$ and $p_\theta m'_\lambda(r) = 0$. Since $p \neq 0$, one has $p_\theta \neq 0$ and we get the solution $\mu(r) = m'_\lambda(r) = 0$ which corresponds to the equator $r = \pi/2$.
- Case $p_r \neq 0$. They correspond to additional singularities whose determination is crucial in relationship with short periodic orbits since every solution has to encircle a singular point. From the previous equation, they exist only in the weak current domain where $\|F_0\|_g \leq 1$. In the round case $m'_\lambda = 0$ we must have $\mu'(r) = 0$.

Existence of long periodic orbits.

Lemma 9. *If the level set $M = 0$ is compact, without singular point and has a central symmetry with respect to the point $(r, p_r) = (\pi/2, 0)$, then it contains a periodic trajectory (r, p_r) of period T, and if $p_r^\pm(0)$ are distinct then we have two distinct geodesics $q^+(\cdot)$ and $q^-(\cdot)$ starting from the same point and intersecting with the same length $T/2$ at a point such that $r(T/2) = \pi - r(0)$.*

Proof. The proof is similar to the Riemannian case to construct long periodic orbits starting from the equator $r(0) = \pi/2$. Indeed, consider the equation (3.15) and assume that $\Delta > 0$. Let $p_r^\pm$ be the two distinct roots and let $q^\pm(\cdot) = (r^\pm(\cdot), \theta^\pm(\cdot))$ be the two corresponding distinct geodesics with initial condition $p^\pm(0)$, starting from $(\pi/2, 0)$ and on the same level set $M + p^0 = 0$. Using the central symmetry, $r^\pm$ are T-periodic and moreover $r^+(T/2) = r^-(T/2)$, $\theta^+(T/2) = \theta^-(T/2)$.

Corollary 1. *Long r-periodic orbits correspond to quasi-periodic geodesics preserving quasi-periodic of the Riemannian case.*

Carathéodory-Zermelo-Goh geodesic representation.

We have

$$X = (\mu(r) + \cos\alpha)\frac{\partial}{\partial r} + \frac{\sin\alpha}{m(r)}\frac{\partial}{\partial\theta}, \quad Y = \frac{\partial}{\partial\alpha},$$

in coordinates $\tilde{q} = (r, \theta, \alpha)$, and we compute

$$[Y, X](\tilde{q}) = \sin\alpha\frac{\partial}{\partial r} - \frac{\cos\alpha}{m(r)}\frac{\partial}{\partial\theta},$$

and

$$[[Y,X],Y](\tilde{q}) = \cos\alpha\frac{\partial}{\partial r} + \frac{\sin\alpha}{m(r)}\frac{\partial}{\partial\theta},$$
$$[[Y,X],X](\tilde{q}) = -\mu'\sin\alpha\frac{\partial}{\partial r} + \frac{m'}{m}(1+\mu\cos\alpha)\frac{\partial}{\partial\theta}.$$

Hence,

$$D(\tilde{q}) = \frac{1}{m(r)},$$
$$D''(\tilde{q}) = \frac{1}{m(r)}(1+\mu(r)\cos\alpha)$$

and

$$\begin{aligned} D'(\tilde{q}) &= \sin\alpha\frac{m'}{m^2}(1+\mu\cos\alpha) - \frac{\mu'}{m}\sin\alpha\cos\alpha \\ &= \frac{\sin\alpha}{m}\left(\frac{m'}{m}(1+\mu\cos\alpha) - \mu'\cos\alpha\right). \end{aligned}$$

The dynamics is

$$\frac{\mathrm{d}r}{\mathrm{d}t} = \mu(r) + \cos\alpha,$$
$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = \frac{\sin\alpha}{m(r)},$$
$$\frac{\mathrm{d}\alpha}{\mathrm{d}t} = -\frac{D'(\tilde{q})}{D(\tilde{q})}.$$

The representation is interesting because it encodes the geometric objects. In particular, one can compare with the case study of the historical example of [5].

Proposition 9. *In the case of a vertical current $\mu(r)$:*

1. *The collinear set is the barrier given by $\mu(r)+\cos\alpha = \sin\alpha = 0$.*
2. *The limit abnormal arcs in strong current domains satisfy $\mu(r)\,m(r)\sin\alpha + 1 = 0$.*
3. *D' vanishes along the collinear set.*

3.3 Applications

3.3.1 Numerical simulations for the simplified Lindblad model

We consider the simplified model of the Lindblad system. We first set $\delta = 1.25$ and $r_0 = \pi/2$. We can observe on Figure 3.3 the geodesic flow. One can see that the flow is trapped between the two regions of strong current since in the North hemisphere, the current is pointing down while in the South hemisphere, it is pointing up. We can compute the cut locus which is simply given by an arc of the initial meridian (by symmetry), see Figure 3.4.

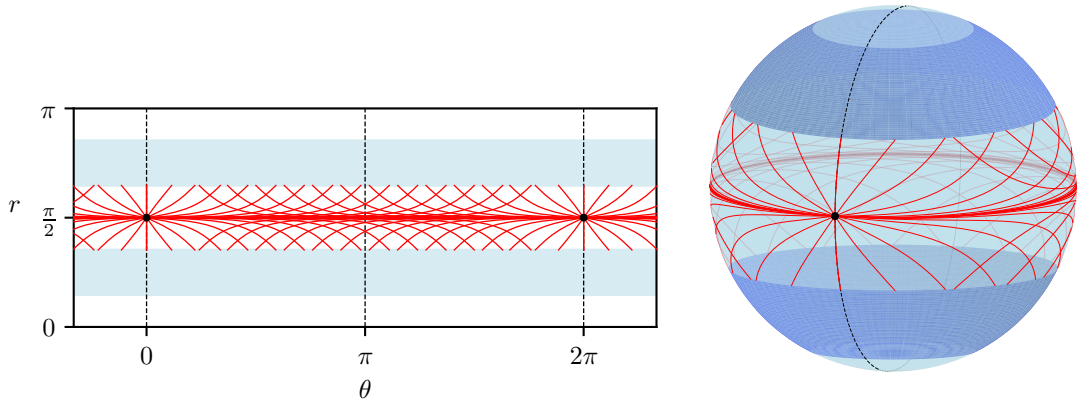


Fig. 3.3: Lindblad problem: $\delta = 1.25$, $r_0 = \pi/2$. Geodesic flow. The red curves correspond to geodesics. The blue strips correspond to the domain of strong current.

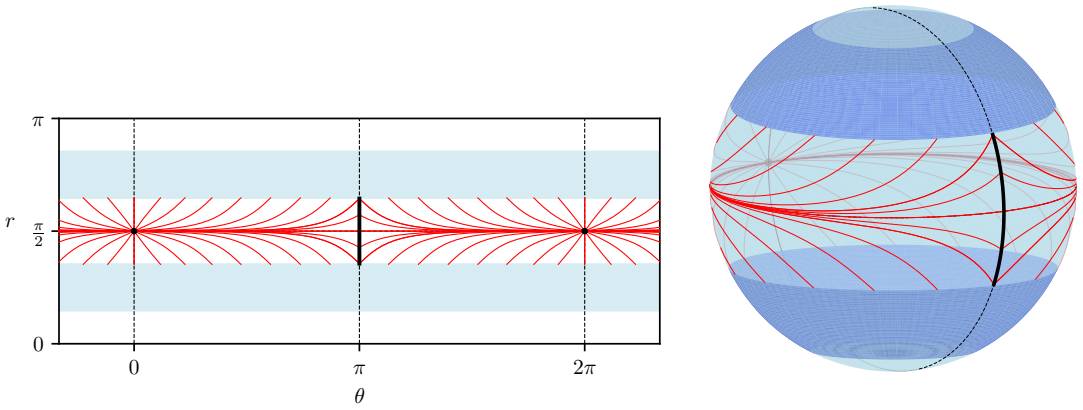


Fig. 3.4: Lindblad problem: $\delta = 1.25$, $r_0 = \pi/2$. Synthesis. The red curves correspond to geodesics. The thick plain black line on the initial meridian is the cut locus. The blue strips correspond to the domain of strong current.

To complete the numerical simulations for the Lindblad problem, we provide geodesic flows in other settings, see Figures 3.6 and 3.7. In Figure 3.6, $\delta = -1.25$ and $r_0 = \pi/2$. This figure can be compared to Figure 3.3. We have represented only the right part of the geodesic flow, that is associated to $p_\theta \geq 0$. When δ is negative, the current is pointing up is the North hemisphere and down in the South one, which explains why the geodesics reach the regions of weak current around the poles. Once the geodesics are in these two regions, they are trapped due to the fact that the current has only a vertical component. The cut locus is more difficult to obtain in this case than for the case where $\delta = 1.25$ because of the folding of the geodesic flow inside the regions of weak current around the poles. In Figure 3.7, on the top, δ is positive while on the sub-figures at the bottom, δ is negative. One the left sub-figures, the initial point is in a region of strong current while for the right sub-figures, the initial point is in a region of weak current around the North pole. We can notice that when δ is positive, then the (hyperbolic) geodesics reach the region of weak current

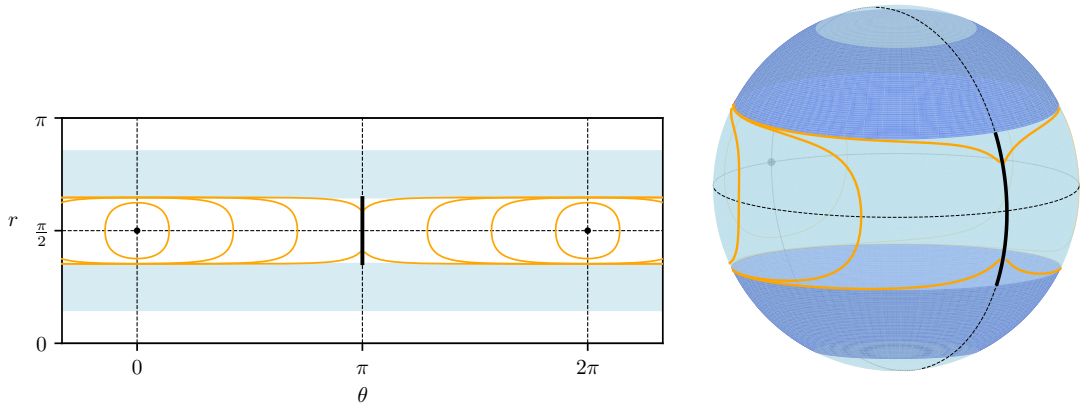


Fig. 3.5: Lindblad problem: $\delta = 1.25$, $r_0 = \pi/2$. Spheres. The orange curves correspond to the spheres at times $t = \{1.0, 3.0, 5.0, 7.2\}$. The thick plain black line on the initial meridian is the cut locus. The blue strips correspond to the domain of strong current.

around the equator and then are trapped, converging to a barrier either in the North hemisphere or the South. When δ is negative the (hyperbolic) geodesics reach a region of weak current around the pole and then are trapped converging again to a barrier.

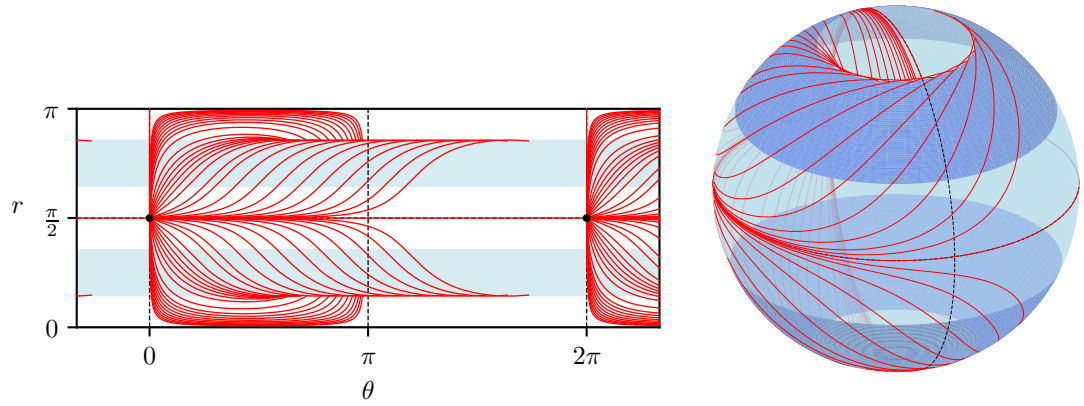


Fig. 3.6: Lindblad problem: $\delta = -1.25$, $r_0 = \pi/2$. Geodesic flow. The red curves correspond to geodesics. The blue strips correspond to the domain of strong current.

3.3.2 The averaged Kepler case

The Riemannian problem related to the averaged Kepler problem in space mechanics (see [3]) can be extended to a metric on a two-sphere of revolution defined in normal coordinates by

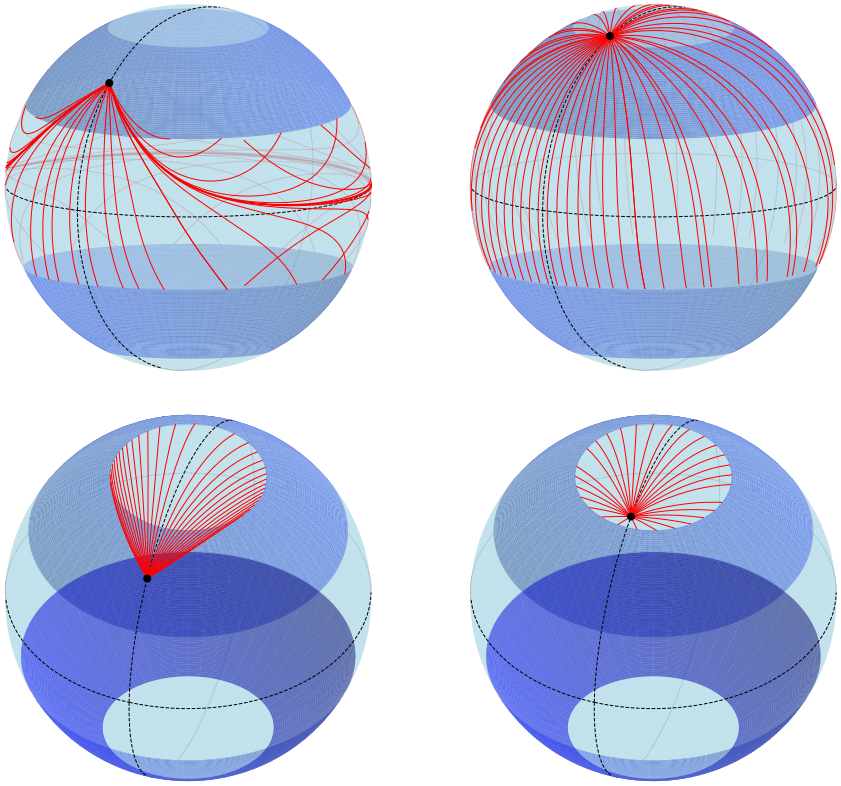


Fig. 3.7: Lindblad problem: $\delta = 1.25$ (Top) and $\delta = -1.25$ (Bottom), $r_0 = \pi/2 + \pi/4$ (Left: strong current at r_0) and $r_0 = \pi/2 + 3\pi/8$ (Right: weak current at r_0). Geodesic flow. The red curves correspond to (hyperbolic) geodesics. The blue strips correspond to the domain of strong current.

$$m^2(r) = \frac{\sin^2 r}{1 - \lambda \sin^2 r}$$

where λ is a homotopic parameter, deforming the round sphere (for $\lambda = 0$) to the singular metric called the *Grushin case* (for $\lambda = 1$) and $\lambda = 4/5$ corresponds to the *averaged Kepler case*. For this case, we will consider a constant current on the covering space. The problem is thus given by

$$F_0 = v\frac{\partial}{\partial\theta}, \quad g = \mathrm{d}r^2 + m^2(r)\mathrm{d}\theta^2,$$

where v is a non-zero constant. Depending on the current at the initial point $q_0 = (r_0, \theta_0)$, we are in the weak (current) case if $\sin^2 r_0 < \frac{1}{v^2+\lambda}$, strong case if $\sin^2 r_0 > \frac{1}{v^2+\lambda}$ and moderate case if $\sin^2 r_0 = \frac{1}{v^2+\lambda}$. In the case where $v^2 + \lambda < 1$, the current will be weak on the whole domain. So we shall assume: $v^2 + \lambda > 1$. The following is a crucial geometric property.

Proposition 10. *On the two-sphere of revolution embedded in $\mathbb{R}^3$, the vector field F_0 defines a linear vector field, tangent to the sphere, and it corresponds to a uniform rotation whose axis is*

the axis of revolution. For the metric the equator solution is also a stationary rotation since $\frac{\mathrm{d}\theta}{\mathrm{d}t}$ *is constant so that the effect of the current can be added to this rotation.*

Integration of the geodesics.

From the previous proposition, the integration follows from the Riemannian case. Introducing the generalized potential, recall that the r-dynamics is given by:

$$\left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2 = 1 - V(r, p_\theta).$$

Taking the ascending branch starting from the equator $r_0 = \pi/2$, we have

$$\frac{\mathrm{d}r}{\mathrm{d}t} = \left(\frac{p_\theta^2\,(1-\lambda\sin^2 r)}{\sin^2 r\,(p^0 + p_\theta v)^2}\right)^{1/2},$$

Since $M = 0$, $\|p\|_g = -(p_\theta v + p^0)$, then, using a time reparameterization, one gets:

$$\frac{\mathrm{d}r}{\mathrm{d}s} = \left(\frac{p_\theta^2\,(1-\lambda\sin^2 r)}{\sin^2 r}\right)^{1/2},$$

which is like the r-dynamics in the Riemannian case, with the addition of v. Then, we can determine the first return mapping to the equator $r_0 = \pi/2$:

$$\frac{\Delta\theta}{2} = \int_{\pi/2}^{r_+} \frac{\partial M/\partial p_\theta}{\partial M/\partial p_r}\mathrm{d}r$$

where r_+ is the maximum of $r(t)$. See Figure 3.8 for an illustration of the geodesic flow.

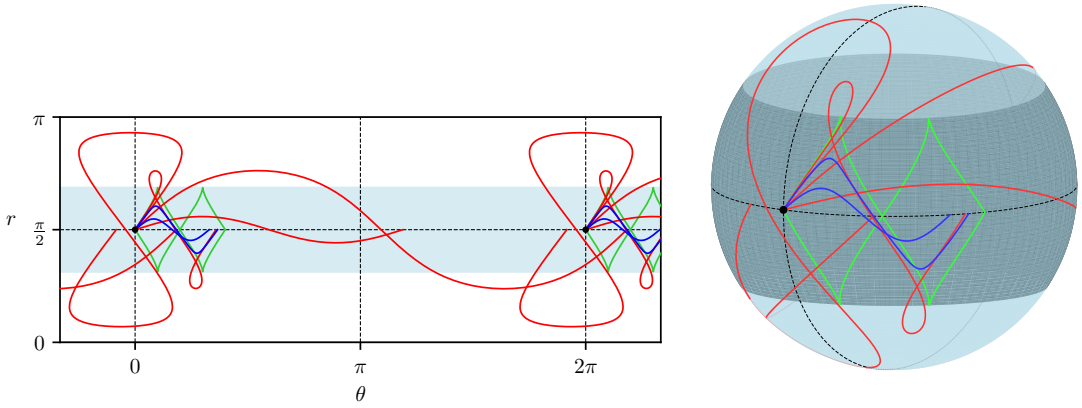


Fig. 3.8: Kepler problem: $\lambda = 4/5$ and $v = 0.8$. Geodesic flow. The red curves correspond to hyperbolic geodesics. The green curves to abnormal and the blue curves to elliptic geodesics. The blue strip corresponds to the domain of strong current.

The geodesic curves are symmetric with respect to the equator, the cone of admissible direction being symmetric with respect to the equator. This leads to the following stratification of the set of geodesics, using the variable p_θ.

Proposition 11. *Assume that $\lambda = 4/5$ and $v = 0.8$, then starting from the equator and considering only the ascending branch, geodesics split into (see also Figure 3.9):*

- *Abnormal given by $p_\theta^a = -1/v$;*
- *Hyperbolic geodesics parameterized by $p_\theta \in (p_\theta^a, m(r_0))$;*
- *Elliptic geodesics parameterized by $p_\theta \in (-m(r_0), p_\theta^a)$.*

Moreover, in the hyperbolic case, the set of geodesics can be stratified in four different classes:

- *The equator which corresponds to $r = \pi/2$, $p_r = 0$ and $p_\theta = m(r_0)$.*
- *The two pseudo-meridians (ascending and descending ones) which correspond, on the covering space, to the non-compact case where $p_\theta = 0$.*
- *Generic r-periodic orbits which split in two different families namely orbits without self-intersections, parameterized by $p_\theta \in (0, m(r_0))$ and orbits with self-intersections, parameterized by $p_\theta \in (p_\theta^a, 0)$ and $\pm p_r(0)$ corresponding to the symmetric orbits.*

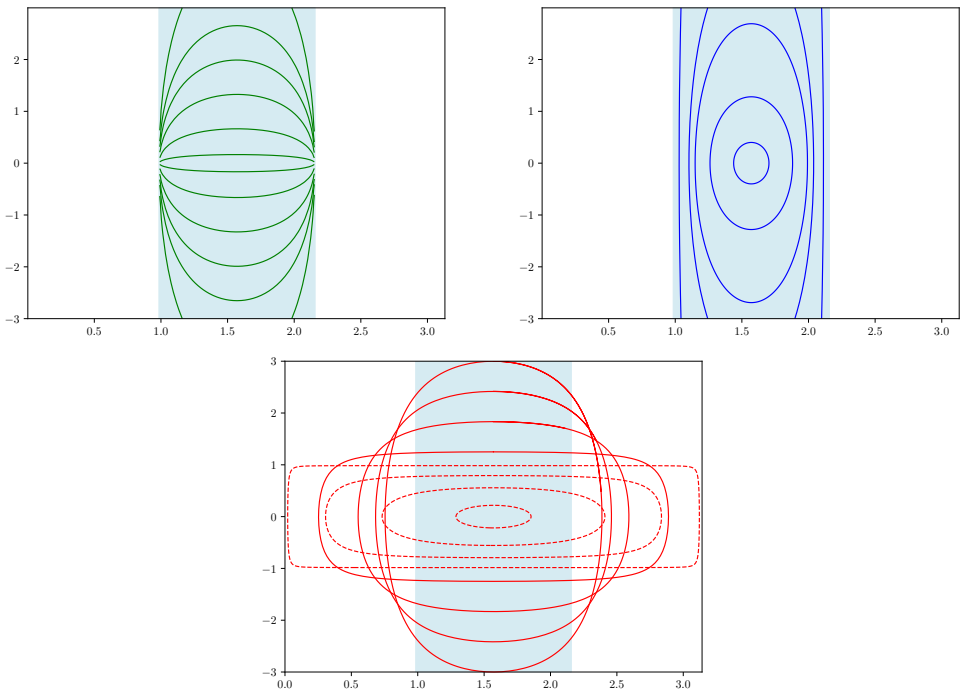


Fig. 3.9: Kepler problem: $\lambda = 4/5$ and $v = 0.8$. Orbits in the (r, p_r) plane. The blue area represents the domain of strong current where the abnormal and elliptic extremals belong. Top-Left: abnormal orbits in green. Top-Right: elliptic orbits in blue. Bottom: hyperbolic orbits in red. The hyperbolic orbits without self-intersections are in dashed lines while orbits with self-intersections are in plain lines. The equator $r = \pi/2$ is a point at $(r, p_r) = (\pi/2, 0)$. The two pseudo-meridians give the transition between the two types of hyperbolic orbits. They correspond to the horizontal lines: $p_r = \pm 1$.

Remark 2. The other geodesics in the flow are obtained by a symmetry with respect to the equator. See Figure 3.10 for the complete classification.

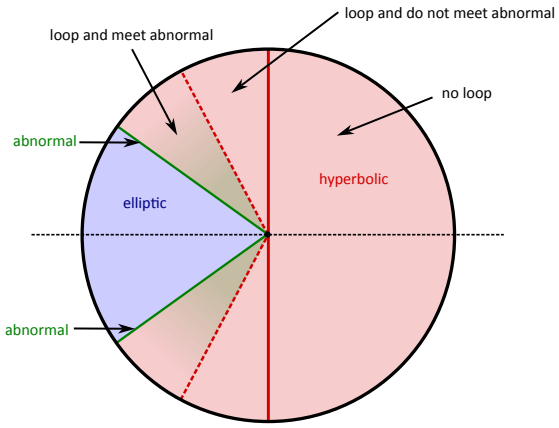


Fig. 3.10: Kepler problem: $\lambda = 4/5$ and $v = 0.8$. Classification of the geodesics using the initial heading angle α_0 with the parameterization $p_\theta = m(r_0)\cos\alpha_0$ and $p_r(0) = \sin\alpha_0$. The term loop stands for self-interesting geodesics. The red vertical line separate the self-interesting hyperbolic geodesics to the hyperbolic geodesics without loops. The red dashed lines separate the hyperbolic geodesics interesting an abnormal to the one without intersection with any abnormal. The abnormals are represented by the green lines. The blue domain corresponds to the elliptic geodesics.

The cut locus in this case will split into two branches. See Figures 3.11 and 3.12 and 3.13. The first branch is associated to the cusp singularity of the abnormal directions, which are symmetric with respect to the equator. The second branch of the cut locus is the persistence of the segment formed by the equator and related to the tame behavior of the first return mapping corresponding to non self-intersecting geodesics. The conjugate points can be numerically evaluated. They exist for different types of geodesics but occur after the intersection of the geodesics with the equator.

Theorem 5. *The cut locus of a point on the equator* ($r_0 = \frac{\pi}{2}$) *has two pairs of symmetric sets. Each decomposes into two branches, the first branch being formed by the abnormal curves occurring in the neighborhood of the cusp point and associated to self-intersecting geodesics and the second branch being a segment of the equator, starting by a cusp point of the conjugate locus and associated to non self-intersecting geodesics.*

3.3.3 The Landau-Lifshitz model for ferromagnetic ellipsoidal samples

Model

This model is borrowed from [8]. We consider hereafter a particular Zermelo-type system modeling the behavior of magnetization in a ferromagnetic sample of ellipsoidal shape. We introduce the magnetization m and an external field u playing the role of a control, both being spatially uniform.

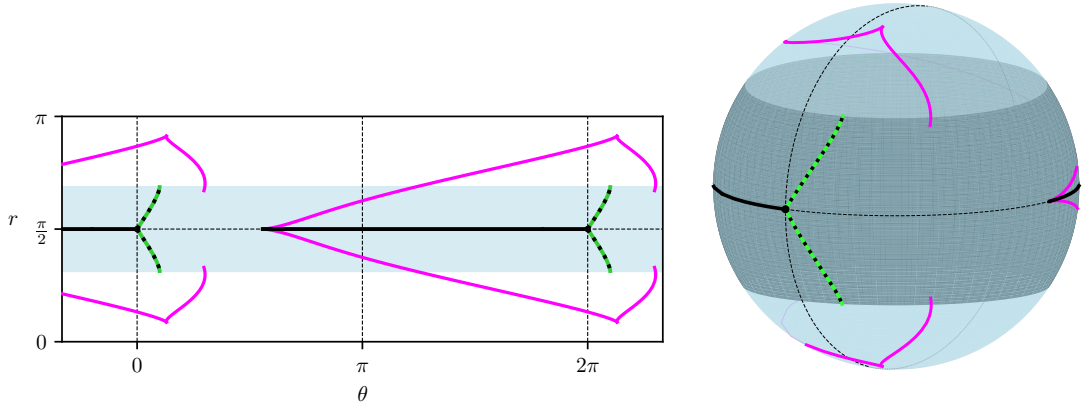


Fig. 3.11: Kepler problem: $\lambda = 4/5$ and $v = 0.8$. Synthesis. The magenta curves correspond to the conjugate locus. The thick plain black line on the equator is one branch of the cut locus. The green curves are part of the two abnormals which are contained in the cut locus, that is why they are also represented by dashed black lines. The blue strip corresponds to the domain of strong current.

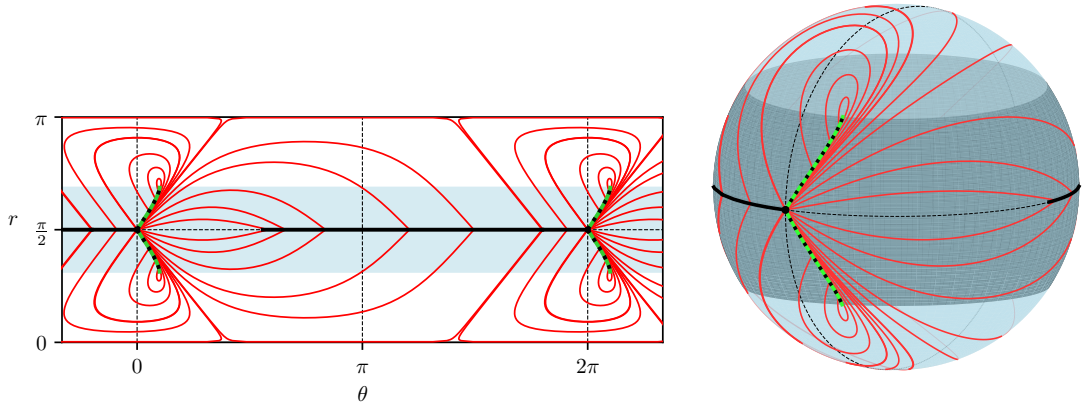


Fig. 3.12: Kepler problem: $\lambda = 4/5$ and $v = 0.8$. Synthesis. The thick plain black line on the equator is one branch of the cut locus. The green curves are part of the two abnormals which are contained in the cut locus, that is why they are also represented by dashed black lines. The red curves correspond to hyperbolic geodesics until their cut points. The blue strip corresponds to the domain of strong current.

Ellipsoidal domains have been much studied in the literature dedicated to ferromagnetism (see [12, 14, 18]).

According to [12, 14], for uniform (in space) magnetizations m on the ellipsoidal sample, the magnetization obeys the Landau-Lifshitz equation

$$\begin{cases} \dfrac{\mathrm{d}m}{\mathrm{d}t} = \alpha\left(h_0(m) - (h_0(m)\cdot m)m\right) - m \wedge h_0(m) \quad \text{in } (0,T) \\ m(0) = m^0 \end{cases} \tag{3.20}$$

where $\alpha > 0$ is a damping parameter, $h_0(m) = -Dm + u$ with a time-dependent external magnetic field u, $T > 0$, $m(t) \in \mathbb{S}^2 \subset \mathbb{R}^3$, $D = \mathrm{diag}(\gamma_1, \gamma_2, \gamma_3)$ denotes a diagonal matrix with nonnegative

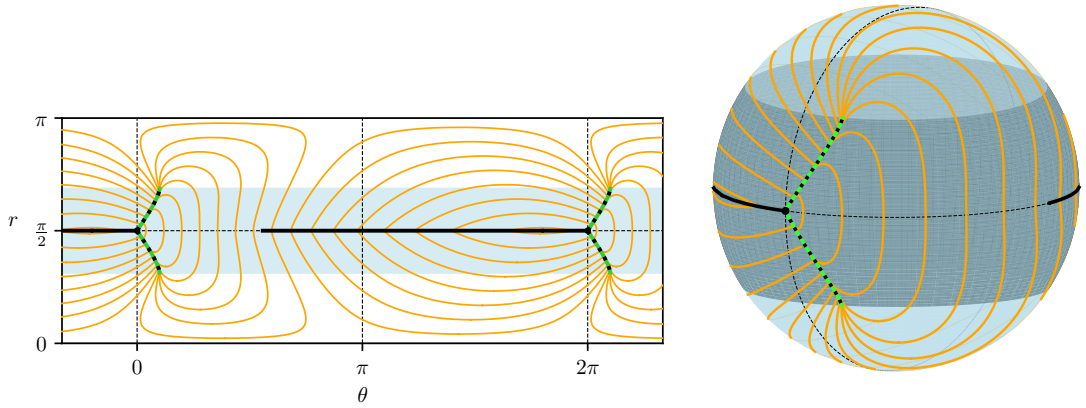


Fig. 3.13: Kepler problem: $\lambda = 4/5$ and $v = 0.8$. Spheres. The orange curves correspond to spheres. One can notice the fan shape of spheres of small radii. The thick plain black line on the initial meridian is the cut locus. The green curves are part of the two abnormals which are contained in the cut locus, that is why they are also represented by dashed black lines. The blue strips correspond to the domain of strong current.

coefficients, where each γ_i $(i = 1, 2, 3)$ is a constant depending only on the semi-axes. We refer for instance to [11] for the dependence of these coefficients on the geometry. Making a change of basis, we assume without loss of generality that $0 \leq \gamma_1 \leq \gamma_2 \leq \gamma_3 \leq 1$.

Reduction to a 2-sphere Zermelo problem

The control applied to the ferromagnetic sample is a control whose maximal intensity $U > 0$ is prescribed, which leads us to write

$$u = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \qquad u_1^2 + u_2^2 + u_3^2 \leq U^2 \quad \text{a.e. in } \mathbb{R}_+.$$

Using adequate changes of unknowns and time reparametrization[6], this leads to a control system of the form

$$\dot{q} = F_0(q) + \sum_{i=1}^{3} v_i F_i(q) \tag{3.21}$$

with $\|v\| \leq 1$.

Since the system evolves on the sphere, we introduce the coordinates

$$m = \begin{pmatrix} \cos r \\ \sin r \cos\theta \\ \sin r \sin\theta \end{pmatrix},$$

and we denote by $q = (r, \theta)$ the polar coordinates on the 2-sphere. Hence, the vector fields in (3.21) are given by

[6] Namely, we consider the new control function $v(\cdot) = u(\cdot)/U$ and operate, with a slight abuse of notation, the change of variable $t \leftarrow t/U$. Moreover, we still denote by γ_i the real numbers γ_i/U.

$$F_0(q) = \begin{pmatrix} \alpha\gamma_1 \cos r \sin r + (\gamma_2 - \gamma_3)\cos\theta\sin\theta\sin r - \\ \qquad\qquad \alpha(\gamma_2\cos^2\theta + \gamma_3\sin^2\theta)\cos r\sin r \\ \alpha(\gamma_2-\gamma_3)\cos\theta\sin\theta - \gamma_1\cos r + (\gamma_3\sin^2\theta + \gamma_2\cos^2\theta)\cos r \end{pmatrix},$$

$$F_1(q) = \begin{pmatrix} -\alpha\sin r \\ 1 \end{pmatrix}, \qquad F_2(q) = \begin{pmatrix} -\sin\theta + \alpha\cos\theta\cos r \\ -\dfrac{\alpha\sin\theta + \cos\theta\cos r}{\sin r} \end{pmatrix},$$

$$F_3(q) = \begin{pmatrix} \cos\theta + \alpha\sin\theta\cos r \\ \dfrac{\alpha\cos\theta - \sin\theta\cos r}{\sin r} \end{pmatrix}.$$

Proposition 12. *The Zermelo navigation problem associated to the Landau-Lifshitz model described above is associated to the maximized Hamiltonian*

$$M = \langle p, F_0(q)\rangle + \sqrt{p_r^2 + \frac{p_\theta^2}{\sin^2 r}} + p^0$$

where the parameters are such that $0 \le \gamma_1 \le \gamma_2 \le \gamma_3$ *and the metric is the standard metric on* S^2 *given by* $g = dr^2 + \sin^2 r d\theta^2$ *with constant curvature 1.*

Proof. Let us denote by $G(q) = (F_1(q), F_2(q), F_3(q))$ the 2×3 matrix formed by concatenating the three vector fields. Denoting

$$e_1 = \begin{pmatrix} \frac{1}{\tan r} \\ \cos\theta \\ \sin\theta \end{pmatrix}, \qquad e_2 = \begin{pmatrix} 0 \\ -\sin\theta \\ \cos\theta \end{pmatrix} \qquad e_3 = \begin{pmatrix} \tan r \\ -\cos\theta \\ -\sin\theta \end{pmatrix},$$

one has

$$\operatorname{Ker} G(q) = \mathbb{R}e_1 \quad \text{and} \quad (\operatorname{Ker} G(q))^\perp = \operatorname{Span}\{e_2, e_3\}.$$

The basis (e_1, e_2, e_3) is orthogonal and direct, and moreover

$$\|e_3\|^2 = 1 + \tan^2 r = \frac{1}{\cos^2 r}.$$

Let us write the control v as

$$v = w_1\frac{e_1}{\|e_1\|} + w_2\frac{e_2}{\|e_2\|} + w_3\frac{e_3}{\|e_3\|}$$

so that

$$Gv = w_2\frac{Ge_2}{\|e_2\|} + w_3\frac{Ge_3}{\|e_3\|} = w_2\begin{pmatrix} 1 \\ \frac{\alpha}{\sin r}\end{pmatrix} + w_3\begin{pmatrix} -\alpha\frac{|\cos r|}{\cos r} \\ \frac{|\cos r|}{\sin r\cos r}\end{pmatrix}.$$

One can assume by symmetry that q belongs to the Northern hemisphere so that $r \in [0, \pi/2]$, where $r = 0$ is the pole. Hence, we have

$$Gv = w_2\begin{pmatrix} 1 \\ \frac{\alpha}{\sin r}\end{pmatrix} + w_3\begin{pmatrix} -\alpha \\ \frac{1}{\sin r}\end{pmatrix}.$$

Hence, the controlled system rewrites as

$$\dot q = F_0(q) + G'w \quad \text{with} \quad G' = \begin{pmatrix} 1 & -\alpha \\ \frac{\alpha}{\sin r} & \frac{1}{\sin r} \end{pmatrix} = (G'_1, G'_2).$$

One has

$$\begin{aligned}\langle p, G'_1\rangle^2 + \langle p, G'_2\rangle^2 &= \left(p_r + \frac{\alpha}{\sin r} p_\theta\right)^2 + \left(-\alpha p_r + \frac{p_\theta}{\sin r}\right)^2 \\ &= (1+\alpha^2)\left(p_r^2 + \frac{p_\theta^2}{\sin^2 r}\right).\end{aligned}$$

Hence, the maximized Hamiltonian reads

$$M = p.F_0(q) + \sqrt{1+\alpha^2}\sqrt{p_r^2 + \frac{p_\theta^2}{\sin^2 r}} + p^0$$

and performing one more renormalization of the parameters γ_i, the proposition is proved.

The dynamics of the Landau-Lifshitz system (3.20) reveal a certain richness when the parameters γ_i and α are varied, as illustrated by a wide variety of trajectory behaviours, see Figure 3.14. In particular, in [11], the existence of basins on the 2-sphere, from which trajectories cannot escape, thus acting as barriers, was demonstrated when the parameter α is chosen above a certain threshold. This phenomenon is illustrated in Figure 3.15, and reflects an obstruction to global controllability.

Geometric properties and computation of $\|F_0\|_g = 1$

Let us use the notation

$$F_0(q) = \mu_1(q)\frac{\partial}{\partial r} + \mu_2(q)\frac{\partial}{\partial \theta}.$$

Proposition 13. *The domains corresponding to $\|F_0\|_g = 1$ are given by $\mu_1(g)^2 + \sin^2 r\,\mu_2(g)^2 = 1$, that is,*

$$(\alpha^2+1)\sin^2 r\left((\gamma_2-\gamma_3)^2\cos^2\theta\sin^2\theta + \cos^2 r\big(\gamma_1-\gamma_3-(\gamma_2-\gamma_3)\cos^2\theta\big)^2\right) = 1.$$

The case of revolution corresponds to $\gamma_2 = \gamma_3$.

The following proposition characterizes the existence of boundaries delimiting strong and weak currents zones (see Fig. 3.16). This comes to investigate the existence of solutions for the equation $\|F_0\|_g = 1$.

Proposition 14.

1. *When $\gamma_2 < \gamma_3$, then a solution exists if the renormalized coefficients satisfy*

$$\frac{4}{(\alpha^2+1)(\gamma_1-\gamma_3)^2} \le 1.$$

2. *In the case of revolution $\gamma_2 = \gamma_3$:*
   a) $\|F_0\|_g = 1$ is equivalent to

$$1 = (\alpha^2+1)(\gamma_1-\gamma_3)^2\sin^2 r\cos^2 r.$$

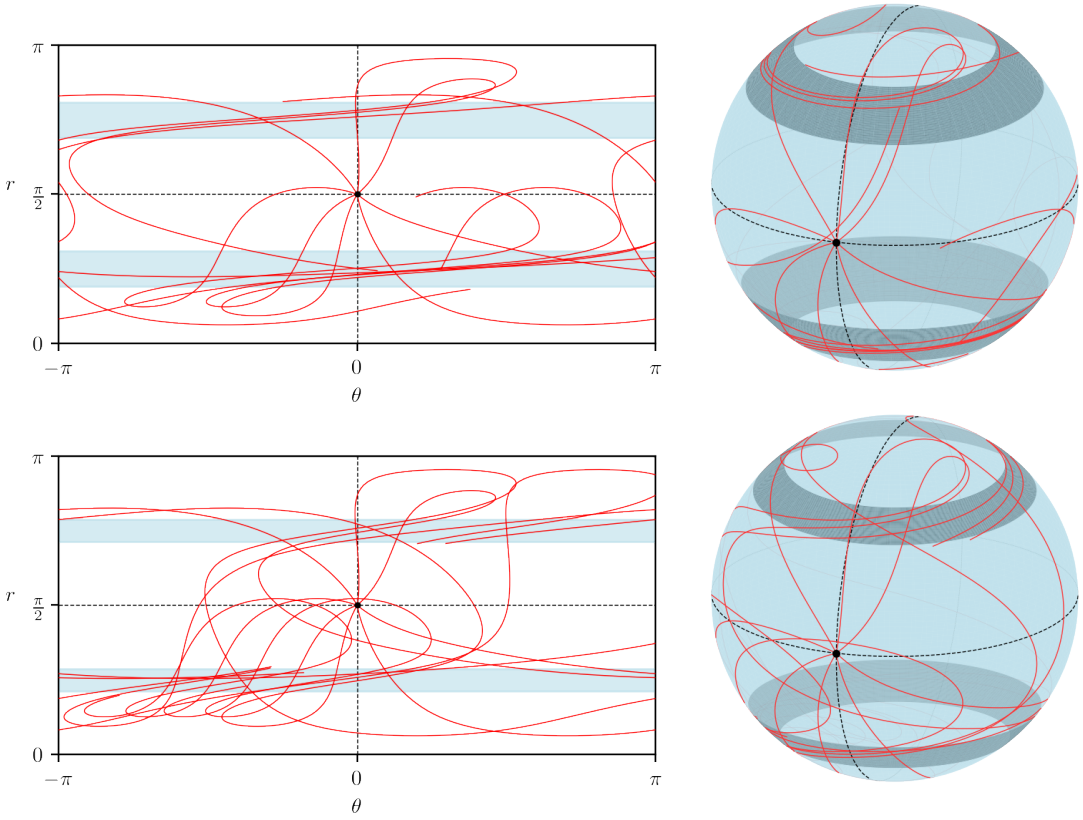


Fig. 3.14: Case of revolution: $\alpha = 1.9$, $\gamma_2 - \gamma_1 = 1$. Example of trajectories. There are no confinement regions here.

b) A solution exists if and only if the renormalized coefficients satisfy

$$\frac{4}{(\alpha^2+1)(\gamma_1-\gamma_3)^2} \leq 1.$$

Proof. We focus on the first point, the second being an easy consequence of the writing of the equation in the particular case $\gamma_2 = \gamma_3$. Hence, let us assume from now on that $\gamma_2 < \gamma_3$. For all $(i,j) \in \{1,2,3\}$, let us set $\gamma_{ij} = \gamma_i - \gamma_j$, and $c = 1/(\alpha^2+1)$. It is straightforward to see that solving $\|F_0\|_g = 1$ is equivalent to the existence of a pair (r,θ) such that $\varphi_\theta(\sin r) = 0$, where

$$\begin{aligned}\varphi_\theta(X) &= (\gamma_{31} - \gamma_{32}\cos^2\theta)^2 X^2 - (\gamma_{32}^2 \cos^2\theta \sin^2\theta + (\gamma_{31} - \gamma_{32}\cos^2\theta)^2)X + c \\ &= (\gamma_{31} - \gamma_{32}\cos^2\theta)^2 X^2 - (\gamma_{21}^2 \cos^2\theta + \gamma_{31}^2 \sin^2\theta)X + c.\end{aligned}$$

Hence, a solution exists if that there exists $\theta \in [0, 2\pi]$ such that the equation $\varphi_\theta(X) = 0$ has a solution in $[0,1]$.

Let us assume temporarily that $\gamma_{21} > 0$ so that the leading coefficient of the polynomial $\varphi_\theta(X)$ is non degenerated. One has

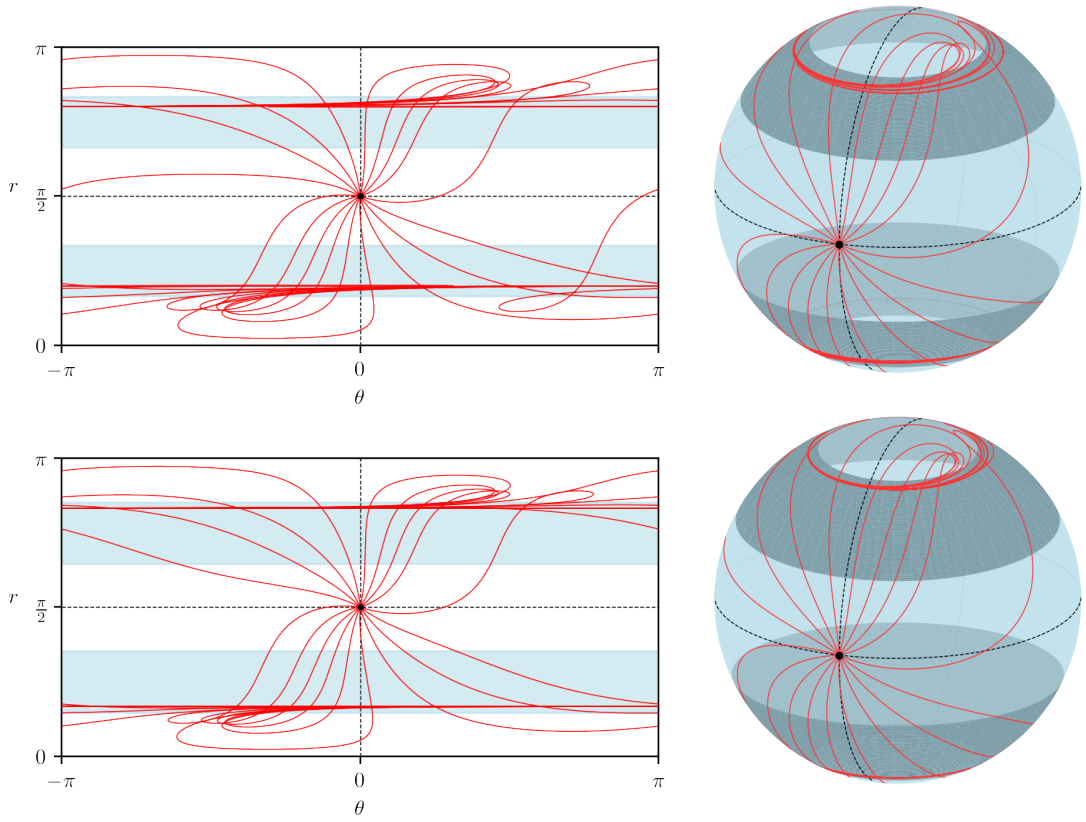


Fig. 3.15: Case of revolution: $\alpha = 2.1$, $\gamma_2 - \gamma_1 = 1$. Example of optimal trajectories, illustrating the confinement regions from which the dynamics cannot escape.

$$\varphi'_\theta(X) = 2(\gamma_{31} - \gamma_{32}\cos^2\theta)^2 X - (\gamma_{21}^2\cos^2\theta + \gamma_{31}^2\sin^2\theta)$$

and therefore, $\varphi'_\theta(0) = -(\gamma_{21}^2\cos^2\theta + \gamma_{31}^2\sin^2\theta) < 0$. The function φ_θ is then convex and decreasing in a neighborhood of 0. A solution thus exists provided that there exists $\theta \in [0, 2\pi]$ such that

$$\min_{X\in\mathbb{R}} \varphi_\theta(X) \leq 0 \quad \text{and} \quad \Big(\varphi'_\theta(1) \geq 0 \quad \text{or} \quad \varphi_\theta(1) \leq 0\Big).$$

We compute

$$\begin{aligned}
\varphi'_\theta(1) &= 2\gamma_{32}^2\cos^4\theta - \gamma_{32}(\gamma_{32} + 2\gamma_{31})\cos^2\theta + \gamma_{31}^2 \\
\varphi_\theta(1) &= c - \gamma_{32}^2\cos^2\theta\sin^2\theta \\
\min_{\mathbb{R}} \varphi_\theta &= c - \frac{(\gamma_{21}^2\cos^2\theta + \gamma_{31}^2\sin^2\theta)^2}{(\gamma_{21}\cos^2\theta + \gamma_{31}\sin^2\theta)^2}
\end{aligned}$$

and we obtain the necessary existence condition by noting that

$$\max_{\mathbb{R}} \varphi_0'(1) = \gamma_{21}^2 > 0 \quad \text{and} \quad \min_{\mathbb{R}} \varphi_0 = c - \frac{\gamma_{31}^2}{4}.$$

Consider now the case $\gamma_{21} = 0$. The equation rewrites

$$\frac{1}{\gamma_{32}^2(\alpha^2+1)} = \sin^2 r \sin^2\theta(1 - \sin^2 r \sin^2\theta).$$

Since

$$\max_{r,\theta} \sin^2 r \sin^2\theta(1 - \sin^2 r \sin^2\theta) = 1/4,$$

the expected conclusion follows.

3.4 Conclusion

In this article we have developed and applied some techniques of geometric optimal control to classify and analyze Zermelo navigation problems on two-spheres of revolution. We have illustrated our results on three case studies, in the fields of quantum control, of orbital transfer and of micromagnetism, providing some numerical simulations. Our findings can be used further to evaluate the fixed time accessibility sets and their boundaries, for instance by combining the techniques of the present paper with a NMPC method (see [17]).

References

1. V. I. Arnold. *Mathematical methods of classical mechanics*, volume 60 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1989. Translated from the Russian by K. Vogtmann and A. Weinstein.
2. A. V. Bolsinov and A. T. Fomenko. *Integrable Hamiltonian systems*. Chapman & Hall/CRC, Boca Raton, FL, 2004. Geometry, topology, classification, Translated from the 1999 Russian original.
3. B. Bonnard and J.-B. Caillau. Geodesic flow of the averaged controlled Kepler equation. *Forum Math.*, 21(5):797–814, 2009.
4. B. Bonnard and M. Chyba. *Singular trajectories and their role in control theory*, volume 40 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 2003.
5. B. Bonnard, O. Cots and B. Wembe. Zermelo navigation problems on surfaces of revolution and geometric optimal control. *ESAIM: Control, Optimisation and Calculus of Variations*, 29:60, 2023.
6. B. Bonnard, J. Rouot and B. Wembe. Accessibility properties of abnormal geodesics in optimal control illustrated by two case studies. *Mathematical Control and Related Fields*, 13(4):1618–1638, 2023.
7. B. Bonnard and D. Sugny, *Time-minimal control of dissipative two-level quantum systems: The integrable case*, SIAM J. Control Optim., **48** (2009), no. 3, 1289–1308.
8. W. F. Brown, *Micromagnetics*, Interscience publishers, 1963, New York.
9. A. E. Bryson. *Applied optimal control: optimization, estimation and control*. Routledge, 2018.
10. C. Carathéodory. *Calculus of variations and partial differential equations of the first order. Part II: Calculus of variations*. Holden-Day, Inc., San Francisco, Calif.-London-Amsterdam, 1967. Translated from the German by Robert B. Dean, Julius J. Brandstatter, translating editor.
11. R. Côte, C. Courtès, G. Ferrière and Y. Privat. Minimal time of magnetization switching in small ferromagnetic ellipsoidal samples. *arXiv preprint arXiv:2301.03839*, 2023.

12. G. Di Fratta. The newtonian potential and the demagnetizing factors of the general ellipsoid. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 472(2190):20160197, 2016.
13. K. R. Meyer and G. R. Hall. *Introduction to Hamiltonian dynamical systems and the N-body problem*, volume 90 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1992.
14. J. A. Osborn. Demagnetizing factors of the general ellipsoid. *Physical review*, 67(11–12):351, 1945.
15. A. Pelayo and S. Vu Ngoc. Symplectic theory of completely integrable hamiltonian systems. *Bulletin of the American Mathematical Society*, 48(3):409–455, 2011.
16. L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko. *The mathematical theory of optimal processes*. Interscience Publishers John Wiley & Sons, Inc., New York-London, 1962. Translated from the Russian by K. N. Trirogoff; edited by L. W. Neustadt.
17. J. B. Rawlings, D. Q. Mayne and M. Diehl, *Model predictive control: theory, computation, and design*, Vol. 2. Madison, WI: Nob Hill Publishing, 2017.
18. D. Takahashi and V. C. Oliveira Jr. Ellipsoids (v1. 0): 3-D magnetic modelling of ellipsoidal bodies. *Geoscientific Model Development*, 10(9):3591–3608, 2017.
19. R. J. Walker. *Algebraic curves*, volume 58. Springer, 1950.
20. J. Williamson. *On the algebraic problem concerning normal forms of linear dynamical systems*. Amer. J. Math. 58 (1936) 141–163. 42.
21. E. Zermelo. Über das navigationsproblem bei ruhender oder veränderlicher windverteilung. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 11(2):114–124, 1931.

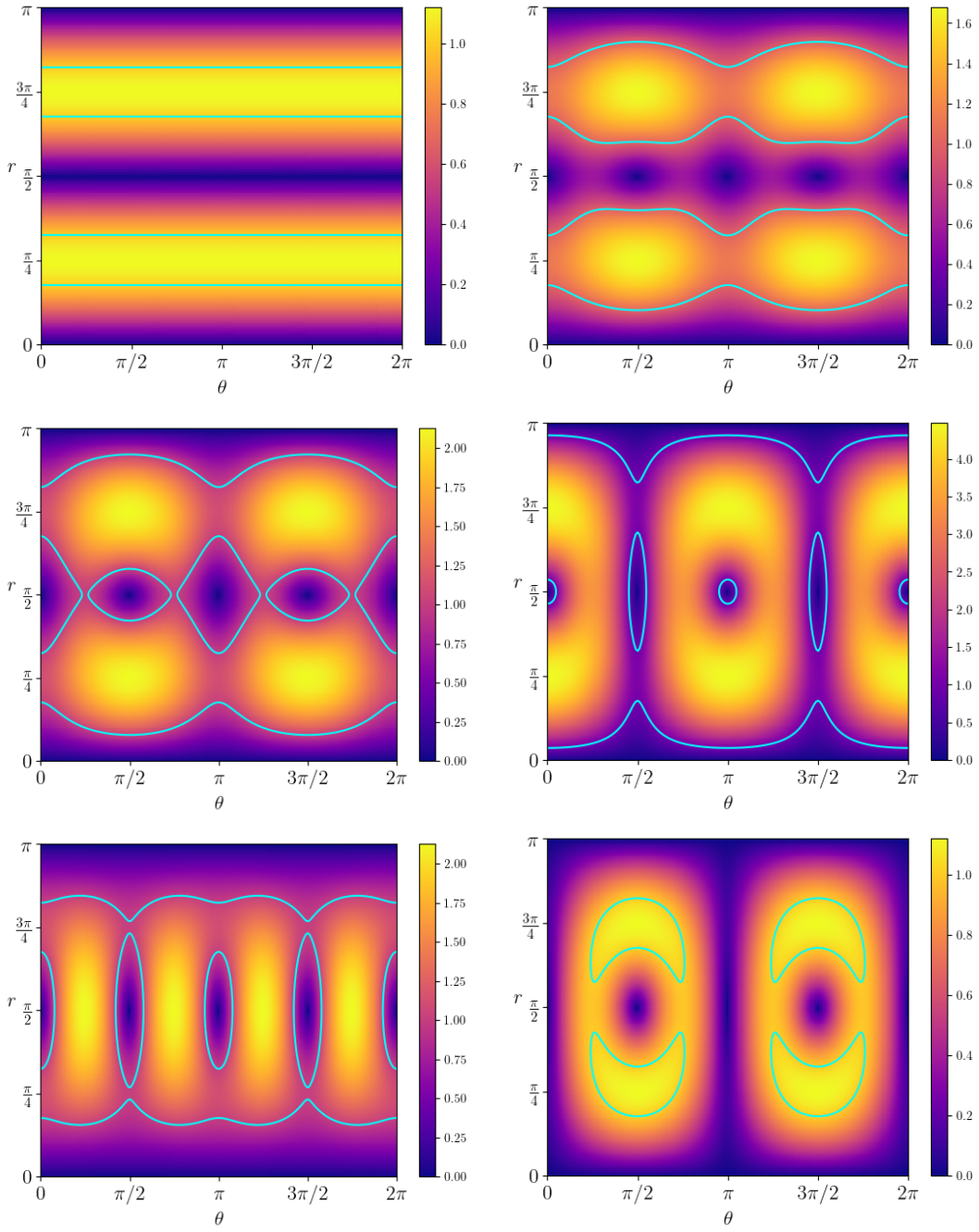


Fig. 3.16: Boundary of weak and strong current domains on the 2-sphere.

Top left: $\alpha = 2$, $\gamma_1 = 1$, $\gamma_2 = 2$, $\gamma_3 = 2$.
Top right: $\alpha = 1$, $\gamma_1 = 2$, $\gamma_2 = 2.5$, $\gamma_3 = 2$.
Middle left: $\alpha = 2$, $\gamma_1 = 1$, $\gamma_2 = 2$, $\gamma_3 = 2.9$.
Middle right: $\alpha = 2$, $\gamma_1 = 1$, $\gamma_2 = 5$, $\gamma_3 = 2$.
Bottom left: $\alpha = 2$, $\gamma_1 = 2.9$, $\gamma_2 = 3.9$, $\gamma_3 = 2$.
Bottom right: $\alpha = 2$, $\gamma_1 = 2.9$, $\gamma_2 = 3.9$, $\gamma_3 = 2$.

4

Time Optimal Control of Ermakov's Equation

Heinz Schättler[1] and Dionisis Stefanatos[2]

[1] Dept. of Electrical and Systems Engineering, Washington University, St. Louis, Mo, 63130 USA
hms@wustl.edu
[2] Materials Science Department, School of Natural Sciences, University of Patras, 26504 Greece
dionisis@post.harvard.edu

Summary. The global solution to the following time-optimal control problem is derived: given $a, b \geq 1$, minimize the transfer time T from $z(0) = 1$ into $z(T) = \gamma > 1$ over all Lebesgue measurable functions $u : [0, T] \to [-a, b]$ subject to the dynamics $\ddot{z} + uz - \frac{1}{z^3} \equiv 0$. This second-order ODE, called Ermakov's equation, models frictionless atom cooling in a harmonic trap. Any of its solutions for the boundary value problem for z achieves a temperature reduction by the factor $\frac{1}{\gamma}$.

dedicated to the memory of Ivan Kupka

4.1 Introduction and Motivation

Given a time-varying (possibly complex) frequency $\omega : [0, T] \to \mathbb{C}$, $t \mapsto \omega(t)$, with $\omega^2(t) \in \mathbb{R}$, and a positive constant ω_0^2, Ermakov's equation [8] is the second-order ODE

$$\ddot{z}(t) + \omega^2(t) z(t) - \frac{\omega_0^2}{z^3(t)} \equiv 0. \tag{4.1}$$

Although non-linear through its driving term, Ermakov's equation can be solved explicitly in terms of the fundamental system for the homogeneous equation [14]: given initial conditions $z(0) = z_0 > 0$ and $\dot{z}(0) = \dot{z}_0$, let x and y be the fundamental system for the homogeneous equation $\ddot{z} + \omega^2(t) z \equiv 0$ that satisfies the initial conditions $x(0) = z_0$, $\dot{x}(0) = \dot{z}_0$ and $y(0) = 0$, $\dot{y}(0) = 1$. Then the solution to Ermakov's equation (4.1) with initial conditions $z(0) = z_0$ and $\dot{z}(0) = \dot{z}_0$ is given by

$$z(t) = +\sqrt{x^2(t) + \left(\frac{\omega_0}{z_0}\right)^2 y^2(t)}. \tag{4.2}$$

Renewed interest in Ermakov's equation stems from the fact that its solutions realize frictionless atom cooling from frequency ω_0 to frequency $\omega_1 < \omega_0$ in a harmonic trap in quantum mechanics [22].

We briefly indicate these connections, but refer to the papers [18, 19] for the details. It is well-known that the Schrödinger equation for the evolution of a wavefunction $\psi(t, x)$ of a particle of mass m in a one-dimensional parabolic trapping potential with frequency $\omega(t)$,

$$i\hbar\frac{\partial\psi}{\partial t} = \left[-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + \frac{1}{2}m\omega^2(t)x^2\right]\psi,$$

can be solved explicitly (in terms of a series of associated eigenfunctions) by separation of variables if $\omega(t) \equiv$ const (e.g., see [5]). Cooling corresponds to lowering the frequency from $\omega_0 = \omega(0)$ to $\omega(T) = \omega_1 < \omega_0$ while frictionless cooling requires that the path $\omega : [0, T] \to \mathbb{C}$ between these two values is chosen so that the populations of all the oscillator levels for $t \geq T$ are equal to those for $t = 0$. Given a solution z to Ermakov's equation, making the following transformation introduced by Kagan, Surkov and Shlyapnikov [10],

$$\widetilde{\phi}(t,\chi) = \sqrt{z(t)}\psi(t, z(t)\chi)\exp\left(-\frac{im}{2\hbar}\chi^2\dot{z}(t)z(t)\right), \tag{4.3}$$

and rescaling time according to

$$\tau(t) = \int_0^t \frac{ds}{z^2(s)} \tag{4.4}$$

with inverse $t = t(\tau)$, the Schrödinger equation for a time-varying frequency $\omega(t)$ with solution $\psi(t, x)$ becomes the Schrödinger equation for $\phi(\tau, \chi) = \widetilde{\phi}(t(\tau), \chi)$ with constant frequency ω_0. This allows us to reduce the problem of frictionless atom cooling in harmonic traps to a study of the solutions to Ermakov's equation [13].

We merely note that, aside from the intrinsic interest in fundamental physics to study systems near absolute zero [12], atom cooling is also of practical relevance as any realization of quantum computing requires low temperatures [1, 4, 7]. Another potential practical application of the problem under study, in the context of quantum thermodynamics, is the optimization of the adiabatic expansion and compression strokes in a quantum heat engine executing the Otto cycle [11].

In this paper, we give a theoretical proof for the global solution to the frictionless atom cooling problem stated in [19]. The case where only controls with real frequencies are used has been studied in [15, 17]. Relaxing this restriction-that is, one allows the trap to become an expulsive parabolic potential for some time intervals-shorter transfer times can be obtained [6]. In the paper [18], frictionless atom cooling was formulated as a minimum-time optimal control problem permitting the frequency to take both real and imaginary values in specified ranges. It was shown that also in this case the optimal solution still consists of bang-bang controls and estimates for the minimum transfer times for various numbers of switchings were given. In our paper [19], and based on a careful analysis of the times between consecutive switchings of extremal bang-bang controls, the synthesis of optimal controlled trajectories for the corresponding time-optimal control problem was described supported by numerical computations. Here this global solution will be proven analytically based on explicit formulas for the transfer times in parameterized families of bang-bang trajectories with n switchings, $n \in \mathbb{N}$. Depending on the numerical values of a and b for the limits of the controls, the solutions are optimal controls with one, respectively two switchings, or, as $\gamma \to \infty$, an increasing number of switchings is required to achieve the minimum transfer times. In the latter case, the optimal number of switchings are determined by cut-loci between the transfer times of these parameterised families of bang-bang extremals. Our proof uses a geometric framework to analyze bang-bang extremals and consists of explicit and, unfortunately, technical calculations. In this paper, the main steps of these calculations will be outlined.

4.2 Frictionless Atom Cooling as an Optimal Control Problem

We rewrite Ermakov's equation as a 2-dimensional system with variables $x_1 = z$ and $x_2 = \frac{\dot{z}}{\omega_0}$, rescale time according to $t_{\text{new}} = \omega_0 t_{\text{old}}$, and introduce the control $u(t) = \left(\frac{\omega(t)}{\omega_0}\right)^2$. This gives us a control system Σ with the following dynamics:

$$\Sigma: \quad \dot{x}_1 = x_2,\ \dot{x}_2 = -ux_1 + \frac{1}{x_1^3},\ -a \le u \le b \quad \text{with } a, b > 0. \tag{4.5}$$

The *state-space* for Σ is the half-space $\mathscr{X} = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 > 0\}$ and the *control set* U is the compact interval $U = [-a, b]$ with a and b positive numbers, i.e., negative values—an expulsive parabolic potential—are allowed [6, 18]. For various reasons [19] it is natural to assume $b \ge 1$ and here we also assume that $a \ge 1$. The latter assumption limits the structure of optimal controls. Throughout this paper we therefore **assume** that $a, b \ge 1$. *Admissible controls* u, $u \in \mathscr{U}$, are Lebesgue measurable functions defined over some compact interval $[0, T]$ with values in U. The terminal time T is free. Given any admissible control $u \in \mathscr{U}$ defined over $[0, T]$ and an arbitrary initial condition $(x_1^0, x_2^0) \in \mathscr{X}$, it follows from the representation (4.2) that the corresponding trajectory $(x_1(t), x_2(t))$ exists on all of $[0, T]$ and remains positive. The system (4.5) can be interpreted as one-dimensional Newtonian motion of a unit-mass particle with position x_1, velocity x_2 and acceleration (force) acting on the particle given by $-ux_1 + \frac{1}{x_1^3}$. This point of view provides useful intuition about the time-optimal solution that we shall compute.

Minimum time frictionless atom cooling is realized by the solution to the following optimal control problem **[OC]**: among all admissible controls $u \in \mathscr{U}$ for the control system Σ, find one that minimizes the transfer time from the initial condition $(x_1(0), x_2(0)) = (1, 0)$ into the terminal state $(x_1(T), x_2(T)) = (\gamma, 0)$, $\gamma > 1$.

We write the dynamics (4.5) in the form $\dot{x} = f(x) + ug(x)$ with drift vector field f and control vector field g given by

$$f(x) = \begin{pmatrix} x_2 \\ \frac{1}{x_1^3} \end{pmatrix} \qquad \text{and} \qquad g(x) = \begin{pmatrix} 0 \\ -x_1 \end{pmatrix}. \tag{4.6}$$

We denote the vector fields corresponding to the constant controls $u(t) \equiv -a$ and $u(t) \equiv b$ by $X = f - ag$ and $Y = f + bg$, respectively, and label the trajectories corresponding to these constant controls as X- and Y-trajectories. Note that $f = \frac{aY + bX}{a+b}$ and $g = \frac{Y - X}{a+b}$ and admissible directions for the control system are convex combination of X and Y:

$$f + ug = \frac{b-u}{a+b} X + \frac{u+a}{a+b} Y.$$

We write XY for a concatenation of an X-trajectory followed by a Y-trajectory and YX for a concatenation in the inverse order. Analogous notations will be used for concatenations of more pieces.

Extremals are controlled trajectories which satisfy the necessary conditions for optimality of the Pontryagin maximum principle (e.g., see [2, 3, 16]). We shall show below that all extremals are finite combinations of X and Y-trajectories. As we shall pronounce the best of all extremals to be optimal, the following result is important to our reasoning.

Proposition 1. *The minimum time frictionless atom cooling problem [OC] has a solution.* □

We briefly outline the argument of the proof: Because of the presence of the explosive potential at $x_1 = 0$, the usual growth assumptions made in standard results on existence of optimal controls (e.g., see [3, 9]) are not satisfied. It can be shown, however, that, given $\gamma > 1$, optimal controlled trajectories lie in a compact subset of the state-space $\mathscr{X}$. For, we shall show below (Proposition 2) that there exists a unique XY-extremal which transfers the point $(1, 0)$ into the target $(\gamma, 0)$ in a finite time $\tilde{T}$. Using the representation (4.2) for the solutions to Ermakov's equation it can be shown that given any admissible controlled trajectory (x, u) defined over the interval $[0, \tilde{T}]$, for all $t \in [0, \tilde{T}]$ it holds that $x_1(t) \leq \sqrt{2}e^{\mu t}$ where $\mu = \max\{1, a, b\}$. Using the explicit geometric shapes of X- and Y-trajectories (see Figure 4.1 below), it is then straightforward to define a compact set $K \subset \mathscr{X}$ which only depends on a, b and $\tilde{T}$ such that the trajectories of all admissible controlled trajectories (x, u) defined over the interval $[0, \tilde{T}]$ lie in K. Once this is shown, existence of an optimal control follows from standard results.

We call an extremal $YX \dots XY$-trajectory with $2n$ switchings a Y-loop with n turns. The theorem below summarizes the structure of the optimal solutions for the problem [OC].

Theorem 1. *We write $v = b + 1$ and $w = a + b$.*

(A) *If $v^2 \leq 4w \Leftrightarrow b \leq 1 + 2\sqrt{a}$, then for all $\gamma > 1$ the one switch XY-trajectories are optimal.*

(B) *For $v^2 > 4w \Leftrightarrow b > 1 + 2\sqrt{a}$, let $\zeta^2 = \frac{v}{2w}\left\{1 - \sqrt{1 - \frac{4w}{v^2}}\right\}$ and define*

$$\varpi(a, b) = \frac{1}{\sqrt{b}} \sin^{-1}\left(\sqrt{\frac{b(1-\zeta^2)}{b-1}}\right) + \frac{1}{\sqrt{a}} \ln\left(\sqrt{(a+1)\zeta^2}\right), \tag{4.7}$$

$$\widehat{\varpi}(a, b) = \frac{1}{\sqrt{a}} \ln\left(2\sqrt{\frac{a}{a+b}}\right) + \frac{1}{\sqrt{b}}\left(\pi - \sin^{-1}\left(\sqrt{\frac{b}{a+b}}\right)\right). \tag{4.8}$$

In this case, the optimal controlled trajectories for the minimum-time frictionless atom cooling problem are as follows:

1. *If $\varpi(a, b) \geq 0$, then the one switch XY-trajectories are optimal for all $\gamma > 1$.*
2. *If $\varpi(a, b) < 0$, then there exists a unique value $\widehat{\gamma}_1$ for which the one-switch extremal XY-trajectory which steers 1 into γ takes the same time as the Y-loop YXY with one turn. The one-switch XY-trajectories are optimal for $1 < \gamma \leq \widehat{\gamma}_1$.*
   a) *If $\widehat{\varpi}(a, b) \geq 0$, then for all $\gamma \geq \widehat{\gamma}_1$ the Y-loop with one turn is optimal.*
   b) *If $\widehat{\varpi}(a, b) < 0$, then for every $n \in \mathbb{N}$, $n > 1$, there exists a unique value $\widehat{\gamma}_n$ for which the Y-loop with $n-1$ turns takes the same time as the Y-loop with n turns. The sequence $\{\widehat{\gamma}_n\}_{n\in\mathbb{N}}$ is strictly monotonically increasing and diverges to ∞ as $n \to \infty$. For $\gamma \in [\widehat{\gamma}_n, \widehat{\gamma}_{n+1}]$, $n \in \mathbb{N}$, Y-loops with n turns are optimal.*

We develop the proof of this result in the remaining sections of the paper. The proof also leads to a simple, straightforward numerical algorithm which allows us to compute the cut-loci $\{\widehat{\gamma}_n\}_{n\in\mathbb{N}}$ and the switching points of the optimal trajectories.

4.3 Preliminary Observations

It is easy to see that optimal controls are bang-bang, i.e., optimal controlled trajectories are finite concatenations of X- and Y-trajectories. The control Hamiltonian H is given by

$$H = H(\lambda_0, \lambda, x, u) = \lambda_0 + \lambda_1 x_2 + \lambda_2 \left(-u x_1 + \frac{1}{x_1^3} \right) \tag{4.9}$$

where λ_0 is a non-negative constant and $\lambda = (\lambda_1, \lambda_2) \in \left(\mathbb{R}^2\right)^*$ is a co-vector. If (x_*, u_*) is an optimal controlled trajectory, then it follows from the necessary conditions for optimality of the maximum principle (e.g., see [2, 3, 16]) that there exists a non-trivial solution $\lambda(t)$ to the adjoint equations,

$$\dot{\lambda}_1(t) = \lambda_2(t) \left(u + \frac{3}{x_1^4} \right) \quad \text{and} \quad \dot{\lambda}_2(t) = -\lambda_1(t), \tag{4.10}$$

such that the Hamiltonian H vanishes identically along (x_*, u_*) and $\lambda(t)$ and optimal controls $u_*(t)$ are given by $u_*(t) = -a$ if $\lambda_2(t) < 0$ and $u_*(t) = b$ if $\lambda_2(t) > 0$. The non-triviality of the multiplier λ implies that $\dot{\lambda}_2(\tau) = -\lambda_1(\tau) \neq 0$ whenever $\lambda_2(\tau) = 0$ and thus the switching function $\Phi(t) = \lambda_2(t)$ changes sign at every zero. Furthermore, zeros of λ_2 cannot accumulate at a finite time τ as otherwise also the zeros of the derivative would accumulate at τ and thus $\lambda(\tau) = 0$. Hence there only exist a finite number of switchings on any compact interval $I \subset [0, \infty)$, i.e., *optimal controls are bang-bang.*

Solving the optimal control problem consists in establishing the precise concatenation sequences for X and Y (how many switchings are there and in what order) and calculating the times between the switchings of the controls. This is a more involved and at times delicate undertaking for which one needs to fully understand the phase portraits of the vector fields X and Y. For a constant control $u(t) \equiv u = \text{const}$, Ermakov's equation is a Hamiltonian system with the Hamiltonian function $\mathscr{H}$ given by

$$\mathscr{H} = \frac{1}{2} \left(x_2^2 + u x_1^2 + \frac{1}{x_1^2} \right). \tag{4.11}$$

Integral curves of the dynamics are the level curves of $\mathscr{H}$. For the trajectory passing through the point $(\alpha, 0)$, $\alpha > 0$, we have that

$$x_2^2 = -u x_1^2 - \frac{1}{x_1^2} + u\alpha^2 + \frac{1}{\alpha^2} = \left(x_1^2 - \alpha^2 \right) \left(\frac{1}{\alpha^2 x_1^2} - u \right). \tag{4.12}$$

If $u \leq 0$, it follows that $x_1(t) \geq \alpha$ along any such trajectory whereas, if $u > 0$, then we obtain that $x_1(t)$ lies between the values α and $\frac{1}{\alpha\sqrt{u}}$. In this case, the trajectories are periodic orbits with these values their extreme points in the x_1-direction. The equilibrium solution is given by $(x_1^*, x_2^*) \equiv \left(\sqrt[4]{\frac{1}{u}}, 0 \right)$. Representative examples of the integral curves are shown in Figure 4.1.

Proposition 2. [19] *The terminal point $(\gamma, 0)$, $\gamma > 1$, is reachable from the initial point $(1, 0)$ through a unique XY-trajectory. This trajectory is an extremal and the time $T_1 = T_1(\gamma)$ to steer $(1, 0)$ into $(\gamma, 0)$ is given by*

$$\begin{aligned} T_1(\gamma) = \frac{1}{\sqrt{a}} \sinh^{-1} \left(\sqrt{\frac{a(\gamma^2 - 1)}{a + b} \cdot \frac{b\gamma^2 - 1}{(a+1)\gamma^2}} \right) \\ + \frac{1}{\sqrt{b}} \sin^{-1} \left(\sqrt{\frac{a\gamma^2 + 1}{a + b} \cdot \frac{b(\gamma^2 - 1)}{b\gamma^4 - 1}} \right). \end{aligned} \tag{4.13}$$

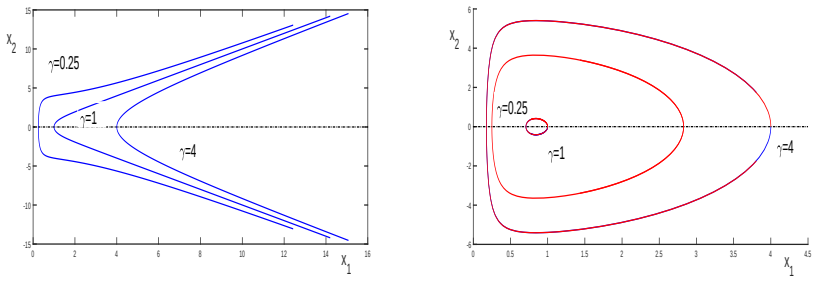


Fig. 4.1: Phaseportraits for the vector fields $X = f - ag$ (left, shown in blue for $a = 1$) and $Y = f + bg$ (right, shown in red for $b = 2$).

Proof. The Y-loop through $(\gamma, 0)$ lies to the left of $(\gamma, 0)$ and contains the initial point $(1, 0)$ in its interior. Hence there exists a unique point (κ, μ) in the upper quadrant, $\mu > 0$, where the forward X-orbit from $(1, 0)$ intersects the backward Y-orbit through $(\gamma, 0)$ and we have $1 < \kappa < \gamma$. The coordinates are easily computed from the relations (c.f., equation (4.12))

$$\mu^2 = (\kappa^2 - 1)\left(\frac{1}{\kappa^2} + a\right) \quad \text{and} \quad \mu^2 = (\kappa^2 - \gamma^2)\left(\frac{1}{\gamma^2\kappa^2} - b\right).$$

Writing $\mathscr{H}$ in the forms

$$\mu^2 - a\kappa^2 + \frac{1}{\kappa^2} = -a + 1 \quad \text{and} \quad \mu^2 + b\kappa^2 + \frac{1}{\kappa^2} = b\gamma^2 + \frac{1}{\gamma^2},$$

it follows that

$$\kappa^2 = \frac{b\gamma^2 + \frac{1}{\gamma^2} + a - 1}{a + b} = \frac{b\gamma^4 + (a-1)\gamma^2 + 1}{\gamma^2(a+b)}. \tag{4.14}$$

Concatenating the X-trajectory with the Y-loop at (κ, μ) generates the unique XY concatenation that steers $(1, 0)$ into $(\gamma, 0)$. It is straightforward to verify that these XY-trajectories are normal extremals and that YX-trajectories cannot steer $(1, 0)$ into $(\gamma, 0)$.

It remains to compute the time $T_1(\gamma)$. The following Lemma from [19] allows us to evaluate the times along X- and Y-trajectories. We let $\mathscr{X}_+ = \{(x_1, x_2) \in \mathscr{X} : x_2 > 0\}$ and $\mathscr{X}_- = \{(x_1, x_2) \in \mathscr{X} : x_2 < 0\}$.

Lemma 1 (Time evolution of x_1). *The time evolution of x_1 along an X-trajectory starting from $(\alpha, 0)$ at time $t = 0$ is given by the following equivalent representations*

$$x_1(t) = \sqrt{\alpha^2 \cosh(\sqrt{a}t) + \frac{1}{a\alpha^2}\sinh(\sqrt{a}t)} \tag{4.15}$$

$$= \sqrt{\frac{1}{2}\left(\alpha^2 - \frac{1}{a\alpha^2}\right) + \frac{1}{2}\left(\alpha^2 + \frac{1}{a\alpha^2}\right)\cosh(2\sqrt{a}t)}, \tag{4.16}$$

while the evolution along a Y-trajectory starting from $(\beta, 0)$ at time $t = 0$ is given by

$$x_1(t) = \sqrt{\beta^2 \cos^2(\sqrt{b}t) + \frac{1}{b\beta^2}\sin^2(\sqrt{b}t)} \tag{4.17}$$

$$= \sqrt{\frac{1}{2}\left(\beta^2 + \frac{1}{b\beta^2}\right) + \frac{1}{2}\left(\beta^2 - \frac{1}{b\beta^2}\right)\cos(2\sqrt{b}t)}. \tag{4.18}$$

Proof. It follows from equation (4.2) that the solution $x_1(t)$ for the vector field X is given by

$$x_1(t) = \sqrt{u(t)^2 + \left(\frac{v(t)}{\alpha}\right)^2}$$

where u and v are the fundamental solutions to the homogeneous equation $\ddot{z} = az$ with initial conditions $u(0) = \alpha$, $\dot{u}(0) = 0$ and $v(0) = 0$, $\dot{v}(0) = 1$. These equations have the solutions $u(t) = \alpha\cosh(\sqrt{a}t)$ and $v(t) = \frac{1}{\sqrt{a}}\sinh(\sqrt{a}t)$ which gives us equation (4.15). Equation (4.16) follows from basic identities for the hyperbolic sine and cosine.

Analogously, the evolution of Y-trajectories from the point $(\beta, 0)$ is governed by

$$x_1(t) = \sqrt{u(t)^2 + \left(\frac{v(t)}{\beta}\right)^2}$$

where $u(t) = \beta\cos(\sqrt{b}t)$ and $v(t) = \frac{1}{\sqrt{b}}\sin(\sqrt{b}t)$. This gives (4.17) and (4.18) follows from standard trigonometric identities. □

Corollary 1. *Given an X-trajectory that starts at $(\alpha, 0) = (x_1(0), x_2(0))$, the time $t > 0$ until the point $(x_1(t), x_2(t)) = (\kappa, \mu) \in \mathscr{X}_+$, is reached is given by*

$$t = \frac{1}{\sqrt{a}}\sinh^{-1}\left(\alpha\sqrt{\frac{a(\kappa^2 - \alpha^2)}{1 + a\alpha^4}}\right). \tag{4.19}$$

Similarly, given a Y-trajectory that starts at $(\beta, 0) = (x_1(0), x_2(0))$, the time $t > 0$ until the point $(x_1(t), x_2(t)) = (\zeta, \xi) \in \mathscr{X}_-$, is reached is given by

$$t = \frac{1}{\sqrt{b}}\sin^{-1}\left(\beta\sqrt{\frac{b(\zeta^2 - \beta^2)}{1 - b\beta^4}}\right). \tag{4.20}$$

Proof. It follows from equation (4.15) that

$$\kappa^2 = \alpha^2\cosh^2(\sqrt{a}t) + \frac{1}{a\alpha^2}\sinh^2(\sqrt{a}t) = \alpha^2 + \left(\alpha^2 + \frac{1}{a\alpha^2}\right)\sinh^2(\sqrt{a}t)$$

and thus

$$\sinh^2(\sqrt{a}t) = \frac{\kappa^2 - \alpha^2}{\alpha^2 + \frac{1}{a\alpha^2}} = \frac{a\alpha^2(\kappa^2 - \alpha^2)}{1 + a\alpha^4}.$$

Similarly, along a Y-trajectory by (4.17) we have that

$$\zeta^2 = \beta^2\cos^2(\sqrt{b}t) + \frac{1}{b\beta^2}\sin^2(\sqrt{b}t) = \beta^2 + \left(\frac{1}{b\beta^2} - \beta^2\right)\sin^2(\sqrt{b}t)$$

and thus

$$\sin^2(\sqrt{b}t) = \frac{\beta^2 - \zeta^2}{\beta^2 - \frac{1}{b\beta^2}} = \frac{b\beta^2(\beta^2 - \zeta^2)}{b\beta^4 - 1}.$$

Using these formulas, we compute $T_1(\gamma)$: setting $\alpha = 1$ in equation (4.19) and using (4.14) we obtain that

$$\frac{a(\kappa^2 - 1)}{1 + a} = \frac{a}{1 + a}\left(\frac{b\gamma^4 + (a-1)\gamma^2 + 1}{\gamma^2(a + b)} - 1\right) = \frac{a}{a + b}\frac{(b\gamma^2 - 1)(\gamma^2 - 1)}{(a + 1)\gamma^2}.$$

The time along the X-segment is therefore given by

$$\frac{1}{\sqrt{a}}\sinh^{-1}\left(\sqrt{\frac{a(\gamma^2 - 1)(b\gamma^2 - 1)}{(a + b)(a + 1)\gamma^2}}\right).$$

The time along the Y-segment is computed analogously from equation (4.20). However, here the trajectory needs to be computed backward from the point $(\gamma, 0)$. Setting $\kappa = \zeta$ and $\beta = \gamma$ in equation (4.20), we obtain

$$\begin{aligned}\frac{b\gamma^2(\gamma^2 - \kappa^2)}{b\gamma^4 - 1} &= \frac{b\gamma^2}{b\gamma^4 - 1}\left(\gamma^2 - \frac{b\gamma^4 + (a-1)\gamma^2 + 1}{\gamma^2(a + b)}\right) \\ &= \frac{b}{a + b}\frac{(\gamma^2 - 1)(a\gamma^2 + 1)}{b\gamma^4 - 1}. \end{aligned} \tag{4.21}$$

Hence the time along the Y-portion is given by

$$\frac{1}{\sqrt{b}}\sin^{-1}\left(\sqrt{\frac{b(\gamma^2 - 1)(a\gamma^2 + 1)}{(a + b)(b\gamma^4 - 1)}}\right).$$

This verifies formula (4.13) and concludes the proof of the proposition. □

Figure 4.2 shows a representative graph of the function T_1.

We merely remark that abnormal extremals (i.e., the multiplier λ_0 is zero) do not exist for the problem. An analysis of the formal conditions for abnormal extremals shows that these must be Y-loops, i.e., closed curves for the constant control $u = b$. Since $\gamma > 1$, these curves cannot satisfy the terminal condition. We henceforth set $\lambda_0 = 1$. Given a normal extremal controlled trajectory (x, u), it follows for any switching time τ that $\lambda_1(\tau)x_2(\tau) = -1$. Hence XY junctions are only possible in $\mathscr{X}_+ = \{x_2 > 0\}$ while YX junctions lie in $\mathscr{X}_- = \{x_2 < 0\}$.

Next we develop the precise switching structures of optimal controls. It depends on the values of the control limits a and b and the formula T_1 for the XY-concatenations provides us with a baseline for the comparison of all these trajectories.

4.4 Switching Structure of Time-optimal Controlled Trajectories

We show that extremals which have more than one switching start and end with a Y-trajectory and explicitly compute the times between switchings. We rely on results from [19] where the same statement has been proven, but give alternate and simpler formulations that will then be used to parameterize these extremals.

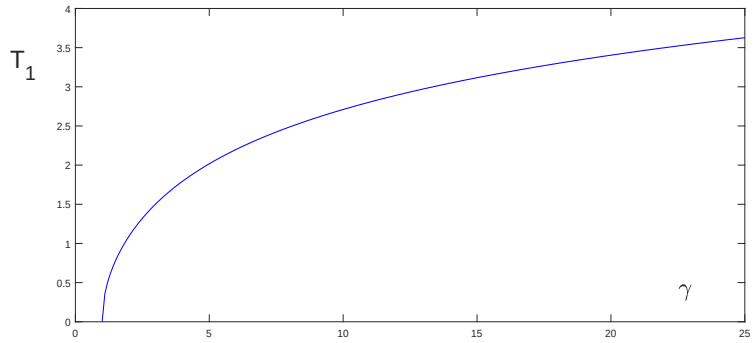


Fig. 4.2: The time $T_1 = T_1(\gamma)$ along the XY-trajectories that steer $(1,0)$ into $(\gamma,0)$ for $a = 1$ and $b = 8$.

The times between consecutive switchings along optimal controls are uniquely determined by specific relations called *conjugate point relations*. Suppose p and q are consecutive switching points of an extremal trajectory, $\overrightarrow{pq}$, and, without loss of generality, assume that the trajectory passes through p at time $t = 0$ and reaches the next switching point q at time τ. Then the vector $g(p)$ is parallel to (linearly dependent with) the vector v that is obtained by moving the vector $g(q)$ backward to the point p along the flow of the trajectory [20, 21]. The vector v is computed by integrating the variational equation of the dynamics backward along the trajectory from time $t = \tau$ to time $t = 0$ with the terminal condition given by $g(q)$ at time $t = \tau$ (e.g., see [2, 16]). The vector fields defining Ermakov's equation generate a finite-dimensional Lie algebra [19] and along the vector fields X and Y we can explicitly solve the equations defining this linear dependence for τ. Let s denote the slope of the line on which the first switching point p lies. We recall that XY-junctions lie in $\mathscr{X}_+$ so that $s > 0$ while YX-junctions lie in $\mathscr{X}_-$ and $s < 0$. It follows from Lemma 3.6 in [19] that for an X-trajectory we have $s < -\sqrt{a} < 0$ and

$$\sinh(2\sqrt{a}\tau) = -\frac{2\sqrt{a}s}{s^2 - a}, \qquad \cosh(2\sqrt{a}\tau) = \frac{s^2 + a}{s^2 - a}; \tag{4.22}$$

whereas for a Y-trajectory it holds that $s > 0$, $\frac{\pi}{2} < \sqrt{b}\tau < \pi$, and

$$\sin(2\sqrt{b}\tau) = -\frac{2\sqrt{b}s}{s^2 + b}, \qquad \cos(2\sqrt{b}\tau) = \frac{s^2 - b}{s^2 + b}. \tag{4.23}$$

Using basic trigonometric and hyperbolic identities, the following equivalent formulas are obtained:

Theorem 2. *Let $p = (x_1, x_2)$, $x_2 = sx_1$, be a switching point for an extremal trajectory and denote the time to reach the next switching point q by $\tau > 0$. If $\overrightarrow{pq}$ is an X-trajectory, then*

$$\sinh(\sqrt{a}\tau) = \sqrt{\frac{a}{s^2 - a}} \quad \textit{and} \quad \cosh(\sqrt{a}\tau) = -\sqrt{\frac{s^2}{s^2 - a}} = -\frac{s}{\sqrt{s^2 - a}}, \tag{4.24}$$

whereas, if $\overrightarrow{pq}$ is a Y-trajectory, then

$$\sin(\sqrt{b}\tau) = \sqrt{\frac{b}{s^2+b}} \quad \textit{and} \quad \cos(\sqrt{b}\tau) = -\sqrt{\frac{s^2}{s^2+b}} = -\frac{s}{\sqrt{s^2+b}}. \tag{4.25}$$

It is an important observation that the switching time τ until the next junction only depends on the slope s of the line $x_2 = sx_1$ on which the first switching point lies. This leads to a strict regime of switching points which allows us to give simple direct formulas for consecutive switching points.

Proposition 3 (consecutive switching points). *Let $\overrightarrow{pq}$ be two consecutive switching points along a Y-trajectory with coordinates $p = (\kappa, \mu)$ and $q = (\zeta, \xi)$ and set $s = \frac{\mu}{\kappa}$. Then it holds that*

$$\zeta = \frac{1}{\sqrt{\mu^2 + b\kappa^2}} \qquad \textit{and} \qquad \xi = -s\zeta = -\frac{\mu}{\kappa}\frac{1}{\sqrt{\mu^2 + b\kappa^2}}. \tag{4.26}$$

Furthermore, if $r = (\lambda, \nu)$ denotes the next switching point along an X-trajectory, then $s = -\frac{\xi}{\zeta}$ and we have that

$$\lambda = \frac{1}{\sqrt{\xi^2 - a\zeta^2}} \qquad \textit{and} \qquad \nu = s\lambda = -\frac{\xi}{\zeta}\frac{1}{\sqrt{\xi^2 - a\zeta^2}}. \tag{4.27}$$

Geometrically, in either case, if the first junction lies on the line $x_2 = sx_1$, then the next junction lies on the line $x_2 = -sx_1$ reflected around the x_1-axis.

Proof. Consider a XYX-trajectory with XY-junction at (κ, μ) and YX-junction at (ζ, ξ). It follows from equation (4.2) that the time-evolution of the x_1-coordinate of the Y-trajectory which starts at (κ, μ) at time $t = 0$ is governed by

$$x_1^2(t) = \left(\kappa\cos(\sqrt{b}t) + \frac{\mu}{\sqrt{b}}\sin(\sqrt{b}t)\right)^2 + \frac{1}{b\kappa^2}\sin^2(\sqrt{b}t). \tag{4.28}$$

Using the formulas for the switching time τ from Theorem 2 we obtain that

$$\begin{aligned}
\zeta^2 &= \kappa^2\cos^2(\sqrt{b}\tau) + 2\frac{\kappa\mu}{\sqrt{b}}\cos(\sqrt{b}\tau)\sin(\sqrt{b}\tau) + \frac{\mu^2}{b}\sin^2(\sqrt{b}\tau) + \frac{1}{b\kappa^2}\sin^2(\sqrt{b}\tau) \\
&= \kappa^2\frac{s^2}{s^2+b} - \frac{\kappa\mu}{\sqrt{b}}\frac{2\sqrt{b}s}{s^2+b} + \frac{\mu^2}{b}\frac{b}{s^2+b} + \frac{1}{b\kappa^2}\frac{b}{s^2+b} \\
&= \left\{\kappa^2 s^2 - 2\mu\kappa s + \mu^2 + \frac{1}{\kappa^2}\right\}\frac{1}{s^2+b} = \frac{1}{\kappa^2(s^2+b)} = \frac{1}{\mu^2 + b\kappa^2}.
\end{aligned}$$

The second relation in equation (4.26) follows from the fact that the two switching points lie on the same level curve for the Hamiltonian function $\mathscr{H}$ for the vector field Y. This gives us that

$$\xi^2 + b\zeta^2 + \frac{1}{\zeta^2} = \mu^2 + b\kappa^2 + \frac{1}{\kappa^2}.$$

As $\frac{1}{\zeta^2} = \mu^2 + b\kappa^2$, it follows that $\left(\frac{\xi}{\zeta}\right)^2 = \frac{1}{\zeta^2\kappa^2} - b = s^2 + b - b = s^2$. Since YX-junctions lie in $\mathscr{X}_-$ whereas XY-junctions lie in $\mathscr{X}_+$, we furthermore have that $\frac{\xi}{\zeta} = -s = \frac{\mu}{\kappa}$.

The computation for X-trajectories is analogous. The time-evolution of the x_1-coordinate of the X-trajectory which starts at (ζ, ξ) at time $t = 0$ is governed by

$$x_1^2(t) = \left(\zeta \cosh(\sqrt{a}t) + \frac{\xi}{\sqrt{a}} \sinh(\sqrt{a}t)\right)^2 + \frac{1}{a\zeta^2} \sinh^2(\sqrt{a}t). \tag{4.29}$$

Here we use we use the same parameter s as in the previous computation for Y, i.e., $s = \frac{\mu}{\kappa} = -\frac{\xi}{\zeta} > 0$. We point out, however, that the letter s in the formulas of Theorem 2 denotes the slope of the line through the first switching point. Hence we need to replace s in those formulas with $-s$. This only makes a difference in the formula for $\sinh(2\sqrt{a}\tau) = 2\sinh(\sqrt{a}\tau)\cosh(\sqrt{a}\tau)$ below. It thus follows that

$$\begin{aligned}
\lambda^2 &= \zeta^2 \cosh^2(\sqrt{a}\tau) + 2\frac{\zeta\xi}{\sqrt{a}} \cosh(\sqrt{a}\tau)\sinh(\sqrt{a}\tau) \\
&\qquad + \frac{\xi^2}{a} \sinh^2(\sqrt{a}\tau) + \frac{1}{a\zeta^2} \sinh^2(\sqrt{a}\tau) \\
&= \zeta^2 \frac{s^2}{s^2 - a} + \frac{\zeta\xi}{\sqrt{a}} \frac{2\sqrt{a}s}{s^2 - a} + \frac{\xi^2}{a} \frac{a}{s^2 - a} + \frac{1}{a\zeta^2} \frac{a}{s^2 - a} \\
&= \left\{\zeta^2 s^2 + 2\xi\zeta s + \xi^2 + \frac{1}{\zeta^2}\right\} \frac{1}{s^2 - a} = \frac{1}{\zeta^2(s^2 - a)} = \frac{1}{\xi^2 - a\zeta^2}
\end{aligned}$$

This verifies the first relation in equation (4.27). As above, the second one follows from the fact that the two switching points lie on the same level curve for the Hamiltonian function $\mathscr{H}$ for the vector field X. This gives us that $\nu^2 - a\lambda^2 + \frac{1}{\lambda^2} = \xi^2 - a\zeta^2 + \frac{1}{\zeta^2}$. Once again, we have $\frac{1}{\lambda^2} = \xi^2 - a\zeta^2$ and thus $\left(\frac{\nu}{\lambda}\right)^2 = \frac{1}{\lambda^2\zeta^2} + a = s^2 - a + a = s^2$. As above, since YX-junctions lie in $\mathscr{X}_-$ whereas XY-junctions lie in $\mathscr{X}_+$, it follows that $\frac{\xi}{\zeta} = -s = \frac{\nu}{\lambda}$. This completes the proof. □

Corollary 2. *Using the same notation as in Proposition 3 we have that*

$$\begin{pmatrix}\lambda \\ \nu\end{pmatrix} = \sqrt{\frac{\mu^2 + b\kappa^2}{\mu^2 - a\kappa^2}} \begin{pmatrix}\kappa \\ \mu\end{pmatrix} = \sqrt{\frac{s^2 + b}{s^2 - a}} \begin{pmatrix}\kappa \\ \mu\end{pmatrix}. \tag{4.30}$$

Similarly, if (ρ, σ) is the next YX-junction, then we have that

$$\begin{pmatrix}\rho \\ \sigma\end{pmatrix} = \sqrt{\frac{\xi^2 - a\zeta^2}{\xi^2 + b\zeta^2}} \begin{pmatrix}\zeta \\ \xi\end{pmatrix} = \sqrt{\frac{s^2 - a}{s^2 + b}} \begin{pmatrix}\zeta \\ \xi\end{pmatrix}. \tag{4.31}$$

In particular, these multiples are inverses of each other and only depend on the square of the slope s of the line $x_2 = sx_1$ on which the first switching point lies.

Proof. From the above formulas we get that

$$\lambda^2 = \frac{1}{\xi^2 - a\zeta^2} = \frac{1}{\zeta^2(s^2 - a)} = \frac{\mu^2 + b\kappa^2}{\left(\frac{\mu}{\kappa}\right)^2 - a} = \frac{\mu^2 + b\kappa^2}{\mu^2 - a\kappa^2}\kappa^2.$$

and

$$\rho^2 = \frac{1}{\nu^2 + b\lambda^2} = \frac{1}{\lambda^2(s^2 + b)} = \frac{\xi^2 - a\zeta^2}{\left(\frac{\xi}{\zeta}\right)^2 + b} = \frac{\xi^2 - a\zeta^2}{\xi^2 + b\zeta^2}\zeta^2.$$

Since both pairs of points lie on the same line, this proves the corollary. □

The following result then completely determines optimal controlled trajectories through their first switching point.

Proposition 4. *Consecutive switching points along an optimal controlled trajectory are not symmetric with respect to the x_1-axis.*

Remark. It is actually possible that extremal controlled trajectories can have consecutive switching points which are symmetric with respect to the x_1-axis along an X-segment. Figure 4.3 shows two such extremal YXY-loops. For this reason, it is not possible to exclude such configurations based on an analysis of the conditions of the maximum principle. Obviously, however, they cannot be optimal.

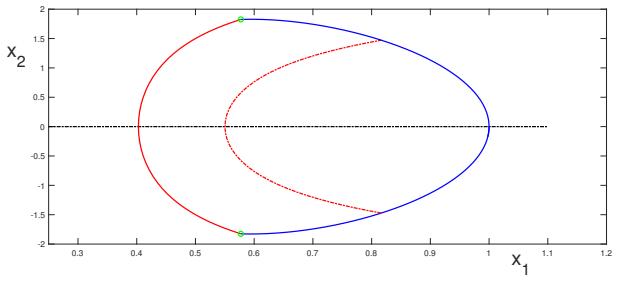


Fig. 4.3: Two examples of extremal YXY-trajectories that close at the initial condition $(1, 0)$ for $a = 1$ and $b = 8$.

Proof. Suppose an extremal controlled trajectory has an X-segment with entry junction at the point (ζ, ξ) and exit junction at (λ, ν) and suppose these junctions are symmetric with respect to the x_1-axis, i.e., $\zeta = \lambda$ and $\xi = -\nu$. Since both X- and Y-trajectories are symmetric with respect to the x_1-axis, it follows that the Y-trajectories through the two junctions (ζ, ξ) and (λ, ν) are the same. As YX-junctions can only lie in $\{x_2 < 0\}$, the forward Y-orbit from (λ, ν) extends at least to the next intersection with the x_1-axis, say $(\gamma, 0)$. Analogously, since XY-junctions can only lie in $\{x_2 > 0\}$, the backward Y-orbit from (ζ, ξ) also extends at least to the previous intersection with the x_1-axis. Since these two Y-orbits lie on the same Y-trajectory, this intersection is the point $(\gamma, 0)$. Hence the extremal controlled trajectory contains a loop starting and ending at $(\gamma, 0)$. This clearly is not time-optimal. The same argument (with obvious modifications in the terminology) applies to Y-junctions which would be symmetric with respect to the x_1-axis. □

Proposition 5. *Optimal controlled trajectories with more than one switching start and end with a Y-trajectory.*

Proof. If the controlled trajectory has more than one switching, then the trajectory has at least one XY and one YX-junction. Denote the coordinates of the first XY-junction by (κ, μ) and those of the first YX-junction by (ζ, ξ). We first note that $\zeta < 1$. This is clear if the trajectory starts with Y. If the trajectory starts with X, we have $\kappa > 1$ and thus also $\zeta = \frac{1}{\sqrt{\mu^2 + b\kappa^2}} < \frac{1}{\sqrt{b\kappa}} < 1$.

It follows from Corollary 2 that the x_1-coordinate is less then 1 for any subsequent YX-junction. Starting from any such point, the only point reachable on the x_1-axis along an X-trajectory lies even further to the left from the x_1-coordinate of the last junction. Hence no points $\gamma > 1$ are

reachable along optimal trajectories with a final X-segment. Thus optimal controlled trajectories end with a Y-segment.

If the optimal concatenation sequence starts with an X-segment, it thus contains at least the sequence $XYXY$. Denote the slope of the line through the origin and the first XY-junction (κ, μ) by s. The first switching point (κ, μ) lies on the X-orbit through $(1, 0)$ and, as the second X-arc has switching points at the beginning and end, it follows that $s^2 > a$ [19, Lemma 3.6]. Thus we have that

$$-a + 1 = \mu^2 - a\kappa^2 + \frac{1}{\kappa^2} = (s^2 - a)\kappa^2 + \frac{1}{\kappa^2} > 0 \tag{4.32}$$

contradicting $a \geq 1$. □

Combined with Proposition 3 this completely determines the structure of extremal controlled trajectories which have more than one switching in terms of their first switching point. Starting with a Y-arc, the first switching point (ζ, ξ) determines the slope of the line through the origin and this switching point. Henceforth, however, we always use s as the positive value, i.e., set $s = -\frac{\xi}{\zeta}$. After this initial junction the trajectory follows X until it meets the line through the origin with slope s for the second time in the non-symmetric point $(\kappa, \mu) \neq (\zeta, -\xi)$ where it switches back to Y. The first intersection corresponds to the symmetric switching point and is discarded. It then follows Y until it reaches the second intersection with the line through the origin with slope $-s$. Again the first intersection corresponds to the symmetric switching point and is discarded. Trajectories continue to iterate through YX- and XY-junctions always switching at the second intersection with the respective lines through the origin with slopes s, respectively $-s$.

Depending on the values for the control limits a and b and the terminal value γ, optimal controls indeed can have a large number of switchings. An intuitive understanding of these trajectories can be obtained by viewing the underlying dynamics as describing the motion of a unit mass particle with position x_1 and velocity x_2. Along a spiral trajectory with multiple loops, instead of moving directly to the target, the particle, although moving away from the target at first, comes close to $x_1 = 0$ where it acquires extra speed through the strong repulsive potential $\frac{1}{x_1^3}$ and thus reaches the target point faster. We start, however, with the following simple case.

Theorem 3. *If $v^2 \leq 4w \Leftrightarrow b \leq 1 + 2\sqrt{a}$, then the one switch XY-extremals are optimal.*

Proof. Any other extremal starts with a YXY-segment. Let (ζ, ξ) denote the first junction and denote by s the (positive) slope of the line through the origin and the XY-junction. We then have that

$$(s^2 + b)\zeta^2 + \frac{1}{\zeta^2} = b + 1. \tag{4.33}$$

Since the X-arc has switchings at the beginning and end, we have $s^2 > a$ and thus

$$\zeta_\pm^2(s^2) = \frac{b+1}{2(s^2+b)} \left\{ 1 \pm \sqrt{1 - \frac{4(s^2+b)}{(b+1)^2}} \right\}. \tag{4.34}$$

This equation has real solutions only if $4(s^2 + b) \leq (b + 1)^2$. In particular, if

$$4(a + b) \geq (b + 1)^2 \quad \Leftrightarrow \quad a \geq \frac{1}{4}(b - 1)^2 \quad \Leftrightarrow \quad b \leq 1 + 2\sqrt{a},$$

then there do not exist extremals that have more than one switch. The case $b = 1+2\sqrt{a}$ is degenerate in the sense that there does exist a unique real solution $\zeta^2 = \frac{b+1}{2(a+b)}$, but as we also have that $s^2 = a$,

it follows that the next switching lies at infinity and thus this does not give rise to an extremal of the control problem neither. Thus, under our assumptions on the control limits, the XY-trajectory that steers $(1,0)$ into $(\gamma,0)$ is the only extremal controlled trajectory. By Proposition 1 the optimal control problem does have a solution and thus this trajectory is globally optimal. □

4.5 Parameterized Families of Y-loops with n turns

Extremals with multiple switchings exist if $s^2 > a$ and $4(s^2+b) < (b+1)^2$. The line $x_2 = -sx_1$ with slope $s = \frac{1}{2}(b-1)$, equivalently $4(s^2+b) = (b+1)^2$, is tangent to the Y-trajectory through $(1,0)$ and forms the lower limit for possible intersections (see Figure 4.4). No intersections exist if $s > \frac{1}{2}(b-1)$. Overall, extremals with multiple switchings exist if and only if the slope satisfies $\sqrt{a} < s < \frac{1}{2}(b-1)$, i.e., if and only if $b > 1 + 2\sqrt{a}$ or, equivalently $4w < v^2$. We henceforth make this assumption.

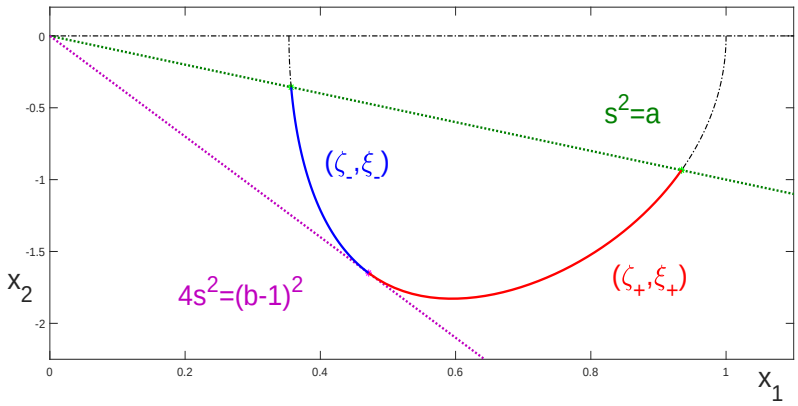


Fig. 4.4: The initial Y-segment with admissible first switching points (ζ,ξ) marked by red and blue colored segments. The colors distinguish between the points $\zeta_\pm^2 = \zeta_\pm^2(s^2)$ where the root is taken positive (red) and where it is taken negative (blue).

We analyze extremal Y-loops by setting up *parameterized families of extremals* [16] and calculating the transfer times for all candidate optimal trajectories. Such trajectories start with a YXY-segment and we denote the first YX-junction by (ζ,ξ) and the first XY-junction by (κ,μ). We write $s = \frac{\mu}{\kappa} = -\frac{\xi}{\zeta}$ for the positive slope of the line on which the XY-junctions lie. The first switching point lies on the segment of the Y-trajectory which is bounded by its two points of intersection with the line with slope $s = -\sqrt{a}$. Solving equation (4.34) for $s^2 = a$ gives us the limits

$$\zeta_{a,\pm}^2 = \frac{v \pm \sqrt{v^2 - 4w}}{2w} = \frac{v}{2w}\left\{1 \pm \sqrt{1 - \frac{4w}{v^2}}\right\}.$$

The curve $\mathscr{S}_1$ where the first switching points can lie is therefore given by

$$\mathscr{S}_1 = \left\{(\zeta,\xi) = \zeta(1,-s):\ \zeta_{a,-} < \zeta < \zeta_{a,+} \Leftrightarrow \sqrt{a} < s \le \frac{1}{2}(b-1)\right\}$$

The geometric set-up is illustrated in Figure 4.4.

It is convenient to introduce the variable $z = \zeta^2 = x_1^2$ and equation (4.33) defines s^2 as a function of the x_1-coordinate in the form

$$s^2 = \left(\frac{1}{\zeta^2} - 1\right)\left(b - \frac{1}{\zeta^2}\right) = \frac{(1-z)(bz-1)}{z^2}, \qquad \zeta_{a,-}^2 \le z \le \zeta_{a,+}^2. \tag{4.35}$$

The solutions $\zeta_\pm^2(s^2)$ to equation (4.34) are equal for $s^2 = \frac{1}{4}(b-1)^2$, i.e., when the line with slope $-s$ is tangent to the Y-trajectory through $(1,0)$. The point of tangency is given for $\breve{z} = \frac{2}{v}$. We also note the following basic relations about the algebraic and geometric means of the interval limits $\zeta_{a,\pm}^2$ and the parameters v and w (for $v^2 > 4w$):

$$p_{\text{in}} \ < \ p_{\text{alg}} \ < \ p_{\text{geo}} \ < \ \breve{p} \ < \quad p_{\text{fin}}. \tag{4.36}$$

while the reverse inequalities hold for their associated z-values:

$$\zeta_{a,-}^2 \ < \ \breve{z} = \frac{2}{v} \ < \ z_{\text{geo}} = \frac{1}{\sqrt{w}} \ < \ z_{\text{alg}} = \frac{v}{2w} \ < \ \zeta_{a,+}^2. \tag{4.37}$$

We parameterize extremal Y-loops through the time p when the first switching along the initial Y-trajectory occurs. It follows from Lemma 1 that

$$\zeta^2 = \cos^2(\sqrt{b}t) + \frac{1}{b}\sin^2(\sqrt{b}t) \qquad \Leftrightarrow \qquad \sin^2(\sqrt{b}t) = \frac{b}{b-1}(1-\zeta^2).$$

Hence the times until the points $(\zeta_{a,\pm}, \xi_{a,\pm})$ on the line $s = -\sqrt{a}$ are reached are given by

$$p_{\text{in}} = \frac{1}{\sqrt{b}}\sin^{-1}\left(\sqrt{\frac{b(1-\zeta_{a,+}^2)}{b-1}}\right) \ < \ p_{\text{fin}} = \frac{1}{\sqrt{b}}\sin^{-1}\left(\sqrt{\frac{b(1-\zeta_{a,-}^2)}{b-1}}\right).$$

The Y-trajectory starting at $(1,0)$ reaches the x_1-axis again in the point $\frac{1}{\sqrt{b}}$ and thus, by Corollary 1, the time to make the semi-loop from $(1,0)$ to the next intersection with the x_1-axis is $\frac{\pi}{2\sqrt{b}}$. Hence the inverse sine is taken along its principal branch and $P = (p_{\text{in}}, p_{\text{fin}}) \subset \left(0, \frac{\pi}{2}\right)$.

Denote the domain of the parameterization by D,

$$D = \{(t,p) : t \ge 0, p \in P\}.$$

It follows from the results in Section 4.4 that the parameter p determines all successive switching points and switching times. We denote the time when the i-th switching occurs by $\tau_i(p)$ and the formulas for the inter-switching times derived in Section 4.4 determine the times $\tau_i = \tau_i(p)$, $i = 2, 3, \ldots$. We formally also write $\tau_0(p) \equiv 0$ and divide the domain D into the subdomains

$$D_i = \{(t,p) \in D : \tau_{i-1}(p) < t \le \tau_i(p)\}, \quad i = 1, 2, \ldots.$$

The control $u(t,p)$ is given by b if $(t,p) \in D_{2i-1}$ and by $-a$ if $(t,p) \in D_{2i}$ and the corresponding trajectories $x = x(t,p)$ are obtained through integration of the dynamics. Explicit formulas for the

controlled trajectories could be given from the formulas for the solutions to Ermakov's equation, but these are neither required nor helpful in the analysis. We set $\lambda_0(p) \equiv 1$ and the multiplier $\lambda = \lambda(t,p)$ is defined as the solution of the associated adjoint equation which satisfies $\lambda_2(p,p) = 0$ and $\lambda_1(p,p) = -\frac{1}{x_2(p,p)}$. The condition on λ_2 simply specifies that there is a switching at time p and the condition on λ_1 enforces that $H \equiv 0$ along the controlled trajectories. This defines a parameterized family of extremals (also in the sense of the definition given in [16]).

The controlled trajectories $(x(\cdot,p), u(\cdot,p))$ defined here for $p \in P$ are extremals for the optimal control problem to steer the initial point $(1,0)$ into a terminal point $(\gamma,0)$, but without restrictions on γ. Indeed, for some values of the control limits, some YXY-extremals steer the initial point into a terminal point with $\gamma < 1$ and thus are not of interest for the time-optimal frictionless atom cooling problem. We shall restrict the family to those extremals that steer the system into a terminal point with value $\gamma > 1$ at the appropriate later time.

Let $\mathscr{E}_{2n}$ denote the parameterized family where the trajectories $x = x(\cdot,p)$ are terminated when they cross the x_1-axis for the n-th time. Thus $\mathscr{E}_{2n}$ consists of all extremal Y-loops with n turns. The last sub-domain then is given by

$$\widehat{D}_{2n} = \left\{(t,p) \in D : \tau_{2n}(p) < t \le \widehat{T}_{2n}(p)\right\}$$

with $\widehat{T}_{2n}(p)$ denoting the time when the extremal reaches the x_1-axis for the n-th time. These times are easily computed. An extremal Y-loop with n turns has $2n$ switchings and consists of $n+1$ Y-arcs and n X-arcs. The time $\tau_{\text{in}}(p)$ along the first Y-segment is the parameter, i.e., $\tau_{\text{in}}(p) = p$. It follows from Theorem 2 that the times along intermediate X- and Y-segments (i.e., between two switching points) are given by

$$\tau_X(p) = \frac{1}{\sqrt{a}} \sinh^{-1}\left(\sqrt{\frac{a}{s^2-a}}\right) \tag{4.38}$$

and

$$\tau_Y(p) = \frac{1}{\sqrt{b}}\left(\pi - \sin^{-1}\left(\sqrt{\frac{b}{s^2+b}}\right)\right). \tag{4.39}$$

These times only depend on the slope s and thus the times along different intermediate X-, respectively Y-segments, are all equal for a particular extremal. We recall that the time τ_Y along an intermediate Y-segment satisfies $\tau_Y > \frac{\pi}{2\sqrt{b}}$ which brings in the π when solving for $\tau_Y(p)$.

Computing the time $\tau_{\text{fin}}(p)$ along the final Y-arc is more involved. We denote the first and second switching points by $(\zeta(p),\xi(p))$ and $(\kappa(p),\mu(p))$, respectively, and write $s(p) = -\frac{\xi(p)}{\zeta(p)} = \frac{\mu(p)}{\kappa(p)}$ for the positive slope defined by these points. Then the switching curves $\mathscr{S}_1$ and $\mathscr{S}_2$ where the first, respectively second switching occur are given by

$$\mathscr{S}_1 = \{(\zeta,\xi) = (\zeta(p),\xi(p)) : p \in P\}$$

and

$$\mathscr{S}_2 = \{(\kappa,\mu) = (\kappa(p),\mu(p)) : p \in P\}.$$

We denote consecutive YX-junctions by $(\zeta_i,\xi_i) = (\zeta_i(p),\xi_i(p))$, $i \in \mathbb{N}$, and consecutive XY-junctions by $(\kappa_i,\mu_i) = (\kappa_i(p),\mu_i(p))$, $i \in \mathbb{N}$. It follows from Proposition 3 that the following relations hold between these points:

$$\kappa_i\zeta_i = \frac{1}{\sqrt{s^2-a}} \qquad \text{and} \qquad \mu_i = s\kappa_i, \tag{4.40}$$

$$\zeta_{i+1}\kappa_i = \frac{1}{\sqrt{s^2+b}} \qquad \text{and} \qquad \xi_{i+1} = -s\zeta_{i+1}. \tag{4.41}$$

The terminal point $\gamma_n = \gamma_n(p)$ which is reached by the trajectory $x = x(\cdot, p)$ after n loops is the solution of the equation that defines the next crossing of the x_1-axis after the n-th XY-junction. The defining equation now takes the form

$$b\gamma_n^2 + \frac{1}{\gamma_n^2} = (s^2+b)\kappa_n^2 + \frac{1}{\kappa_n^2} = (s^2-a)\zeta_n^2 + (s^2+b)\kappa_n^2 \tag{4.42}$$

and γ_n is the solution that satisfies $\gamma_n > \kappa_n$. Using Corollary 1, the time $\tau_{\text{fin}}(p)$ along the final Y-segment, $\tau_{\text{fin}}(p) = \widehat{T}_{2n}(p) - \tau_{2n}(p)$, is given by the same expression as in equation (4.21), but with γ_n instead of γ:

$$\tau_{\text{fin}}(p) = \frac{1}{\sqrt{b}} \sin^{-1}\left(\sqrt{\frac{b\gamma_n^2(\gamma_n^2-\kappa_n^2)}{b\gamma_n^4-1}}\right). \tag{4.43}$$

Overall, we therefore have the following formulas:

Proposition 6. *The total time along the parameterized Y-loop with n turns for parameter p is given by*

$$\begin{aligned}\widehat{T}_{2n}(p) &= \tau_{\text{in}}(p) + n\tau_X(p) + (n-1)\tau_Y(p) + \tau_{\text{fin}}(p)\\ &= p + \frac{n}{\sqrt{a}} \sinh^{-1}\left(\sqrt{\frac{a}{s^2-a}}\right) + \frac{n-1}{\sqrt{b}}\left(\pi - \sin^{-1}\left(\sqrt{\frac{b}{s^2+b}}\right)\right)\\ &\qquad + \frac{1}{\sqrt{b}} \sin^{-1}\left(\sqrt{\frac{b\gamma_n^2(\gamma_n^2-\kappa_n^2)}{b\gamma_n^4-1}}\right).\end{aligned} \tag{4.44}$$

Using these formulas it is straightforward to compute the optimal number of switchings numerically. The parameterized curves $p \mapsto (\gamma_n(p), \widehat{T}_{2n}(p))$ can be computed recursively using equations (4.40) and (4.41) and then solving for γ_n. Cut-loci can simply be read off by plotting these curves in (γ, T)-space.

Figure 4.5 shows the times T_1, T_2, T_4 and T_6 for the control limits $a = 1$ and various values of b. For $b = 8$ only one cut-locus $\widehat{\gamma}_1$ exists while there are infinitely many for $b = 20$ and $b = 50$ as the number of loops is increased. For small terminal values γ the one-switching strategy XY is optimal and starting with the value $\widehat{\gamma}_1$ for the first cut-locus, the one-turn YXY-strategy is faster. In the first example shown in Figure 4.5 this then always is the optimal solution while there exist further cut-loci $\widehat{\gamma}_2$ and $\widehat{\gamma}_4$ as the number of loops is increased in the other two examples. Numerical values are given in Table 4.1. The one switch trajectory XY is optimal for terminal values $1 < \gamma \le \widehat{\gamma}_1$, the YXY-extremals are optimal for $\widehat{\gamma}_1 \le \gamma \le \widehat{\gamma}_2$, $YXYXY$-extremals are optimal for $\widehat{\gamma}_2 \le \gamma \le \widehat{\gamma}_4$, and for $\gamma \ge \widehat{\gamma}_4$, Y-loops with three turns are the best among the extremal controlled trajectories shown in this range.

4.6 Geometric Properties of the Parameterised Switching Curves

We state (without proofs) some relevant properties of the switching curves. It follows from Proposition 3 and Corollary 2 that the switching curve $\mathscr{S}_i$, $i \ge 1$, are images of the first switching curve

b	$\varpi(1,b)$	$\widehat{\varpi}(1,b)$	optimal structure	$\widehat{\gamma}_1$	$\widehat{\gamma}_2$	$\widehat{\gamma}_4$
8	-0.24	0.27	XY and YXY	3.44	–	–
20	-0.79	-0.43	cut-loci	1.51	17.21	119.01
50	-1.39	-1.03	cut-loci	1.19	9.59	49.76

Table 4.1: Some numerical values.

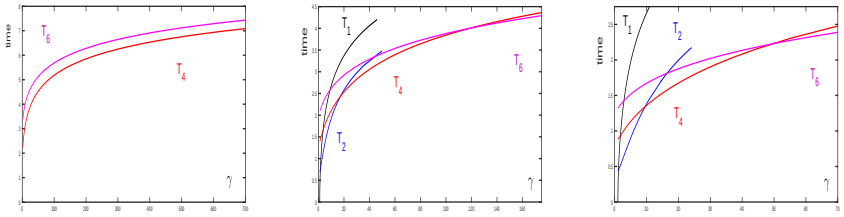


Fig. 4.5: Graphs of the function $\gamma \mapsto T_1(\gamma)$ for the XY-trajectory and $\gamma \mapsto T_{2i}(\gamma)$, $i = 1, 2, 3$ corresponding to Y-loops with 1, 2 and 3 turns for $a = 1$ and some values of b. The left panel shows the graphs of the functions T_4 and T_6 for $b = 8$ and no cut-locus exists for $\gamma \leq 700$. The middle ($b = 20$) and right panel ($b = 50$) show cases with infinitely many cut-oci as $\gamma \to \infty$.

$\mathscr{S}_1$:

$$\mathscr{S}_{2i+1} = \left\{ (\zeta_{i+1}, \xi_{i+1}) = \left(\sqrt{\frac{s^2 - a}{s^2 + b}} \right)^i (\zeta, \xi) : \ (\zeta, \xi) \in \mathscr{S}_1 \right\} \tag{4.45}$$

and

$$\mathscr{S}_{2i} = \left\{ (\kappa_i, \mu_i) = \left(\sqrt{\frac{s^2 + b}{s^2 - a}} \right)^{i-1} \frac{(1, s)}{\zeta \sqrt{s^2 - a}} : \ (\zeta, \xi) \in \mathscr{S}_1 \right\} \tag{4.46}$$

All switching curves $\mathscr{S}_i$, $i \geq 1$, are embedded 1-dimensional manifolds, i.e., have no self-intersections. Figure 4.6 shows the second switching curve $\mathscr{S}_2$ and illustrates the flow of X-trajectories between the first and second switching curves. After the switching, Y-trajectories come down to the x_1-axis and thus it is clear that the mapping $p \mapsto \gamma(p)$ is $2:1$ for large γ. Figure 4.7 shows two examples of the third switching curve $\mathscr{S}_3$. The geometric shapes seen in these figures are characteristic for all even, respectively odd switching curves.

The following theorem summarizes the geometric properties of the flow of X-, respectively Y-trajectories between the switching curves. Its proof consists of rather lengthy, but explicit calculations. As it turns out that conjugate points play no role in the structure of the globally optimal trajectories, we merely state this result.

Theorem 4. *For each $i \geq 1$, $i \in \mathbb{N}$, there exists a unique parameter $\tilde{p}_{2i+1}$ for which the vector field X is tangent to the switching curve $\mathscr{S}_{2i+1}$ and a unique parameter $\tilde{p}_{2i}$ where the vector field Y is tangent to the switching curve $\mathscr{S}_{2i}$. The sequence of parameters $\{\tilde{p}_i\}_{i \geq 2}$ is strictly monotonically increasing and bounded above by $\breve{p}$, the parameter value where the two solutions $\zeta_{\pm}^2$ agree. For all points on $\mathscr{S}_i$ corresponding to parameters $p > \tilde{p}_i$ X- and Y-trajectories cross the switching*

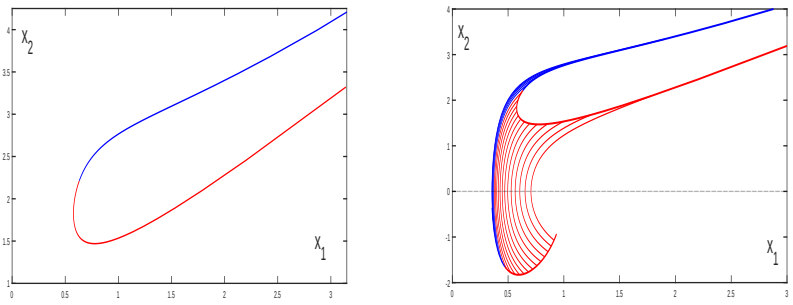


Fig. 4.6: The switching curve $\mathscr{S}_2$ for the XY-junctions for $a = 1$ and $b = 8$ (left) and the flow F from the first to the second switching curve along X-trajectories. Red curves correspond to trajectories starting at points ζ_+^2 where the root is taken with the positive sign and the blue curves correspond to trajectories starting at points ζ_-^2 where the root is taken with the negative sign.

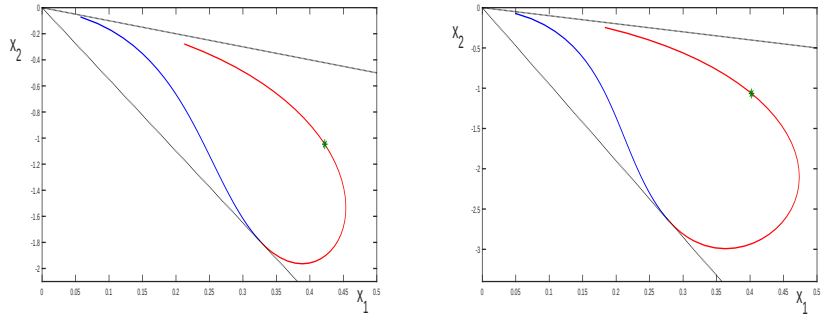


Fig. 4.7: The third switching curve $\mathscr{S}_3$ for $a = 1$ and $b = 12$ (left) and for $a = 1$ and $b = 20$ (right). Also shown (as black lines) are the line $s = -\frac{1}{2}(b-1)$ and $s = -\sqrt{a}$ which gives the asymptotic behavior of the switching curve as $p \to p_{\text{in}}$ and $p \to p_{\text{fin}}$. All switching curves $\mathscr{S}_{2i+1}$, $p \in P$, lie in the sector defined by these two lines and every line with slope $-s$, $\sqrt{a} < s < \frac{1}{2}(b-1)$, intersects $\mathscr{S}_{2i+1}$ in exactly two points.

curve transversally and local optimality properties are preserved ('transversal crossings' [16]). For all parameters p in the open interval $(\tilde{p}_{i-1}, \tilde{p}_i)$ X- and Y-trajectories point to the same side of the switching curve ('transversal fold' [16]). In this case the switching curve is an envelope for the control problem and the switching points are conjugate points where optimality ceases. Switching points defined by parameter values $p > \breve{p}$ (where the root is taken with the negative sign) are always transversal crossings. □

4.7 The First Cut-locus

We analyze the difference in the time T_1 along the one switch XY-extremals and the time $\widehat{T}_2$ along the YXY-extremals in the parameterized family $\mathscr{E}_2$. These times have already been computed, but while T_1 is given as a function of the terminal condition γ, $\widehat{T}_2$ is only known as a function of the parameter p. Once we restrict the parameters to $p \geq \tilde{p} = \tilde{p}_2$, the parameter where the vector field Y is tangent to the second switching curve $\mathscr{S}_2$ (c.f., Theorem 4), then the correspondence between the parameters p and the terminal values $\gamma \geq \tilde{\gamma}$ is 1:1. We write

$$\gamma_1 : [\tilde{p}, p_{\text{fin}}) \to [\tilde{\gamma}, \infty)\,,\ p \mapsto \gamma_1(p) \quad \text{and} \quad \pi_1 : [\tilde{\gamma}, \infty) \to [\tilde{p}, p_{\text{fin}})\,,\ \gamma \mapsto \pi_1(\gamma)$$

for the corresponding inverse mapping with the subscript 1 indicating that the YXY-extremal makes one loop. Thus, for a given $\gamma \geq \tilde{\gamma}$, $\pi_1(\gamma)$ is the parameter p such that $\gamma_1(p) = \gamma$. For $p \geq \tilde{p}$ we thus have $T_2(\gamma) = \widehat{T}_2\,(\pi_1(\gamma))$. Analogously, $\widehat{T}_1 = T_1 \circ \gamma_1$. We also define the time-difference ΔT as a function of the terminal condition γ, i.e.,

$$\Delta T(\gamma) = T_2(\gamma) - T_1(\gamma) = (\widehat{T}_2 \circ \pi_1)(\gamma) - T_1(\gamma),$$

and $\Delta\widehat{T} = \Delta T \circ \gamma_1 = \widehat{T}_2 - \widehat{T}_1$ for the time-difference as a function of the parameter p.

We recall from equation (4.44) that

$$\widehat{T}_2(p) = p + \frac{1}{\sqrt{a}} \sinh^{-1}\left(\sqrt{\frac{a}{s^2-a}}\right) + \frac{1}{\sqrt{b}} \sin^{-1}\left(\sqrt{\frac{b\gamma^4}{b\gamma^4-1} \cdot \frac{\gamma^2-\kappa^2}{\gamma^2}}\right)$$

with s, κ and γ functions of p while equation (4.13) gives us that

$$\begin{aligned} T_1(\gamma) = \frac{1}{\sqrt{a}} \sinh^{-1}&\left(\sqrt{\frac{a(\gamma^2-1)}{a+b} \cdot \frac{b\gamma^2-1}{(a+1)\gamma^2}}\right) \\ &+ \frac{1}{\sqrt{b}} \sin^{-1}\left(\sqrt{\frac{(a\gamma^2+1)}{a+b} \cdot \frac{b(\gamma^2-1)}{b\gamma^4-1}}\right). \end{aligned}$$

We also write $\widehat{\sigma}_X = \sigma_X \circ \gamma_1$ and $\widehat{\sigma}_Y = \sigma_Y \circ \gamma_1$ for the corresponding expressions as functions of the parameter.

Figure 4.8 compares the times T_1 and T_2 for $a = 1$ and $b = 8$. For small values of γ the one-switch XY-extremal is faster whereas the YXY-trajectories are faster for larger values of γ. While the first assertion is unconditionally true, the latter one need not always hold, but depends on relations between the control limits a and b. However, there exists at most one point of intersection. The corresponding value $\widehat{\gamma}_1$ is called the first cut-locus.

Lemma 2. *For γ close enough to the initial condition $\gamma = 1$, the YXY-trajectory that steers $(1, 0)$ into $(\gamma, 0)$ takes longer than the one switch XY-trajectory that does the same. The XY-trajectories are optimal for $\gamma > 1$ near $\gamma = 1$.* □

Proof. If $v^2 \geq 8w$, it can be shown that there exists a unique parameter $\bar{p} \in P$ for which the extremal YXY-loop returns to the initial point $\gamma = 1$, ($\gamma_1(\bar{p}) = 1$) and thus $\widehat{T}_2(\bar{p}) > 0 = \widehat{T}_1(\bar{p}) = T_1(1)$. For γ close enough to $\gamma = 1$ the XY-trajectory will therefore simply be faster by continuity.

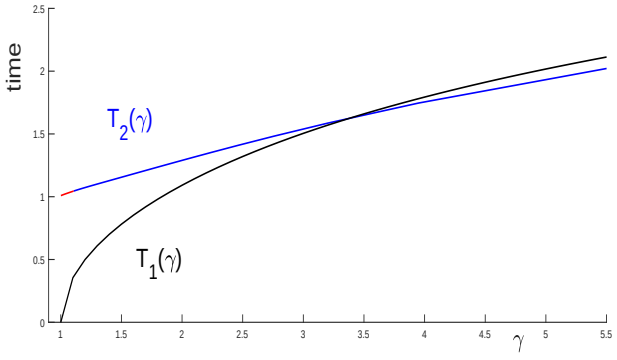


Fig. 4.8: The total times $T_1 = T_1(\gamma)$ (shown in black) for the family of XY-extremals and $T_2 = T_2(\gamma)$ (shown in red/blue) for the YXY-extremals in the family $\mathscr{E}_2$ as a function of the terminal value γ, $\gamma \geq \bar{\gamma}$, for $a = 1$ and $b = 8$.

On the other hand, if $v^2 < 8w$, only terminal points with $\gamma \geq \tilde{\gamma} > 1$ are reachable with YXY-extremals. As the minimum time frictionless atom cooling problem has a solution, and since the XY-trajectory is the only extremal which reaches γ for $\gamma < \bar{\gamma}$, the one switch control is the optimal control. □

Theorem 5. *The time difference $\Delta T = T_2 - T_1$ is strictly decreasing for $\gamma \geq \tilde{\gamma}$.*

Proof. The proof is analytical and we use the clock-form $\omega = \frac{dx_1}{x_2}$ and Stokes's theorem to evaluate the difference in the time between the XY- and YXY-trajectories that steer $(1,0)$ into the same terminal point $(\gamma, 0)$. Along the YXY-trajectory, however, x_2 is also negative and crosses the x_1-axis while the clock-form has a singularity at $x_2 = 0$. This precludes a straightforward application of Stokes's theorem and we need to work around the singularity. The easiest way is to use symmetries to avoid the singularity all together. This is possible since for both X- and Y-trajectories the times along segments that are symmetric to the x_1-axis are equal. This is a consequence of the explicit descriptions of solutions for X- and Y-trajectories that were derived in Lemma 1 and the corresponding formulas for the times along such segments.

Given $\gamma \geq \tilde{\gamma}$, as before, we denote the coordinates of the YX-junction of the YXY-trajectory by (ζ, ξ) and we let $j = (\zeta, -\xi)$ be the symmetric image of this point about the x_1-axis. Also, let $r = (\kappa, \mu)$ be the XY-junction of the YXY-trajectory and let s denote the point where the XY-junction of the one switch XY-trajectory occurs. Since both trajectories have the same terminal point they agree along the Y-segment of the XY-trajectory, i.e., from s onward. The concatenation of the YXY-trajectory from $(1,0)$ into s with the backward X-segment of the XY-trajectory forms a closed curve Γ. Let $\breve{\Gamma}$ denote the curve that is obtained from Γ by reflecting the portions of the curve that lie in $\mathscr{X}_- = \{x_2 < 0\}$ into $\mathscr{X}_+ = \{x_2 > 0\}$ around the x_1-axis (see Figure 4.9). The portion of the X-segment of the YXY-trajectory that connects (ζ, ξ) with $j = (\zeta, -\xi)$ is symmetric with respect to the x_1-axis and we cancel it in this transformation. We orient $\breve{\Gamma}$ clockwise (mathematically negatively), i.e., as the concatenation of the X-segment from j to r, followed by the Y-segment from r to s, the backward X-trajectory from s to the initial point $(1,0)$ and the

backward Y-trajectory from $(1,0)$ to j. The last piece is the reflected initial Y-segment of the YXY-trajectory from $(1,0)$ to (ζ,ξ).

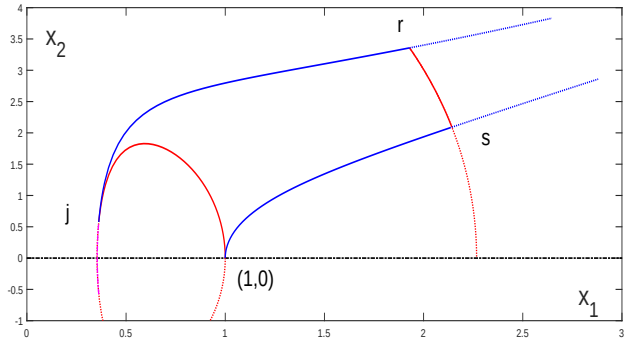


Fig. 4.9: The region R measuring the time-difference $\Delta T = T_2 - T_1$.

Writing $\tau_{Z:x\to y}$ for the time to go from x to y along a Z-trajectory and using the clock form ω, we can express the time difference ΔT in the following equivalent form:

$$\begin{aligned}\Delta T &= \tau_{Y:(1,0)\to(\zeta,\xi)} + \tau_{X:(\zeta,\xi)\to(\zeta,-\xi)=j} + \tau_{X:j\to r} + \tau_{Y:r\to s} - \tau_{X:(1,0)\to s}\\ &= \tau_{X:(\zeta,\xi)\to(\zeta,-\xi)} + \left\{\tau_{X:j\to r} + \tau_{Y:r\to s} - \tau_{X:(1,0)\to s} - \tau_{Y:(1,0)\to(\zeta,-\xi)}\right\}\\ &= \tau_{X:(\zeta,\xi)\to(\zeta,-\xi)} + \int_{\breve{\Gamma}} \omega. \end{aligned} \tag{4.47}$$

We show that both of these expressions are strictly monotonically decreasing in γ. For the first term, $\tau_{X:(\zeta,\xi)\to(\zeta,-\xi)}$, this is an explicit computation that we postpone briefly until after the second term has been dealt with.

Let $R = R(\gamma)$, $\gamma \geq \bar{\gamma} \geq \tilde{\gamma}$, denote the region enclosed by the closed curve $\breve{\Gamma}$. Since this curve is oriented negatively (clockwise), it follows from Stokes's theorem that we have that

$$\int_{\breve{\Gamma}} \omega = -\int_R d\omega = -\int_R \left(-\frac{1}{x_2^2}\right) dx_2 \wedge dx_1 = -\int_R \frac{dx_1 dx_2}{x_2^2} < 0. \tag{4.48}$$

The region R contains the point $(1,0)$ where ω is singular. This can easily be avoided: instead of using the X-trajectory from $(1,0)$ to s, cut off the singularity near $(1,0)$ at the point on the Y trajectory that connects $(\zeta,-\xi)$ to $(1,0)$ which has x_2-coordinate $x_2 = \varepsilon$. Starting from that point integrate X until this trajectory meets the second Y-leg of the YXY-trajectory and thus close the loop this way. This curve avoids the singularity and Stokes's theorem is applicable. As we take the limit $\varepsilon \to 0$, the times along the respective Y and X-segments converge and remain finite. Hence $\int_{\breve{\Gamma}} \omega$ is well-defined. So is then $\int_R \frac{dx}{x_2^2}$ in the limit.

It follows from the geometric properties of the X and Y-trajectories that the region R increases monotonically: if $\gamma_1 < \gamma_2$, then $R(\gamma_1) \subsetneqq R(\gamma_2)$. For, the parameters associated with these γ-values

satisfy $p_1 = \pi(\gamma_1) < \pi(\gamma_2) = p_2$ and thus the X-portion of the trajectory $x(\cdot, p_2)$ lies to the left of the X-portion of $x(\cdot, p_1)$. This increases the region R near that boundary curve. Furthermore, for $p \geq \tilde{p}$ we also have that the XY-junction (κ_2, μ_2) for p_2 lies further to the right of the XY-junction (κ_1, μ_1) for p_1 so that the Y-trajectory emanating from (κ_2, μ_2) does not intersect the X-portion for parameter p_1. This precisely is the meaning of the fact that switching points for parameters $p > \tilde{p}$ are *transversal crossings*. Hence the region $R(\gamma)$ is strictly increasing for $\gamma \geq \tilde{\gamma}$. As the negative function $-\frac{1}{x_2^2}$ thus is integrated over a larger set, the integral $\int_{\check{\Gamma}} \omega$ strictly decreases in γ.

It remains to analyze the first term. Since the initial condition (ζ, ξ) lies on the Y-trajectory through $(1, 0)$, here it is advantageous to parameterize the function by p, the time along this trajectory. We write $(\zeta, \xi) = (\zeta(p), \xi(p))$, $0 \leq p \leq p_{\text{fin}}$, and use the representation of solutions for X-trajectories to compute the time

$$\theta : [0, p_{\text{fin}}) \to [0, \infty), \qquad p \mapsto \theta(p) = \tau_{X:(\zeta(p),\xi(p))\to(\zeta(p),-\xi(p))}. \tag{4.49}$$

Clearly, $\theta(0) = 0$ and θ is positive otherwise. While θ thus increases near $p = 0$, from a certain parameter onward, θ is strictly decreasing. Intuitively, as p increases, the speed along the X-trajectory increases and eventually this compensates for the initial increase in the length of the integral curve of X from $(\zeta(p), \xi(p))$ to $(\zeta(p), -\xi(p))$. Once the minimum point for the initial Y-trajectory has been passed, this curve even decreases in length leading to a strong decrease in this function as shorter segments are traversed at increased speeds.

Lemma 3. *The function θ is strictly decreasing for $p \geq p_{\text{qu}} = \frac{\pi}{4\sqrt{b}}$. (The parameter/time p_{qu} corresponds to completing a quarter of the full Y-loop measured in the total time along the trajectory.)*

Proof. The time-evolution of the x_1-coordinate of the X-trajectory which starts at (ζ, ξ) at time $t = 0$ is governed by (see equation (4.29))

$$x_1^2(t) = \left(\zeta \cosh(\sqrt{a}t) + \frac{\xi}{\sqrt{a}} \sinh(\sqrt{a}t)\right)^2 + \frac{1}{a\zeta^2} \sinh^2(\sqrt{a}t).$$

We therefore have that

$$\begin{aligned}\zeta^2 = \zeta^2 \cosh^2(\sqrt{a}\theta) + 2\frac{\zeta\xi}{\sqrt{a}} \cosh(\sqrt{a}\theta) \sinh(\sqrt{a}\theta) \\ + \frac{\xi^2}{a} \sinh^2(\sqrt{a}\theta) + \frac{1}{a\zeta^2} \sinh^2(\sqrt{a}\theta)\end{aligned}$$

which gives us that

$$\left(\zeta^2 + \frac{\xi^2}{a} + \frac{1}{a\zeta^2}\right) \sinh^2(\sqrt{a}\theta) = -2\frac{\zeta\xi}{\sqrt{a}} \cosh(\sqrt{a}\tau) \sinh(\sqrt{a}\theta)$$

and thus

$$\tanh(\sqrt{a}\theta) = -\frac{2\sqrt{a}\zeta\xi}{a\zeta^2 + \xi^2 + \frac{1}{\zeta^2}}. \tag{4.50}$$

The points $(\zeta, \xi) = (\zeta(p), \xi(p))$ lie on the Y-trajectory through $(1, 0)$ and the parameter p represents the time along this trajectory. Using the standard coordinates (x_1, x_2) and $s = -\frac{x_2}{x_1}$ we rewrite this expression as

$$\tanh(\sqrt{a}\theta) = -\frac{2\sqrt{a}x_1x_2}{ax_1^2 + x_2^2 + \frac{1}{x_1^2}} = \frac{2\sqrt{a}s}{s^2 + a + \frac{1}{x_1^4}}$$

and differentiating this equation gives us that

$$\frac{\sqrt{a}\theta'(p)}{\cosh^2(\sqrt{a}\tau)} = 2\sqrt{a}\frac{\Phi(x_1,s)}{\left(s^2 + a + \frac{1}{x_1^4}\right)^2}$$

with

$$\begin{aligned}\Phi(x_1,s) &= s'\left(s^2 + a + \frac{1}{x_1^4}\right) - s\left(2ss' - 4\frac{x_2}{x_1^5}\right)\\ &= s'\left(a - s^2 + \frac{1}{x_1^4}\right) - 4\frac{s^2}{x_1^4}.\end{aligned}$$

Using

$$\begin{aligned}s' &= -\frac{d}{dt}\left(\frac{x_2}{x_1}\right) = -\frac{\dot{x}_2x_1 - x_2\dot{x}_1}{x_1^2}\\ &= \frac{x_2^2 - x_1\left(-bx_1 + \frac{1}{x_1^3}\right)}{x_1^2} = s^2 + b - \frac{1}{x_1^4},\end{aligned}$$

we obtain that

$$\Phi(x_1,s) = -(s^2+b)(s^2-a) - \frac{1}{x_1^4}\left((s^2+a) + (s^2-b) + \frac{1}{x_1^4}\right).$$

At the initial point ($p = 0$ or $x_1 = 1$), the function $p \mapsto s^2 - b + \frac{1}{x_1^4}$ vanishes if $b = 1$ and is negative if $b > 1$. As it is strictly increasing,

$$\frac{d}{dp}\left(s^2 - b + \frac{1}{x_1^4}\right) = 2ss' - 4\frac{x_2}{x_1^5} = 2s\left(s' + \frac{2}{x_1^4}\right) = 2s\left(s^2 + b + \frac{1}{x_1^4}\right) > 0,$$

there exists a unique parameter value $p_{\rm qu}$ where $s^2 - b + \frac{1}{x_1^4} = 0$ and this expression is positive for $p > p_{\rm qu}$. For $b > 1$ we have that $p_{\rm qu} = \frac{\pi}{4\sqrt{b}}$. For, it follows from

$$x_2^2 + bx_1^2 + \frac{1}{x_1^2} = b + 1 \qquad \text{and} \qquad x_2^2 - bx_1^2 + \frac{1}{x_1^2} = 0$$

that $x_1^2 = \frac{b+1}{2b} = z_{\rm qu}$ and thus, substituting into the formula for the solution, we obtain that

$$\frac{b+1}{2b} = \cos^2(\sqrt{b}p) + \frac{1}{b}\sin^2(\sqrt{b}p) = 1 - \frac{b-1}{b}\sin^2(\sqrt{b}p)$$

which gives us that

$$\sin^2(\sqrt{b}p) = \frac{1}{2} \qquad \Leftrightarrow \qquad \sqrt{b}p = \frac{\pi}{4}.$$

Hence, for $p \geq p_{\rm qu}$, we have that $s^2 - b + \frac{1}{x_1^4} \geq 0$ and thus

$$\Phi(x_1, s) \leq -(s^2 + b)(s^2 - a) - \frac{s^2 + a}{x_1^4} < 0.$$

Thus the function $\theta(p) = \tau_{X:(\zeta,\xi)\to(\zeta,-\xi)}$ is strictly decreasing on this interval. This proves the Lemma. □

The statement of the theorem then is an immediate consequence of the bound $\frac{\pi}{4\sqrt{b}} < p_{\text{alg}} \leq \tilde{p}$ where p_{alg} is the parameter value corresponding to the algebraic mean of the values $\zeta_{a,-}^2$ and $\zeta_{a,+}^2$. This holds since for the corresponding z-values we have that (c.f., 4.37)

$$z_{\text{qu}} = \frac{v}{2b} > \frac{v}{2(a+b)} = \frac{v}{2w} = z_{\text{alg}} > \frac{2}{v} = \breve{z} > \tilde{z}.$$

The parameters obey the reverse inequality relations and thus this completes the proof of the Theorem. □

The fact that the time difference $\Delta T = T_2 - T_1$ decreases monotonically does not guarantee that an intersection of the graphs of T_1 and T_2 exists. Indeed, the limit of ΔT as $\gamma \to \infty$ can be positive.

Theorem 6. *It holds that* $\lim_{\gamma\to\infty} \Delta T(\gamma) = \lim_{p\to p_{\text{fin}}} \Delta\widehat{T}(p) = \varpi(a, b)$ *where*

$$\varpi(a, b) = \frac{1}{\sqrt{b}} \sin^{-1}\left(\sqrt{\frac{b(1-\zeta_{a,-}^2)}{b-1}}\right) + \frac{1}{\sqrt{a}} \ln\left(\sqrt{(a+1)\zeta_{a,-}^2}\right) \tag{4.51}$$

and, as before, $\zeta_{a,-}^2 = \frac{v}{2w}\left\{1 - \sqrt{1 - \frac{4w}{v^2}}\right\}$.

Proof. We consider all expressions (such as ζ, κ, γ or s^2) as functions of p, but we do not write the argument in order to simplify the notation. Recall that

$$\Delta\widehat{T}(p) = p + \tau_X(p) - \widehat{\sigma}_X(p) + \tau_{\text{fin}}(p) - \widehat{\sigma}_Y(p).$$

In the limit $p \to p_{\text{fin}}$ we have that $s^2 \to a$ and $\zeta^2 \to \zeta_{a,-}^2$. Hence it follows from equation (4.20) that

$$\lim_{p\to p_{\text{fin}}} p = p_{\text{fin}} = \frac{1}{\sqrt{b}} \sin^{-1}\left(\sqrt{\frac{b(1-\zeta_{a,-}^2)}{b-1}}\right).$$

Along the intermediate X-arc of the YXY-trajectory we have that $\frac{1}{\kappa^2} = \zeta^2(s^2 - a) \to 0$ and thus both κ and $\gamma > \kappa$ diverge to ∞ as $p \to p_{\text{fin}}$. The points (κ, μ) and $(\gamma, 0)$ lie on the final Y-segment and thus it also holds that $(s^2 + b)\kappa^2 + \frac{1}{\kappa^2} = b\gamma^2 + \frac{1}{\gamma^2}$. Taking the limit $p \to p_{\text{fin}}$ it therefore follows that

$$\lim_{p\to p_{\text{fin}}} \left\{\frac{\kappa^2}{\gamma^2}\right\} = \frac{b}{a+b} \tag{4.52}$$

and thus

$$\lim_{p\to p_{\text{fin}}} \left\{\frac{b\gamma^4}{b\gamma^4 - 1} \cdot \frac{\gamma^2 - \kappa^2}{\gamma^2}\right\} = \frac{a}{a+b} = \lim_{p\to p_{\text{fin}}} \left\{\frac{(a\gamma^2 + 1)}{a+b} \cdot \frac{b(\gamma^2 - 1)}{b\gamma^4 - 1}\right\}.$$

Hence the times τ_{fin} and $\widehat{\sigma}_Y$ along the final Y-segments for the YXY- and the XY-trajectories converge to the same limit

$$\frac{1}{\sqrt{b}}\sin^{-1}\left(\sqrt{\frac{a}{a+b}}\right)$$

and these terms cancel.

Each of the terms $\tau_X(p)$ and $\widehat{\sigma}_X(p)$ diverges to ∞ as $p \to p_{\rm fin}$, but their difference has a finite limit. Since

$$\sinh^{-1}(z) = \ln\left(z + \sqrt{1+z^2}\right) = \ln\left(z\left[1+\sqrt{1+\frac{1}{z^2}}\right]\right), \tag{4.53}$$

the difference $\tau_X(p) - \widehat{\sigma}_X(p)$ is given by

$$\begin{aligned}
&\frac{1}{\sqrt{a}}\left(\sinh^{-1}\left(\sqrt{a}\zeta\kappa\right) - \sinh^{-1}\left(\sqrt{\frac{a(\gamma^2-1)}{a+b}\cdot\frac{b\gamma^2-1}{(a+1)\gamma^2}}\right)\right)\\
&\quad= \frac{1}{\sqrt{a}}\ln\left(\frac{\sqrt{a}\zeta}{\sqrt{\frac{a}{(a+b)(a+1)}\frac{(\gamma^2-1)(b\gamma^2-1)}{\gamma^4}}}\times\frac{\kappa}{\gamma}\times\right.\\
&\qquad\qquad\left.\times\frac{\left[1+\sqrt{1+\frac{1}{a\zeta^2\kappa^2}}\right]}{\left[1+\sqrt{1+\frac{(a+b)(a+1)}{a}\frac{\gamma^2}{(\gamma^2-1)(b\gamma^2-1)}}\right]}\right)\\
&\quad\to \frac{1}{\sqrt{a}}\ln\left(\sqrt{(a+b)(a+1)}\zeta_{a,-}\sqrt{\frac{b}{a+b}}\frac{1}{\sqrt{b}}\right) = \frac{1}{\sqrt{a}}\ln\left(\sqrt{a+1}\zeta_{a,-}\right)
\end{aligned}$$

as $p \to p_{\rm fin}$. Here we used that $\zeta \to \zeta_{a,-}$, $\zeta\kappa = \frac{1}{\sqrt{s^2-a}} \to \infty$, and $\frac{\kappa}{\gamma} \to \sqrt{\frac{b}{a+b}}$. This proves the result. □

Corollary 3. *If $\varpi(a,b) \geq 0$, then for all $\gamma > 1$ the time along the XY-trajectory that steers $(1,0)$ into $(\gamma,0)$ is faster than the time along the YXY-extremal. If $\varpi(a,b) < 0$, then there exists a unique value $\widehat{\gamma}_1$, the* first cut-locus*, where the total time along the XY- and YXY-trajectory that steer $(1,0)$ into $(\widehat{\gamma}_1,0)$ are equal. For $\gamma < \widehat{\gamma}_1$ the XY-trajectory is faster while the YXY-trajectory is faster for $\gamma > \widehat{\gamma}_1$.*

Proof. Since the time difference ΔT is strictly monotonically decreasing for $\gamma \geq \bar{\gamma}$, if $\varpi(a,b) \geq 0$, then it follows that $\Delta T(\gamma) > 0$ for all $\gamma \geq \bar{\gamma}$. On the other hand, since ΔT is positive for γ near 1, if $\varpi(a,b) < 0$ then ΔT has a unique zero. □

4.8 Cut-loci between Y-loops with n Turns

We assume that $\varpi(a,b) < 0$ so that a first cut-locus characterized by the parameter $\widehat{p}_1$ and terminal value $\widehat{\gamma}_1$ exists. The total time along the parameterized $YX\cdots XY$-extremal for parameter p with n loops is given by (c.f., Proposition 6)

$$\widehat{T}_{2n}(p) = p + \frac{n}{\sqrt{a}}\sinh^{-1}\left(\sqrt{\frac{a}{s^2-a}}\right) + \frac{n-1}{\sqrt{b}}\left(\pi - \sin^{-1}\left(\sqrt{\frac{b}{s^2+b}}\right)\right)$$

$$+\frac{1}{\sqrt{b}}\sin^{-1}\left(\sqrt{\frac{b\gamma_n^2(\gamma_n^2-\kappa_n^2)}{b\gamma_n^4-1}}\right)$$

where (κ_n, μ_n) denotes the last XY-junction, $s = \frac{\mu_n}{\kappa_n}$ is the slope of the line on which the XY-junctions lie and γ_n is the terminal value, all functions of the parameter p. We recall that $\gamma_n = \gamma_n(p)$ is the solution of the equation (4.42) which satisfies $\gamma_n > \kappa_n$ and that it can be expressed in the form

$$\gamma_n^2(p) = \frac{1}{2b}\left((s^2-a)\zeta_n^2 + (s^2+b)\kappa_n^2 + \sqrt{\left[(s^2-a)\zeta_n^2 + (s^2+b)\kappa_n^2\right]^2 - 4b}\right). \tag{4.54}$$

Like for YXY-extremal trajectories, there is a 1:1 correspondence between the parameters $p \geq \tilde{p}_{2n}$ and terminal values $\gamma \geq \tilde{\gamma}_{2n}$: including the index n in the notation we write

$$\gamma_n : [\tilde{p}_{2n}, p_{\text{fin}}) \to [\tilde{\gamma}_{2n}, \infty), \qquad p \mapsto \gamma_n(p),$$

respectively,

$$\pi_n : [\tilde{\gamma}_{2n}, \infty) \to [\tilde{p}_{2n}, p_{\text{fin}}), \qquad \gamma \mapsto \pi_n(\gamma),$$

for the corresponding inverse mapping: given $\gamma \geq \tilde{\gamma}_{2n}$, $\pi_n(\gamma)$ is the parameter p such that $\gamma_n(p) = \gamma$.

We again define the time-difference ΔT_n as a function of the terminal condition γ. While the parameterized times $\widehat{T}_{2n}(p)$ matter for analyzing the time-difference $\Delta T_n(\gamma)$, the parameters in the functions $\widehat{T}_{2n}$ and $\widehat{T}_{2(n-1)}$ are different when a fixed terminal condition γ is considered. We write $T_{2n} = \widehat{T}_{2n} \circ \pi_n$ for the total time along the Y-loop with n turns and terminal condition $(\gamma, 0)$,

$$T_{2n} : [\tilde{\gamma}_{2n}, \infty) \to [0, \infty), \qquad \gamma \mapsto T_{2n}(\gamma) = \widehat{T}_{2n}(\pi_n(\gamma)),$$

and we define the time-differences as

$$\Delta T_n = T_{2n} - T_{2(n-1)} = \widehat{T}_{2n} \circ \pi_n - \widehat{T}_{2(n-1)} \circ \pi_{n-1}. \tag{4.55}$$

The graphs of the times T_{2i}, $i = 1, 2, 3$, in Figure 4.5 exhibit the following qualitative features:

- For small terminal values γ we have that $T_{2n}(\gamma) > T_{2(n-1)}(\gamma)$, i.e., Y-loops with additional switchings are slower.
- The time difference $\Delta T_n = T_{2n} - T_{2(n-1)}$ is monotonically decreasing.
- For large terminal values γ we have that $T_{2n}(\gamma) < T_{2(n-1)}(\gamma)$, i.e., Y-loops with additional switchings are faster.

These are indeed generally valid statements and we outline how some of these results are proven starting with the asymptotic behavior at infinity.

Proposition 7. *It holds that* $\lim_{\gamma\to\infty} \Delta T_n(\gamma) = \widehat{\varpi}(a, b)$ *where*

$$\widehat{\varpi}(a,b) = \frac{1}{\sqrt{a}}\ln\left(2\sqrt{\frac{a}{a+b}}\right) + \frac{1}{\sqrt{b}}\left(\pi - \sin^{-1}\left(\sqrt{\frac{b}{a+b}}\right)\right). \tag{4.56}$$

This limit is independent of n, the number of loops.

Proof. We break up the time difference into the sum of the differences along the initial segment, the intermediate X- and Y-arcs, and the final segment in the form $\Delta T_n = \Delta_{\text{in}} + \Delta_X + \Delta_Y + \Delta_{\text{fin}}$ where

$$\Delta_{\text{in}}(\gamma) = \pi_n(\gamma) - \pi_{n-1}(\gamma), \tag{4.57}$$

$$\Delta_X(\gamma) = \frac{n}{\sqrt{a}} \sinh^{-1}\left(\sqrt{\frac{a}{s_n^2 - a}}\right) - \frac{n-1}{\sqrt{a}} \sinh^{-1}\left(\sqrt{\frac{a}{s_{n-1}^2 - a}}\right), \tag{4.58}$$

$$\begin{aligned}\Delta_Y(\gamma) = {} & \frac{n-1}{\sqrt{b}}\left(\pi - \sin^{-1}\left(\sqrt{\frac{b}{s_n^2 + b}}\right)\right) \\ & - \frac{n-2}{\sqrt{b}}\left(\pi - \sin^{-1}\left(\sqrt{\frac{b}{s_{n-1}^2 + b}}\right)\right),\end{aligned} \tag{4.59}$$

and

$$\begin{aligned}\Delta_{\text{fin}}(\gamma) = {} & \frac{1}{\sqrt{b}} \sin^{-1}\left(\sqrt{\frac{b\gamma^2(\gamma^2 - \kappa_n^2)}{b\gamma^4 - 1}}\right) \\ & - \frac{1}{\sqrt{b}} \sin^{-1}\left(\sqrt{\frac{b\gamma^2(\gamma^2 - \kappa_{n-1}^2)}{b\gamma^4 - 1}}\right).\end{aligned} \tag{4.60}$$

In these expressions, given γ, $\pi_i(\gamma)$, $i = n-1$ or $i = n$, denotes the parameter p_i such that $\gamma_i(p) = \gamma$, κ_i is the x_1-coordinate of the last XY-junction and $s_i = s \circ \pi_i$ is the slope of the line on which the XY-junctions lie. All these expressions are functions of the terminal condition γ. We emphasize that the indices $n-1$ and n do not denote consecutive points on a specific trajectory, but rather the last switching points on different Y-loops corresponding to $n-1$, respectively n turns in the overall trajectory.

The times along the vector field Y, i.e., Δ_{in}, Δ_Y and Δ_{fin}, have well-defined limits as $\gamma \to \infty$ that are easily computed: As $\gamma \to \infty$, also $\kappa_i \to \infty$ and it follows from equation (4.46) that $\lim_{\gamma\to\infty} s_i(\gamma) = \sqrt{a}$. The parameters therefore converge to either p_{in} or p_{fin}. Since the parameters satisfy $p > \widehat{p}_1$, the first cut-locus, it follows that $\pi_i(\gamma) \to p_{\text{fin}}$. Hence we have that $\lim_{\gamma\to\infty} \pi_i(\gamma) = p_{\text{fin}}$ and thus $\lim_{\gamma\to\infty} \Delta_{\text{in}}(\gamma) = 0$. Furthermore, $\lim_{\gamma\to\infty} s_i(\gamma) = \sqrt{a}$ implies that

$$\lim_{\gamma\to\infty} \Delta_Y(\gamma) = \frac{1}{\sqrt{b}}\left(\pi - \sin^{-1}\left(\sqrt{\frac{b}{a+b}}\right)\right). \tag{4.61}$$

Since the point (κ_i, μ_i) lies on the final Y-segment, it holds that $(s_i^2 + b)\kappa_i^2 + \frac{1}{\kappa_i^2} = b\gamma^2 + \frac{1}{\gamma^2}$ which, as in the case of YXY-extremals, gives us that

$$\lim_{\gamma\to\infty}\left\{\frac{\kappa_i^2}{\gamma^2}\right\} = \frac{b}{a+b}. \tag{4.62}$$

Hence

$$\lim_{\gamma\to\infty}\left\{\frac{b\gamma^2(\gamma^2 - \kappa_i^2)}{b\gamma^4 - 1}\right\} = \lim_{\gamma\to\infty}\left\{\frac{b\gamma^4}{b\gamma^4 - 1} \cdot \frac{\gamma^2 - \kappa_i^2}{\gamma^2}\right\} = \frac{a}{a+b}$$

and thus, like in the case of the first cut-locus, the final terms cancel each other: $\lim_{\gamma\to\infty} \Delta_{\text{fin}}(\gamma) = 0$.

It is less trivial to analyze the difference in the times along the X-segments: each of these terms separately diverges to infinity and the differences between the indices $n-1$ and n are subtle. While the trajectory with n loops makes an additional turn, these turns lie closer to the origin and therefore the speed of the trajectories is faster. Overall, the time along X-trajectories is indeed smaller if more loops are made. Again using (4.53), it follows that

$$\begin{aligned}\Delta_X(\gamma) &= \frac{n}{\sqrt{a}}\sinh^{-1}\left(\sqrt{\frac{a}{s_n^2-a}}\right) - \frac{n-1}{\sqrt{a}}\sinh^{-1}\left(\sqrt{\frac{a}{s_{n-1}^2-a}}\right)\\ &= \frac{1}{\sqrt{a}}\left\{\ln\left(\sqrt{\frac{a}{s_n^2-a}}\left[1+\sqrt{1+\frac{s_n^2-a}{a}}\right]\right)^n\right.\\ &\qquad\left.-\ln\left(\sqrt{\frac{a}{s_{n-1}^2-a}}\left[1+\sqrt{1+\frac{s_{n-1}^2-a}{a}}\right]\right)^{n-1}\right\}\\ &= \frac{1}{\sqrt{a}}\ln\left(\sqrt{a}\frac{\left(s_{n-1}^2-a\right)^{\frac{n-1}{2}}}{\left(s_n^2-a\right)^{\frac{n}{2}}}\cdot\frac{\left[1+\sqrt{1+\frac{s_n^2-a}{a}}\right]^n}{\left[1+\sqrt{1+\frac{s_{n-1}^2-a}{a}}\right]^{n-1}}\right).\end{aligned}$$

Since $s_i^2 \to a$ as $\gamma\to\infty$ we have that

$$\lim_{\gamma\to\infty}\frac{\left(1+\sqrt{1+\frac{s_n^2-a}{a}}\right)^n}{\left(1+\sqrt{1+\frac{s_{n-1}^2-a}{a}}\right)^{n-1}} = 2$$

and it remains to compute the limit

$$\lim_{\gamma\to\infty}\frac{\left(s_{n-1}^2-a\right)^{\frac{n-1}{2}}}{\left(s_n^2-a\right)^{\frac{n}{2}}}.$$

This will be accomplished by means of the following fundamental relation for the slope s_n:

Lemma 4. [19] *Let $\psi_\gamma = b\gamma^2+\frac{1}{\gamma^2}$ for $\gamma\geq 1$. For $\gamma\geq\breve{\gamma}=\gamma_n(\breve{p})$, the slope $s=s_n(\gamma)$ (of the line on which the XY-junctions lie for an extremal controlled Y-loop with n turns) satisfies the following equation:*

$$\frac{\psi_1+\sqrt{\psi_1^2-4(s^2+b)}}{\psi_\gamma+\sqrt{\psi_\gamma^2-4(s^2+b)}} = \left(\frac{s^2-a}{s^2+b}\right)^n. \tag{4.63}$$

Proof of the Lemma. Given $\gamma\geq\tilde{\gamma}_{2n}=\gamma_n(\tilde{p}_{2n})$, the parameterised family $\mathscr{E}_{2n}$ contains a unique extremal Y-loop that makes n turns and ends at γ. The parameter $p=\pi_n(\gamma)\geq\breve{p}$ is given by (c.f., equation (4.20))

$$p=\pi_n(\gamma)=\frac{1}{\sqrt{b}}\sin^{-1}\left(\frac{b(1-\zeta_1^2)}{b-1}\right)$$

where ζ_1^2, κ_n^2 and $s^2=s_n^2$ are the solutions to the following three equations:

$$(s^2+b)\zeta_1^2 + \frac{1}{\zeta_1^2} = b+1, \tag{4.64}$$

$$(s^2+b)\kappa_n^2 + \frac{1}{\kappa_n^2} = b\gamma^2 + \frac{1}{\gamma^2}, \tag{4.65}$$

$$(s^2-a)\kappa_n^2 = \frac{1}{\zeta_n^2} = \left(\frac{s^2+b}{s^2-a}\right)^{n-1} \frac{1}{\zeta_1^2}. \tag{4.66}$$

Equation (4.64) states that ζ_1 lies on the initial Y-trajectory through $(1,0)$, equation (4.65) states that κ_n lies on the last Y-arc through $(\gamma,0)$ and equation (4.66) relates the first and n-th switching points. As $p \geq \breve{p}$, we thus have that (writing, for the moment, κ for the x_1-coordinate of the *first* XY-junction on the Y-loop with n turns)

$$\begin{aligned}
\zeta_1^2 &= \frac{v}{2(s^2+b)}\left\{1 - \sqrt{1 - \frac{4(s^2+b)}{v^2}}\right\}, \\
\kappa_n^2 &= \kappa^2\left(\frac{s^2+b}{s^2-a}\right)^{n-1} = \frac{1}{\zeta_1^2(s^2+b)}\left(\frac{s^2+b}{s^2-a}\right)^n \\
&= \left(\frac{s^2+b}{s^2-a}\right)^n \cdot \frac{2}{v} \cdot \frac{1+\sqrt{1-\frac{4(s^2+b)}{v^2}}}{\frac{4(s^2+b)}{v^2}} \\
&= \left(\frac{s^2+b}{s^2-a}\right)^n \cdot \frac{1}{2(s^2+b)}\left\{v + \sqrt{v^2 - 4(s^2+b)}\right\}
\end{aligned}$$

and also

$$\kappa_n^2 = \frac{1}{2(s^2+b)}\left\{b\gamma^2 + \frac{1}{\gamma^2} + \sqrt{\left(b\gamma^2 + \frac{1}{\gamma^2}\right)^2 - 4(s^2+b)}\right\}.$$

Equating the expressions for κ_n^2 and noting that $v = \psi_1$ equation (4.63) follows. This proves the lemma. □

Equation (4.63) allows us to determine the asymptotic behavior of $s_i^2 - a$ in the limit $\gamma \to \infty$. We have that

$$\begin{aligned}
\lim_{\gamma\to\infty} \gamma^2\left(s_i^2 - a\right)^i &= \lim_{\gamma\to\infty} \frac{\psi_1 + \sqrt{\psi_1^2 - 4(s_i^2+b)}}{\frac{1}{\gamma^2}\left(\psi_\gamma + \sqrt{\psi_\gamma^2 - 4(s_i^2+b)}\right)}\left(s_i^2+b\right)^i \\
&= \frac{\left[v + \sqrt{v^2-4w}\right]w^i}{\lim_{\gamma\to\infty}\left\{b + \frac{1}{\gamma^4} + \sqrt{\left(b+\frac{1}{\gamma^4}\right)^2 - \frac{4(s_i^2+b)}{\gamma^4}}\right\}} \\
&= \frac{v+\sqrt{v^2-4w}}{2b} w^i.
\end{aligned}$$

Hence

$$\lim_{\gamma\to\infty} \frac{\left(s_{n-1}^2 - a\right)^{\frac{n-1}{2}}}{\left(s_n^2-a\right)^{\frac{n}{2}}} = \lim_{\gamma\to\infty} \sqrt{\frac{\gamma^2\left(s_{n-1}^2-a\right)^{n-1}}{\gamma^2\left(s_n^2-a\right)^n}} = \frac{1}{\sqrt{w}} = z_{\text{geo}}$$

and thus

$$\lim_{\gamma\to\infty} \Delta_X(\gamma) = \frac{1}{\sqrt{a}} \ln\left(2\sqrt{\frac{a}{a+b}}\right).$$

This verifies equation (4.56) and proves the result. □

It remains to show the monotonicity of the functions ΔT_n. This is quite a lengthy calculation which leads to a surprisingly simple final result. Essentially, we need to compute the derivative of the function $T_{2n} = \widehat{T}_{2n} \circ \pi_n$, i.e., the time along Y-loops with n turns, as a function of the terminal condition γ. We recall that

$$T_{2n}(\gamma) = \pi_n(\gamma) + \frac{n}{\sqrt{a}} \sinh^{-1}\left(\sqrt{\frac{a}{s_n^2 - a}}\right) + \frac{n-1}{\sqrt{b}}\left(\pi - \sin^{-1}\left(\sqrt{\frac{b}{s_n^2+b}}\right)\right) + \frac{1}{\sqrt{b}} \sin^{-1}\left(\sqrt{\frac{b\gamma^2(\gamma^2-\kappa_n^2)}{b\gamma^4 - 1}}\right)$$

where κ_n denotes the x_1-coordinate of the last XY-junction and $s_n = s \circ \pi_n$ is the slope of the line on which the XY-junctions lie. Here, however, κ_n and s_n are functions of γ. We also write ζ_n for the x_1-coordinate of the *first* YX-junction along this Y-loop with n turns and use a prime to denote the derivative with respect to γ. The following two formulas state (without proof) some intermediate steps of the computations which then lead to the simple final form in Theorem 7.

$$T'_{2n}(\gamma) = s'_n \left\{ \frac{1}{s_n^2 + b - \frac{1}{\zeta_n^4}} - \frac{n}{s_n^2 - a} + \frac{n-1}{s_n^2+b} + \frac{1}{s_n^2 + b - \frac{1}{\kappa_n^4}} \right\} - \frac{4\gamma}{b\gamma^4 - 1} \cdot \frac{s_n}{(s_n^2+b)\kappa_n^2 - \frac{1}{\kappa_n^2}}$$

and

$$2s_n s'_n \left\{ \frac{n}{s_n^2+b} - \frac{n}{s_n^2 - a} + \frac{2}{\left(\psi_\gamma + \sqrt{\psi_\gamma^2 - 4(s_n^2+b)}\right)\sqrt{\psi_\gamma^2 - 4(s_n^2+b)}} - \frac{2}{\left(\psi_1 + \sqrt{\psi_1^2 - 4(s_n^2+b)}\right)\sqrt{\psi_1^2 - 4(s_n^2+b)}} \right\} = \frac{\psi'_\gamma}{\sqrt{\psi_\gamma^2 - 4(s_n^2+b)}}.$$

Define the function

$$\theta : (1,\infty) \to \left(\frac{1}{2}(b-1), \infty\right) \qquad \gamma \mapsto \theta(\gamma) = \frac{1}{2}\left(b\gamma^2 - \frac{1}{\gamma^2}\right) = \frac{1}{4}\gamma\psi'_\gamma.$$

For $\gamma > 1$, the simple identity $b\gamma^4 - (b-1)\gamma^2 - 1 = (\gamma^2 - 1)(b\gamma^2 + 1) > 0$ gives us the lower bound $\theta(\gamma) > \frac{1}{2}(b-1)$ and, since the slope s_n is bounded above by $\frac{1}{2}(b-1)$, we therefore have that $s_n(\gamma) < \theta(\gamma)$ for all $\gamma > 1$. Using this notation, we have the following result:

Theorem 7. *For $\gamma \geq \breve{\gamma}$ (i.e., when taking the negative sign in the square-root when solving for ζ) it holds that*

$$T'_{2n}(\gamma) = \frac{1}{\gamma}\frac{\sqrt{\theta^2(\gamma) - s_n^2(\gamma)}}{\theta(\gamma)s_n(\gamma)} = \frac{1}{\gamma}\sqrt{\frac{1}{s_n^2(\gamma)} - \frac{1}{\theta_n^2(\gamma)}}. \tag{4.67}$$

In particular, this derivative only depends on the terminal value γ and the slope s_n of the line on which the XY-junctions lie.

The following lemma (which is quite intuitive geometrically) then provides the last piece in the proof.

Lemma 5. *Given γ, suppose there exist extremal trajectories which make $n-1$, respectively n turns and end at $(\gamma, 0)$. Let $s_{n-1} = s_{n-1}(\gamma)$ and $s_n = s_n(\gamma)$ denote the respective (positive) slopes of the associated lines on which the XY-junctions lie. Then we have that $s_n > s_{n-1}$.* □

Corollary 4. *For every index $n \geq 1$, the function $\Delta T_n = T_{2n} - T_{2(n-1)}$ is strictly decreasing for $\gamma \geq \breve{\gamma}$.*

Proof. By Theorem 7 the time derivative of T_{2n} can be expressed in the form $T'_{2n}(\gamma) = \Upsilon(\gamma, s_n(\gamma))$ with $\Upsilon = \Upsilon(\gamma, s)$ given by

$$\Upsilon(\gamma, s) = \frac{\sqrt{\theta^2(\gamma) - s^2}}{\gamma\theta(\gamma)s} = \frac{1}{\gamma}\sqrt{\frac{1}{s^2} - \frac{1}{\theta^2(\gamma)}}. \tag{4.68}$$

Hence

$$\Delta T'_n(\gamma) = \Upsilon(\gamma, s_n(\gamma)) - \Upsilon(\gamma, s_{n-1}(\gamma))$$

and as the sequence $\{s_n\}_{n\in\mathbb{N}}$ is monotonically increasing with values in the interval $\big(\sqrt{a}, \frac{1}{2}(b-1)\big)$, it merely remains to show that the function Υ is monotonically decreasing in s for $\gamma > 1$. This is elementary:

$$\frac{\partial\Upsilon}{\partial s}(\gamma, s) = -\frac{1}{\gamma}\frac{\frac{1}{s^3}}{\sqrt{\frac{1}{s^2} - \frac{1}{\theta^2(\gamma)}}} = -\frac{1}{\gamma s^2}\frac{\theta(\gamma)}{\sqrt{\theta^2(\gamma) - s^2}} < 0.$$

This verifies the corollary. □

Taken all together, Theorem 1 follows.

References

1. D. Aharonov, W. van Dam, J. Kempe, Z. Landau, S. Lloyd, and O. Regev, Adiabatic quantum computation is equivalent to standard quantum computation, *SIAM J. Computation* **37**, (2007), pp. 166–194.
2. B. Bonnard and M. Chyba, *Singular Trajectories and Their Role in Control Theory*, Mathématiques & Applications, Vol. 40, Springer-Verlag, Paris, 2003.
3. A. Bressan and B. Piccoli, *Introduction to the Mathematical Theory of Control*, American Institute of Mathematical Sciences (AIMS), 2007.
4. S. Bize, et al., Cold atom clocks and applications, *J. Physics B: Atomic, Molecular and Optical Physics*, **38**, (2005), pp. S449–S468.
5. A. Bohm, *Quantum Mechanics: Foundations and Applications*, Springer Verlag, New York, 1979.
6. X. Chen, A. Ruschhaupt, S. Schmidt, A. del Campo, D. Guéry-Odelin, and J.G. Muga, Fast optimal frictionless atom cooling in harmonic traps: Shortcut to adiabaticity *Physical Review Letters*, **104**, (2010) 063002.

7. J.I. Cirac and P. Zoller, New frontiers in quantum information with atoms and ions, *Physics Today*, **57**, (2004), pp. 38-–44.
8. V.P. Ermakov, Second-order differential equations. Integrability conditions in closed form [in Russian], *Universitetskie Izvestiya, Kiev*, **9**, (1880), pp. 1—25.
9. W.H. Fleming and R.W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, 1975.
10. Y. Kagan, E.L. Surkov, and G.V. Shlyapnikov, Evolution of a Bose-condensed gas under variations of the confining potential, *Physical Reviews A*, **54**, (1996), pp. R1753–1756.
11. R. Kosloff and Y. Rezek, The quantum harmonic Otto cycle, *Entropy*, **19**, (2017), 136, https://doi.org/10.3390/e19040136
12. A.E. Leanhardt, T.A. Pasquini, M. Saba, A. Schirotzek, Y. Shin, D. Kielpinski, D.E. Pritchard and W. Ketterle, Adiabatic and evaporative cooling of Bose-Einstein condensates below 500 picokelvin, *Science*, **301**, (2003), pp. 1513-–1515.
13. H.R. Lewis and W.B. Riesenfeld, An exact quantum theory of the time-dependent harmonic oscillator and of a charged particle in a time-dependent electromagnetic field, *J. of Mathematical Physics*, **10**, (1969), pp. 1458–1473.
14. E. Pinney, The nonlinear differential equation $y''+p(x)y+c\frac{1}{y^3}=0$, *Proc. of the American Mathematical Society*, **1** (1950), pg. 681.
15. P. Salamon, K.H. Hoffmann, Y. Rezek, and R. Kosloff, Maximum work in minimum time from a conservative quantum system, *Physical Chemistry Chemical Physics*, **11**, (2009), pp. 1027-–1032.
16. H. Schättler and U. Ledzewicz, *Geometric Optimal Control*, Springer, New York, 2012.
17. D. Stefanatos, Minimum-time transitions between thermal equilibrium states of the quantum parametric oscillator, *IEEE Transactions on Automatic Control*, **62**, (2017), pp. 4290–4297.
18. D. Stefanatos, J. Ruths, and J.S. Li, Frictionless atom cooling in harmonic traps: a timeoptimal approach, *Physical Reviews A*, **82**, (2010), 063422.
19. D. Stefanatos, H. Schättler and J.S. Li, Minimum-time frictionless atom cooling in harmonic traps, *SIAM J. on Control and Optimization - SICON*, **49**(6), 2011, pp. 2440—2462
20. H.J. Sussmann, Time-optimal control in the plane, in: *Feedback Control of Linear and Nonlinear Systems*, Lecture Notes in Control and Information Sciences, Vol. 39, Springer-Verlag, Berlin, (1982), pp. 244–260.
21. H.J. Sussmann, The structure of time-optimal trajectories for single-input systems in the plane: the C^∞ nonsingular case, *SIAM J. Control Optim.*, **25** (1987), pp. 433–465.
22. C.E. Wieman, D.E. Pritchard and D.J. Wineland, Atom cooling, trapping, and quantum manipulation, Reviews of Modern Physics, **71**, (1999), pp. S253–262.

5

Optimal Geometric Control of a Quadcopter

Monique Chyba[1] and Christopher Gray[1]

Department of Mathematics, University of Hawaii at Manoa 2565 McCarthy Mall Honolulu, Hawaii 96822
chyba@hawaii.edu,ch.b.gray@gmail.com

This paper highlights the peculiar behavior that is the Fuller phenomenon for the motion planning of quadcopters. The subject of chattering arcs on optimal control was introduced to me by Ivan while I was a post-doc in Paris at the end of the 90's. Still today I marvel at his love for mathematics and relentless dedication to it. (Monique Chyba)

Summary. In recent years, with the rise of affordable commercial grade quadcopters, there has been a lot of research done on modeling the motion of quadcopters, however much of it has been focused on creating robust control schemes for quadcopters. This In this paper we first introduce the equations of motion of a quadcopter. The goal of the paper is, using an affine control simplified model version of the equations of motion, to study the time minimization problem with an emphasis on singular extremals. We observe numerically the possibility of Fuller phenomenon.

5.1 Introduction

Unmanned Aerial Vehicles (UAVs) trace their history back to World War I where the US Navy and Army experimented with aerial torpedoes and flying bombs, but it wasn't until 1938 when what we would consider a modern drone would be deployed by the US Navy [9]. In modern times UAV's play a vital role in tracking and detection of opposing military groups and remote monitoring of strategic locations. The focus of UAVs has also expanded to include many commercial, emergency response and environmental uses including but not limited to agriculture, monitoring lava flows during the 2018 Kīlauea volcano eruption and even surveying of coral reefs. Recently there is also a strong interest in increasing the usage of quadcopters for agriculture. When multi-spectrometers or thermal cameras are attached they can be used to monitor health of the crops or find and track lost livestock. For instance, starting in 2002, commercial quadcopters equipped with different sensors were used to study the dynamics, including erosion, of the coastlines of various Hawai'i Beaches. In the last decade there was even a push to use quadcopters to survey infested trees in Hawai'i to stop the spread of a fungus called Rapid 'Ōhi'a Death. While UAVs can be quite sophisticated with complex payload instrumentation, in this paper we will focus on quadcopters which use only a few sensors and engines, with most just having a position tracker, an accelerometer and a way to control the angular velocity of the rotors.

The contribution of this paper is twofold. First, we develop a comprehensive set of equations of motion for quadcopters. Those can then be used to control the motion of these UAVs. The equations should encompass enough of the physical factors to obtain an accurate and realistic model that can be used to calculate the controls necessary to move within an acceptable margin of error over a given time, while also being simple enough that new calculations to compensate for new real world data can be done quickly. Second, we focus on optimal control. Optimization is important for autonomous vehicles since it allows the vehicle to complete a mission in minimum time or by using minimum energy. While the optimal controls depend on the cost that is being minimized, singular trajectories are intrinsic to the system. It is well-known that they can play an important role in the construction of optimal paths. In this paper we characterize the singular trajectories for quadcopters by studying the Lie Algebra associated to the drift and control vector fields of the equations of motion. We also simulate regular extremals for the time minimization problem with bounded controls. It is well known that in situations like this the optimal control is a concatenation of singular arcs and bang arcs. With our simulations, we observe numerically the Fuller phenomenon for our system.

The outline of the paper is as follows. In section 5.2 we discuss and set up the equations of motion of a quadcopter that we will be using. Section 5.3 focuses on specific motions and their properties. In section 5.4 we state the Pontryagin Maximum Principle, and apply it to our situation for the time minimization problem. We finish by introducing definitions and results to characterize properties of optimal paths with an emphasis on singular extremals. Finally in section 5.5 we make concluding remarks and discuss potential future research.

5.2 Dynamics of Quadcopters

In this section we develop the equations of motion under certain reasonable symmetry assumptions regarding the shape of the quadcopter body. including the equations explicitly in coordinates. To do this we will be replicating some of the work done in [6]. We also add restrictions on the equilibrium solution of the system and restrictions on movement along the body axis. We finally include some simulations.

5.2.1 Rigid Body Equations

We assume the quadcopter is a rigid body in a fluid; later we will make certain additional assumptions explicit. As in [6] we will be closely following the derivations in [7, 15]. The configuration space [1] is:

$$Q = SE(3) \cong \mathbb{R}^3 \ltimes SO(3) \tag{5.1}$$

equipped with coordinates (b, R) where $b \in \mathbb{R}^3$ represents the position of the center of mass of the body in space and $R \in \mathbb{R}^{3\times 3}$ is a rotation matrix representing the orientation of the body aligned with the principal axes of inertia. This coordinate system corresponds to some inertial frame of reference. We use the notation of the *hat map*, the Lie algebra isomorphism:

$$\hat{}: (\mathbb{R}^3, \times) \to (\mathfrak{so}(3), [,]) \tag{5.2}$$

given by $\hat{y}z = y \times z$, equivalently:

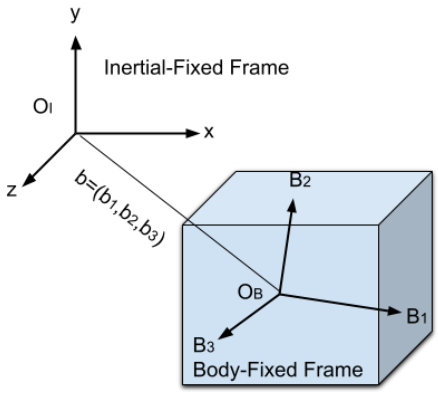


Fig. 5.1: Inertial (O_i) and Body-Fixed (O_B) Reference Frames.

$$\hat{y} = \begin{pmatrix} 0 & -y_3 & y_2 \\ y_3 & 0 & -y_1 \\ -y_2 & y_1 & 0 \end{pmatrix}. \tag{5.3}$$

To the rigid body, we associate a non-inertial body-fixed frame whose axes coincide with the body's principle axes of inertia. The kinematic equations are given by [15]:

$$\dot{b} = Rv \tag{5.4}$$

$$\dot{R} = R\hat{\Omega}. \tag{5.5}$$

and we express momenta between the inertial and body frames according to:

$$p = RP \tag{5.6}$$

$$\pi = R\Pi + \hat{b}p, \tag{5.7}$$

where p is the translational momentum in the inertial frame, P is the translational momentum in the body frame, π is the angular momentum in the inertial frame and Π is the angular momentum in the body frame [15]. Differentiating equations (5.6) and (5.7), we obtain:

$$\dot{p} = \dot{R}P + R\dot{P} \tag{5.8}$$

$$\dot{\pi} = \dot{R}\Pi + R\dot{\Pi} + \dot{b} \times p + b \times \dot{p}. \tag{5.9}$$

To rewrite these equations in a single frame, we must introduce additional quantities. In the body frame we have:

$$E_F = \sum_{i=1}^{k} R^{-1} f_i \tag{5.10}$$

$$E_\tau = \sum_{i=1}^{l} R^{-1} \tau_i, \tag{5.11}$$

where E_F is the total external force in the body frame, f_i is the i^{th} external force in the inertial frame, for $i = 1, \ldots, k$, E_τ is the total external torque in the body frame and τ_i is the i^{th} external torque in the inertial frame, for $i = 1, \ldots, l$. Now in the inertial frame, the dynamics are solely due to external forces and torques:

$$\dot{p} = \sum_{i=1}^{k} f_i \tag{5.12}$$

$$\dot{\pi} = \sum_{i=1}^{k} \hat{x}_i f_i + \sum_{i=1}^{l} \tau_i, \tag{5.13}$$

where x_i is the vector from inertial frame origin to line of action of f_i. We then solve (5.8), (5.9) for $\dot{P}$ and $\dot{\Pi}$ to get:

$$\dot{P} = R^{-1}(\dot{p} - \dot{R}P) \tag{5.14}$$

$$\dot{\Pi} = R^{-1}(\dot{\pi} - \dot{R}\Pi - \dot{b} \times p - b \times \dot{p}). \tag{5.15}$$

For (5.14) if we then substitute using (5.5) and (5.12) we get:

$$\dot{P} = R^{-1}\Big(\sum_{i=1}^{k} f_i - R\hat{\Omega}P\Big) \tag{5.16}$$

which is equivalent to:

$$\dot{P} = E_f + \hat{P}\Omega. \tag{5.17}$$

For (5.15) if we substitute using (5.13), (5.5), (5.4) and (5.12) we get:

$$\dot{\Pi} = R^{-1}\Big(\sum_{i=1}^{k} \hat{x}_i f_i + \sum_{i=1}^{l} \tau_i - R\hat{\Omega}\Pi - (Rv) \times p - b \times \Big(\sum_{i=1}^{k} f_i\Big)\Big) \tag{5.18}$$

which is equivalent to:

$$\dot{\Pi} = \hat{\Pi}\Omega + \hat{P}v + E_\tau + R^{-1}\sum_{i=1}^{k}(x_i - b) \times f_i. \tag{5.19}$$

This gives the evolution of the momenta in the body frame, but the equations mix momenta and velocities. In order to arrive at equations in terms of solely momenta or velocities, we must find explicit formulas relating the two. This can be achieved by the Legendre transform [15]:

$$P = \frac{\partial T}{\partial v}, \qquad \Pi = \frac{\partial T}{\partial \Omega}, \tag{5.20}$$

where T is the total kinetic energy of the system.

To make this explicit, we introduce the kinetic energies of the body and the fluid:

$$T_{body} = \frac{1}{2}\begin{pmatrix} v \\ \Omega \end{pmatrix}^t \begin{pmatrix} mI_3 & -m\hat{r}_{C_G} \\ m\hat{r}_{C_G} & J_b \end{pmatrix}\begin{pmatrix} v \\ \Omega \end{pmatrix} \tag{5.21}$$

$$T_{fluid} = \frac{1}{2}\begin{pmatrix} v \\ \Omega \end{pmatrix}^t \begin{pmatrix} M_f & C_f^t \\ C_f & J_f \end{pmatrix} \begin{pmatrix} v \\ \Omega \end{pmatrix} \tag{5.22}$$

Adding both we obtain the total kinetic energy of the system:

$$T = T_{body} + T_{fluid} = \frac{1}{2}\begin{pmatrix} v \\ \Omega \end{pmatrix}^t \begin{pmatrix} mI_3 + M_f & -m\hat{r}_{C_G} + C_f^t \\ m\hat{r}_{C_G} + C_f & J_b + J_f \end{pmatrix} \begin{pmatrix} v \\ \Omega \end{pmatrix} \tag{5.23}$$

where

- m is the mass of the body,
- r_{C_G} is the vector from the center of gravity to the body frame origin,
- J_b is the body inertia tensor,
- J_f is the added mass inertia tensor from displacing the surrounding fluid,
- M_f is the added mass from displacing the surrounding fluid,
- C_f is the added mass cross terms from displacing the surrounding fluid and
- I_3 is the 3×3 identity matrix.

Taking the partial differential of (5.23) in terms of v and Ω, and then using the Legendre transform, it simplifies into:

$$P = \frac{\partial T}{\partial v} = (mI_3 + M_f)v + \frac{1}{2}(C_f^t + C_f)\Omega \tag{5.24}$$

$$\Pi = \frac{\partial T}{\partial \Omega} = \frac{1}{2}(C_f^t + C_f)v + (J_b + J_f)\Omega. \tag{5.25}$$

These equations show the relations between velocities and momenta. We now assume the body has an especially nice shape, which is somewhat realistic for most quadcopters.

Assumption 1 *Assume the body has three planes of symmetry and the principle axes of inertial coincide with the body frame axes. When considering gravity in Section 5.2.2, we also assume that this symmetry extends to the mass density of the body.*

Assumption 2 *Assume the center of gravity of the body coincides with the origin of the body frame. This implies J_b, J_f, and M_f are all diagonal and $\hat{r}_{C_G} = C_f = 0$ [15].*

With these two assumptions, we can show that:

$$M\dot{v} = E_f + \hat{P}\Omega - \dot{M}v \tag{5.26}$$

$$J\dot{\Omega} = \widehat{J\Omega}\Omega + \widehat{Mv}v + E_\tau + R^{-1}\sum_{i=1}^{k}(x_i - b) \times f_i - \dot{J}\Omega \tag{5.27}$$

and thus we obtain the derivatives of the velocities expressed solely in terms of positions and velocities, without momenta. Finally, since for quadcopters the fluid is air, we make the following simplifying assumption.

Assumption 3 *Assume the added fluid is negligible. This implies $M_f = J_f = 0$. Thus $J = J_b$ and $M = mI_3$ which are constant matrices and thus $\dot{J} = \dot{M} = 0$ and the fictitious force $Mv \times v$ vanishes: $Mv \times v = 0$.*

Combining this with (5.4), (5.5), (5.26), (5.27), we obtain the complete equations of motion for a rigid body in air as a first order system on TQ coordinatized by (b, R, v, Ω):

$$\dot{b} = Rv \tag{5.28}$$

$$\dot{R} = R\hat{\Omega} \tag{5.29}$$

$$m\dot{v} = mv \times \Omega + E_F \tag{5.30}$$

$$J\dot{\Omega} = (J\Omega) \times \Omega + E_\tau + R^{-1} \sum_{i=1}^{k} (x_i - b) \times f_i. \tag{5.31}$$

5.2.2 External Forces and Torques

Next we explicitly work out the external terms E_F and E_τ for a quadcopter in the body frame. We have four rotors with the first located on the arm in the negative second body axis of the body frame and then going counter-clockwise as seen when viewing the quadcopter from the positive third body axis of the body frame, see Figure 5.2.

These rotors produce thrust, with the ith rotor having angular velocity ω_i and thus producing thrust $K_r\omega_i^2$, where K_r is the thrust coefficient [12]. Here we are implicitly treating the ω_is as our controls, which we also assume are all non-negative.

External Forces

For the external forces, we follow [16]. Three forces are to be considered: drag, gravity and thrust.

Drag We assume that the air density is constant for given altitude, and thus the coefficient of drag, which depends on body configuration and orientation, is constant in the body frame. Thus using the common quadratic model, the force due to drag in the body frame is:

$$R^{-1} f_1 = -\text{diag}(v_1|v_1|, v_2|v_2|, v_3|v_3|)C_D, \tag{5.32}$$

where v_j is the jth component of the linear velocity v and C_D is the vector of translational drag coefficients in the body frame.

Gravity The force due to gravity in the body frame is:

$$R^{-1} f_2 = -mgR^{-1} e_3^I, \tag{5.33}$$

where g is the gravitational constant and the third inertial frame basis vector e_3^I represents the direction of gravitational attraction.

Thrust Lastly, the force produced by thrust in the body frame is:

$$R^{-1} f_3 = \sum_{i=1}^{4} e_3^B K_r \omega_i^2, \tag{5.34}$$

where the third body frame basis vector e_3^B represents the direction in which each of the rotors produce thrust.

Thus the total external force is given by $E_F = R^{-1}(f_1 + f_2 + f_3)$.

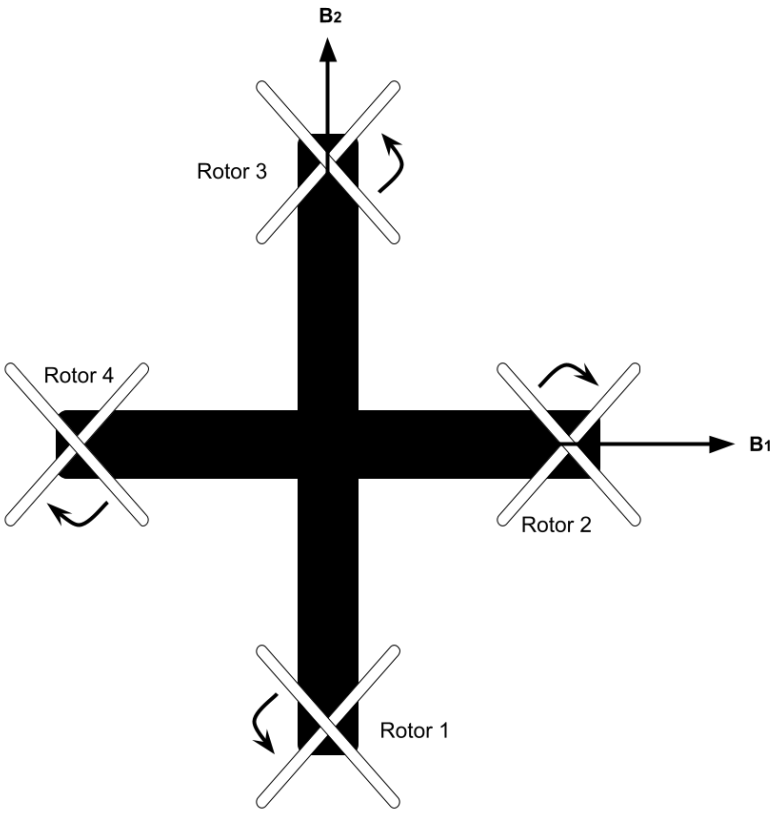


Fig. 5.2: Rotor positions as viewed looking down onto the quadcopter from the positive third body frame axis.

External Torques

Ostensibly, each of these forces induces a torque, represented by the last term $R^{-1}\sum_{i=1}^{k}(x_i - b)\times f_i$ in equation (5.31). For $R^{-1}f_1$ and $R^{-1}f_2$, no torque is induced as the line of force passes through the center of mass of the body due to our symmetry assumptions. The force $R^{-1}f_3$, however, represents the sum of four individual forces located at each of the four rotors. Each of these forces induces a torque and their sum gives the net torque generated by the rotors [17]:

$$\tau_f = \begin{pmatrix} K_r d(\omega_3^2 - \omega_1^2) \\ K_r d(\omega_4^2 - \omega_2^2) \\ K_d \sum_{i=1}^{4}(-1)^{i+1}\omega_i^2 \end{pmatrix}, \tag{5.35}$$

where K_r is the lift coefficient, K_d is the propeller drag coefficient and d is the distance from the quadcopter's center of mass to the rotation axis of each rotor and ω_i is the angular velocity of the i^{th} rotor.

The torque in the body frame generated by the drag is [16]:

$$R^{-1}\tau_1 = -\mathrm{diag}(\Omega_1|\Omega_1|, \Omega_2|\Omega_2|, \Omega_3|\Omega_3|)C_\tau, \tag{5.36}$$

where Ω_j is the jth component of the angular velocity Ω and C_τ is the vector of rotational drag coefficients in the body frame. The torque due to the gyroscopic effects is given by [5]:

$$R^{-1}\tau_2 = \sum_{i=1}^{4} \Omega \times J_r(-1)^{i+1}(0, 0, \omega_i)^t, \tag{5.37}$$

where J_r is the moment of inertia for a rotor. The total external torque is the sum of both: $E_\tau = R^{-1}(\tau_1 + \tau_2)$.

5.2.3 Equations of Motion for a Quadcopter

Using the prior developments, we obtain the complete set of dynamical equations of motion for a quadcopter:

$$\dot{b} = Rv \tag{5.38}$$

$$\dot{R} = R\hat{\Omega} \tag{5.39}$$

$$m\dot{v} = mv \times \Omega + R^{-1}(f_1 + f_2 + f_3) \tag{5.40}$$

$$J\dot{\Omega} = (J\Omega) \times \Omega + \tau_f + R^{-1}(\tau_1 + \tau_2). \tag{5.41}$$

where f_1, f_2, f_3, τ_f, τ_1 and $R^{-1}\tau_2$ are given by (5.32) through (5.37). We introduce $\eta = (b_1, b_2, b_3, \phi, \theta, \psi)^t$ on $Q = SE(3)$ representing the position in the inertial frame and rotations of the quadcopter to go from body frame to inertial frame. The positions b_1, b_2, b_3 are the standard coordinates for $\mathbb{R}^3$. The angles ϕ, θ, ψ are Tait-Bryan angles, known in aeronautics as *roll, pitch* and *yaw* respectively and sometimes referred to as Euler angles, though Classical Euler angles represent rotations around only two of the three given axis while Tait-Bryan angles represent rotations around all three in a given sequence.

Proposition 1. *In coordinates the equations of motion for a quadcopter take the form:*

$$\dot{b}_1 = v_1 C_\psi C_\theta + v_2(C_\psi S_\theta S_\phi - S_\psi C_\phi) + v_3(C_\psi S_\theta C_\phi + S_\psi S_\phi) \tag{5.42}$$

$$\dot{b}_2 = v_1 S_\psi C_\theta + v_2(S_\psi S_\theta S_\phi + C_\psi C_\phi) + v_3(S_\psi S_\theta C_\phi - C_\psi S_\phi) \tag{5.43}$$

$$\dot{b}_3 = -v_1 S_\theta + v_2 C_\theta S_\phi + v_3 C_\theta C_\phi \tag{5.44}$$

$$\dot{\phi} = \Omega_1 + \Omega_2 S_\phi \frac{S_\theta}{C_\theta} + \Omega_3 C_\phi \frac{S_\theta}{C_\theta} \tag{5.45}$$

$$\dot{\theta} = \Omega_2 C_\phi - \Omega_3 S_\phi \tag{5.46}$$

$$\dot{\psi} = \Omega_2 \frac{S_\phi}{C_\theta} + \Omega_3 \frac{C_\phi}{C_\theta} \tag{5.47}$$

$$\dot{v}_1 = v_2\Omega_3 - v_3\Omega_2 - \frac{1}{m}v_1|v_1|C_{D_1} + gS_\theta \tag{5.48}$$

$$\dot{v}_2 = v_3\Omega_1 - v_1\Omega_3 - \frac{1}{m}v_2|v_2|C_{D_2} - gC_\theta S_\phi \tag{5.49}$$

$$\dot{v}_3 = v_1\Omega_2 - v_2\Omega_1 + \frac{1}{m}\left(\sum_{i=1}^{4} K_r\omega_i^2 - v_3|v_3|C_{D_3}\right) - gC_\theta C_\phi \tag{5.50}$$

$$\dot{\Omega}_1 = \frac{1}{J_1}\left[(J_2 - J_3)\Omega_2\Omega_3 + J_r\sum_{i=1}^{4}(-1)^{i+1}\Omega_2\omega_i + K_r d(\omega_3^2 - \omega_1^2) - \Omega_1|\Omega_1|C_{\tau_1}\right] \tag{5.51}$$

$$\dot{\Omega}_2 = \frac{1}{J_2}\left[(J_3 - J_1)\Omega_1\Omega_3 - J_r\sum_{i=1}^{4}(-1)^{i+1}\Omega_1\omega_i + K_r d(\omega_4^2 - \omega_2^2) - \Omega_2|\Omega_2|C_{\tau_2}\right] \tag{5.52}$$

$$\dot{\Omega}_3 = \frac{1}{J_3}\left[(J_1 - J_2)\Omega_1\Omega_2 + K_d\sum_{i=1}^{4}(-1)^{i+1}\omega_i^2 - \Omega_3|\Omega_3|C_{\tau_3}\right]. \tag{5.53}$$

Definition 1. *We refer to the model given by equations (5.42)-(5.53) as the complete model.*

5.2.4 Parameter Values for Simulations

We provide in Table 5.1 the parameters values we use in this paper for all simulations. Most of the values were taken from [11], only the values for K_r, C_D and C_τ differ with the first two taken from [4] instead.

Constant	Value	Unit
m	0.468	Kg
C_D	$[5.5670, 5.5670, 6.3540] \times 10^{-4}$	N/m/s
g	9.81	m/s^2
J_r	4.856×10^{-3}	Nm/rad^2/s^2
K_r	2.9842×10^{-5}	Nm/rad/s
d	0.225	m
J	$[4.856, 4.856, 8.801] \times 10^{-3}$	Nm/rad/s^2
C_τ	$5 \times [5.5670, 5.5670, 6.3540] \times 10^{-2}$	Nm/rad/s
K_d	1.140×10^{-7}	Nm/rad/s

Table 5.1: Parameter values used for our simulations.

5.3 Basic Motions for Quadcopters

We first explore some simple motions for a quadcopter. Such analysis is lacking in the literature, and it is actually non trivial to present even basic motions for a quadcopter that are realistic and account for all forces.

5.3.1 Hovering Equilibrium

We define a hovering equilibrium and explain how to achieve it using the rotors.

Proposition 2. *An equilibrium solution to the dynamics of a quadcopter must satisfy* $\phi = \theta = v_1 = v_2 = v_3 = \Omega_1 = \Omega_2 = \Omega_3 = 0$*, and the four rotor's angular velocities* ω_i *need to be equal, and satisfy* $\omega_i^2 = \frac{1}{4K_r}mg$ *for* $i = 1, \cdots, 4$.

Proof. At an equilibrium point the right hand-sides of equations (5.42) to (5.53) must be zero. Since R is a rotation, it is invertible for any ϕ, θ and ψ. Thus since $\dot{b} = Rv$, we have that $\dot{b} \equiv 0$ implies $v \equiv 0$. Equation (5.48) then becomes $\dot{v}_1 = gS_\theta$ which means that $\theta = 0$. Similarly equation (5.49) becomes $\dot{v}_2 = -gS_\phi$ which gives that $\phi = 0$. It implies that equations (5.45), (5.46) and (5.47) become $\dot{\phi} = \Omega_1 = 0$, $\dot{\theta} = \Omega_2 = 0$ and $\dot{\psi} = \Omega_3 = 0$. From equations (5.51) and (5.52) we deduce that $K_r d(\omega_3^2 - \omega_1^2) = K_r d(\omega_4^2 - \omega_2^2) = 0$ and thus $\omega_3 = \omega_1$ and $\omega_4 = \omega_2$. From equation (5.53) we also have that $K_d \sum_{i=1}^4 (-1)^{i+1}\omega_i^2 = 0$ and thus $\omega_1 = \omega_2 = \omega_3 = \omega_4$. Finally, since $\dot{v}_3 = 0$, the rotors need to balance the gravity acceleration which implies that $\frac{1}{m}(\sum_{i=1}^4 K_r\omega_i^2) - g = 0$ which is equivalent to $\sum_{i=1}^4 \omega_i^2 = 4\omega_1^2 = \frac{m}{K_r}g$.

The next characterization follows directly from Proposition 2.

Corollary 1. *The set of equilibrium solutions to the dynamics of a quadcopter is* $\mathcal{Q} \subseteq Q$ *where*

$$\mathcal{Q} = \{(b_1, b_2, b_3, \phi, \theta, \psi, v_1, v_2, v_3, \Omega_1, \Omega_2, \Omega_3) |\ \phi = \theta = v_i = \Omega_i = 0\} \tag{5.54}$$

is a four parameter family of our configuration space with all the rotors angular velocities being equal to $\frac{1}{4K_r}mg$.

By Corollary 1 an equilibrium corresponds to a quadcopter staying in one place in the air with the angles corresponding to pitch and roll at zero and the thrust generated by the rotors balancing the gravity force ($\sum_{i=1}^4 K_r\omega_i^2 = mg$). This is typically the position that a quadcopter takes before starting a pre-planned mission, see Figure 5.3. We thus introduce the following definition.

Definition 2 (Hovering Equilibrium). *We define an equilibrium solution to the dynamics of a quadcopter as a hovering equilibrium. It must satisfy the conditions given in Proposition 2.*

Remark 1. We note that since, outside of the thrust created by the rotors, the only forces in the equations given by (5.48) through (5.53) are the drag (proportional to the associated velocity squared) and the gravity. Hence in order to reach a hovering equilibrium without relying on drag only, which is proportional to the velocity and thus can be quite low, we need to set the rotor's angular velocity ω_i such that the force created by the rotors plus the force of gravity is in the opposite direction to the direction of the motion.

5.3.2 Yaw Rotational Motion

We here study a purely rotational motion around the third axis in the body-fixed frame. We do assume we are at a hovering equilibrium. This motion is common in practice for instance when the quadcopter needs to execute a sequence of transects along a square and at the same time keep some prescribed orientations.

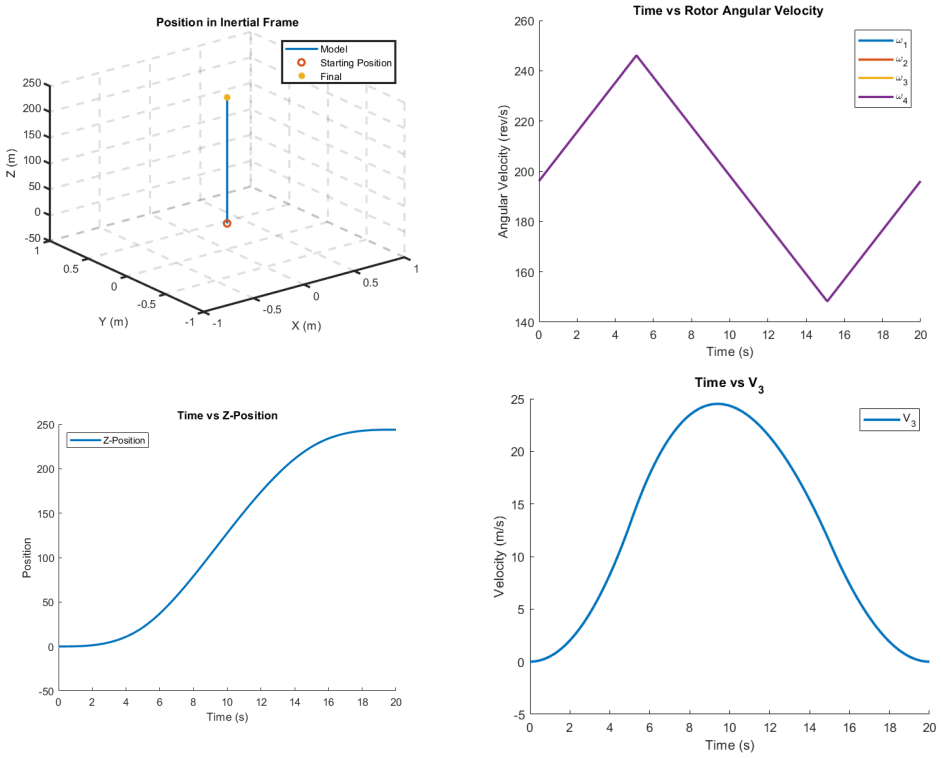


Fig. 5.3: Simulation of a quadcopter starting from a hovering equilibrium at $b_0 = (0, 0, 0)$ and then moving vertically before reaching and stabilizing at another hovering equilibrium at $b = (0, 0, 243.9388)$ over 20 seconds.

Proposition 3. *Assume the quadcopter is at a hovering equilibrium. Then the rotors for a pure yaw rotation must satisfy:*

$$\omega_1^2 = \omega_3^2 = \frac{mg}{2K_r} - \omega_2^2 \qquad \textit{and} \qquad \omega_4^2 = \omega_2^2, \tag{5.55}$$

where ω_2 is a free parameter.

Proof. We are assuming the quadcopter's configuration belongs to $\mathscr{Q}$ at the beginning of the motion and that throughout the rotation all variables are kept constants except Ω_3 and ψ. This imposes:

$$\dot{b}_i = 0, \qquad i = 1, \cdots 3 \tag{5.56}$$

$$\dot{\phi} = \dot{\theta} = v_1 = v_2 = \Omega_1 = \Omega_2 = 0, \tag{5.57}$$

for the duration of the motion. Then, from equations (5.51) and (5.52) we deduce that $\omega_1^2 = \omega_3^2$ and $\omega_2^2 = \omega_4^2$. Since we have $\theta = \phi = 0$ throughout the motion, we deduce that $\dot{b}_3 = v_3 = 0$ and therefore $\dot{v}_3 = \frac{1}{m}\sum_{i=1}^4 K_r\omega_i^2 - g = 0$. Using that $\frac{1}{m}\sum_{i=1}^4 K_r\omega_i^2 = \frac{2K_r}{m}(\omega_1^2 + \omega_2^2) = g$ we obtain the desired result.

Figure 5.4 represents such pure rotation. It shows the yaw rotation (left plot) as well as the angular velocity of the rotors (right plot).

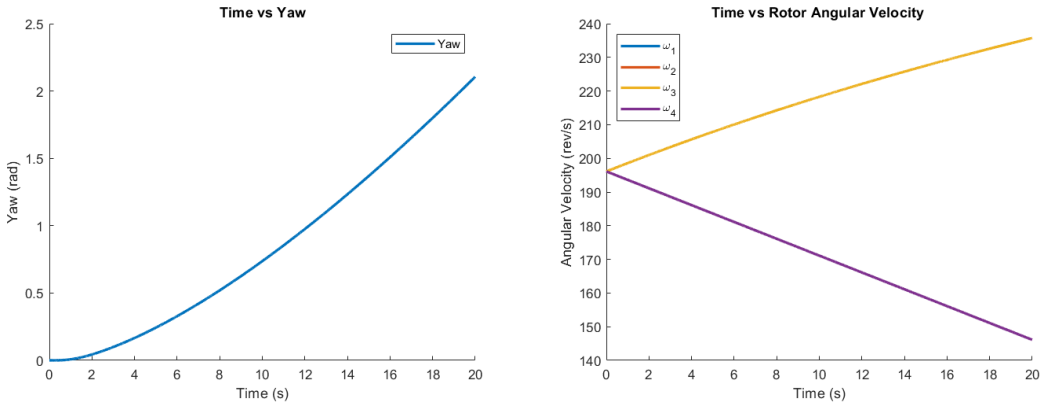


Fig. 5.4: Simulation of a quadcopter starting at hovering equilibrium at $\eta = (0, 0, 100, 0, 0, 0)^t$ and rotating clockwise around the z-axis. It can be actually verified that the condition between ω_1 and ω_2 in (5.55) is satisfied.

5.3.3 Translational Motion Restricted to the Plane of the First Two Axes

The following result establishes a condition that needs to be satisfied for a quadcopter to move from a hovering equilibrium in the horizontal xy-plane.

Proposition 4. *Assume the quadcopter is at a hovering equilibrium. To move in the xy-plane the quadcopter needs to generate a non-zero pitch or a roll.*

Proof. From Proposition 2 we know that if the system is at a hovering equilibrium, the roll and pitch angles must be at zero: $\phi(0) = \theta(0) = 0$. If we assume that there is no change in either the pitch and the roll, we have $\phi(t) = \theta(t) = 0$ along the entire trajectory starting at the hovering equilibrium. This means using equations (5.45), (5.46) and (5.47) that:

$$\dot{\phi}(t) = \Omega_1(t) = 0 \tag{5.58}$$

$$\dot{\theta}(t) = \Omega_2(t) = 0 \tag{5.59}$$

for all t. In addition, we also have $\dot{\psi}(0) = \Omega_3(0) = 0$. Equations (5.48) and (5.49) then become:

$$\dot{v}_1 = v_2\Omega_3 - \frac{1}{m}v_1|v_1|C_{D_1} \tag{5.60}$$

$$\dot{v}_2 = -v_1\Omega_3 - \frac{1}{m}v_2|v_2|C_{D_2}. \tag{5.61}$$

throughout the trajectory. But since $v_1(0) = v_2(0) = 0$ at the hovering equilibrium we deduce that $\dot{v}_1(t) = \dot{v}_2(t) = 0$. Thus from (5.42) and (5.43) we can get $\dot{b}_1 = \dot{b}_2 = 0$ and the quadcopter does not move in the xy-plane.

Remark 2. While motion in the xy-plane requires a non-zero pitch or roll, a change in altitude is possible without a change in the angles. Indeed, as it can be seen in equation (5.50) the velocity $v_3(t)$ might be altered using the acceleration of the rotors to move the quadcopter vertically along the z-axis.

The next two lemmas will discuss the conditions on the pitch and roll for a motion along the x-axis.

Lemma 1. *Assume the quadcopter is at a hovering equilibrium. For a pure translation motion along the first body axis with $\dot{\psi}(t) = 0$ for all t, the quadcopter needs to generate a non-zero pitch.*

Proof. For a pure translation motion along the first body axis with $\dot{\psi} = 0$ we need $v_2 = v_3 = 0$ for the trajectory and thus by (5.48) we get:

$$\dot{v}_1 = -\frac{1}{m} v_1 |v_1| C_{D_1} + g S_\theta. \tag{5.62}$$

If we start at a hovering equlibrium then there must exist some t such that $\theta(t) \neq 0$ to have $\dot{v}_1(t) \neq 0$.

Remark 3. Using (5.49) we can prove that $\phi \neq 0$ for a pure translational movement in the second body axis under the same conditions since $C_\theta = 0$ is not possible for the domain of θ.

Lemma 2. *Assume the quadcopter is at a hovering equilibrium. For a pure translation motion along the first body axis with $\dot{\psi}(t) = 0$ for all t, then $\phi(t) = 0$ for all t.*

Proof. Assuming that $\psi(t) = v_2(t) = v_3(t) = 0$ for all t then from (5.47), (5.49) and (5.50) we get:

$$\dot{\psi} = \Omega_2 S_\phi + \Omega_3 C_\phi = 0 \tag{5.63}$$

$$\dot{v}_2 = -v_1 \Omega_3 - g C_\theta S_\phi = 0 \tag{5.64}$$

$$\dot{v}_3 = v_1 \Omega_2 + \frac{1}{m} \sum_{i=1}^{4} K_r \omega_i^2 - g C_\theta C_\phi = 0, \tag{5.65}$$

by multiplying (5.45) by C_θ.

Now assume that there exist a $t > 0$ such that $\phi(t) \neq 0$.

For that t we obtain by solving (5.63) for Ω_2, that we have $\Omega_2 = -\Omega_3 \frac{C_\phi}{S_\phi}$. Substituting this into (5.65) gives us:

$$-v_1 \Omega_3 \frac{C_\phi}{S_\phi} + \frac{1}{m} \sum_{i=1}^{4} K_r \omega_i^2 - g C_\theta C_\phi = 0. \tag{5.66}$$

Solving (5.64) for $-v_1 \Omega_3$ gives us $-v_1 \Omega_3 = g C_\theta S_\phi$, which if we then substitute that into (5.66) we get by canceling the S_ϕ:

$$g C_\theta C_\phi + \frac{1}{m} \sum_{i=1}^{4} K_r \omega_i^2 - g C_\theta C_\phi = 0. \tag{5.67}$$

Simplifying thus gives us that $\omega_i = 0$ for all $i = 1 \ldots 4$.

If we use this in (5.65) we get that $v_1 \Omega_2 - g C_\theta C_\phi = 0$ which if we solve for $g C_\theta$ we get:

$$gC_\theta = \frac{1}{C_\phi} v_1 \Omega_2. \tag{5.68}$$

Substituting (5.68) into (5.64) and simplifying by multiplying by C_ϕ gives:

$$-v_1 \Omega_3 C_\phi + v_1 \Omega_2 S_\phi = 0. \tag{5.69}$$

If $v_1 \neq 0$ then (5.69) becomes $\Omega_2 S_\phi - \Omega_3 C_\phi = 0$. But by (5.63) we know that $\Omega_2 S_\phi + \Omega_3 C_\phi = 0$, which cannot happen on the domain of ϕ if $\phi \neq 0$, so $v_1 = 0$.

But If $v_1 = 0$ then (5.64) becomes $-gC_\theta S_\phi = 0$ which means that $\phi = 0$ for that t and thus contradicting our assumption.

Remark 4. A similar process can show that $\theta(t) = 0$ for all t if we want to have a pure translational motion along the second body axis with $\dot{\psi}(t) = 0$ for all t.

The next proposition provides the rotor's acceleration for a motion along the first body axis. Note that to derive this result we neglect the gyroscopic forces, which will be explained in detail in Section 5.3.4.

Proposition 5. *Assume the quadcopter is at a hovering equilibrium, and neglect the gyroscopic forces. For a pure translational motion along the first body axis with $\dot{\psi} = 0$ the rotors must satisfy:*

$$\omega_1^2 = \omega_3^2 = \frac{m}{4K_r}(gC_\theta - v_1 \Omega_2) \qquad \textit{and} \qquad \omega_1^2 = \frac{1}{2}(\omega_2^2 + \omega_4^2). \tag{5.70}$$

It is interesting to note that for this motion the angular velocity of the rotors is given as a feedback of the state. There is also a free variable as ω_2 (or equivalently ω_4) can be chosen arbitrarily.

Proof. The goal is to move along the first body axis, with no motion in the other body axes. Along such motion, we have $\dot{\psi}$ as well as the linear velocities v_2 and v_3 as zero throughout this pure translational motion. From Lemma 2 we also have that $\phi(t) = 0$ for all t and that a pitch is however necessary. From the equation for $\dot{\psi}$ we get that $\Omega_3 = 0$ which then implies that $\Omega_1 = 0$ during the entire motion as well. We deduce from the equation for $\dot{\Omega}_1$ that $\omega_1^2 = \omega_3^2$. Then, we have that $\omega_1^2 = \frac{1}{2}(\omega_2^2 + \omega_4^2)$. Using the fact that $v_3 = \dot{v}_3 = 0$ during the motion since there is no change we obtain:

$$\omega_1^2 = \frac{m}{4K_r}(gC_\theta - v_1 \Omega_2). \tag{5.71}$$

Looking at the equations of motion it can be verified that the linear and angular velocities in the direction of a translation in the first body axis when starting from a hovering equilibrium are given by:

$$\dot{v}_1 = \frac{1}{m} v_1 |v_1| C_{D_1} + gS_\theta \tag{5.72}$$

$$\dot{\Omega}_2 = -\Omega_2 |\Omega_2| C_{\tau_2} + K_r d. \tag{5.73}$$

Remark 5. A similar calculation provides the conditions for a pure translation in the second body axis when starting from a hovering equilibrium. It will require to produce a roll instead of a pitch and the rotors must satisfy:

$$\omega_2^2 = \omega_4^2 = \frac{m}{4K_r}(gC_\phi - v_2 \Omega_1) \qquad \text{and} \qquad \omega_2^2 = \frac{1}{2}(\omega_1^2 + \omega_3^2). \tag{5.74}$$

Figure 5.5 displays a translational motion in the first body axis. It can be observed that it requires a pitch (top right plot). The body fixed frame velocity can be found in the bottom left plot and it can be observed that only v_1 varies. The angular velocities of the rotors satisfy equation (5.70) from Proposition 5. It can be observed that due to the pitch there is a motion in the inertial z-axis, this fact is due to the unique way quadcopters are propelled.

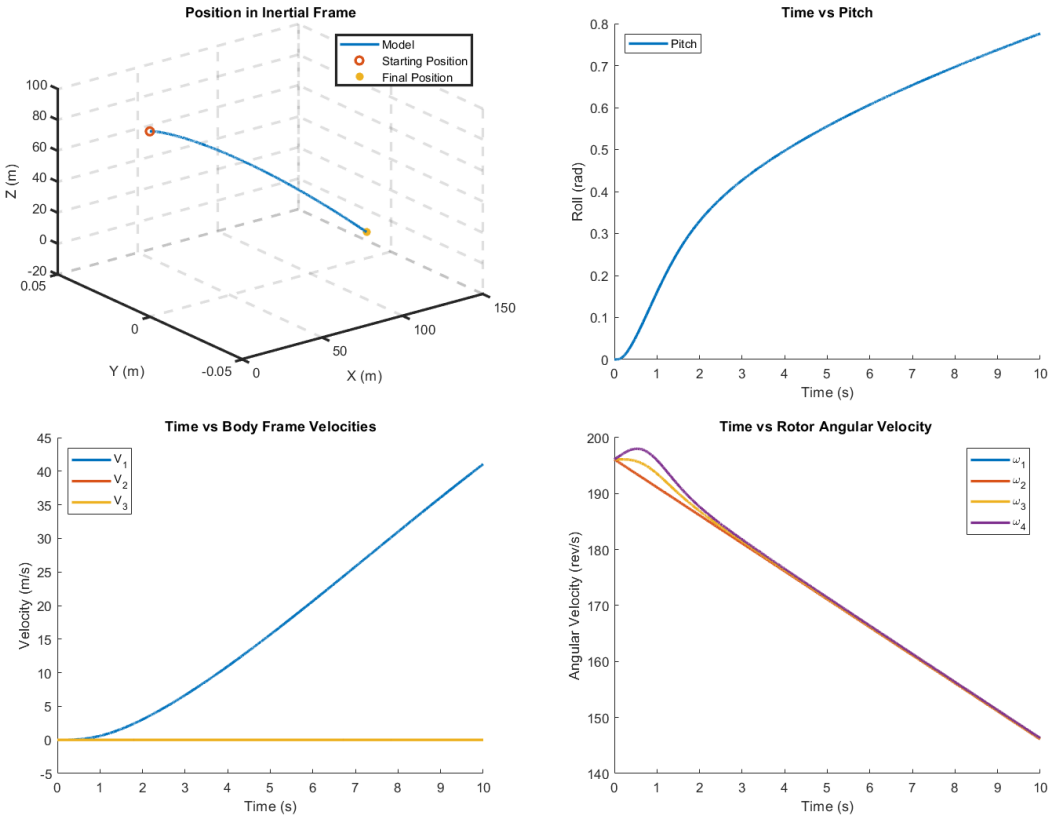


Fig. 5.5: Simulation of a quadcopter starting at hovering equilibrium at $\eta_0 = (0, 0, 100, 0, 0, 0)$ and going in the positive first body axis direction for 10 seconds, ending at $\eta_1 = (134.8851, 0, -2.9653, 0, 0.7762, 0)$.

5.3.4 Quadcopter as an Affine Control System

The goal of this section is to introduce an affine control system associated to the equations of motion for quadcopters. The components of the control are given by the ω_i, the rotor angular velocities. As it can be seen in the equations of motion in Proposition 1 they appear as quadratic terms as well as linear terms in the expressions for the $\dot{\Omega}_i$'s. Affine control systems have a structure that allows for

analytical work to characterize properties of extremals in the optimal control problem. To derive an affine control system from our equations of motion we will neglect the gyroscopic torque from the equations of motion, similar to [12, 16]. This is equivalent to neglecting the torque $R^{-1}\tau_2$, see expression (5.37), from equation (5.41). We make additional simplifications based on the work in [5, 8]. The first simplification assumes that the translational drag is linear in the velocity, and the second one is neglecting the angular drag.

Based on those simplifications, the drift of the affine control system for the simplified model, F_0 takes the form:

$$F_0(X) = \begin{pmatrix} v_1 C_\psi C_\theta + v_2(C_\psi S_\theta S_\phi - S_\psi C_\phi) + v_3(C_\psi S_\theta C_\phi + S_\psi S_\phi) \\ v_1 S_\psi C_\theta + v_2(S_\psi S_\theta S_\phi + C_\psi C_\phi) + v_3(S_\psi S_\theta C_\phi - C_\psi S_\phi) \\ -v_1 S_\theta + v_2 C_\theta S_\phi + v_3 C_\theta C_\phi \\ \Omega_1 + \Omega_2 S_\phi \frac{S_\theta}{C_\theta} + \Omega_3 C_\phi \frac{S_\theta}{C_\theta} \\ \Omega_2 C_\phi - \Omega_3 S_\phi \\ \Omega_2 \frac{S_\phi}{C_\theta} + \Omega_3 \frac{C_\phi}{C_\theta} \\ v_2\Omega_3 - v_3\Omega_2 - \frac{1}{m} v_1 C_{D_1} + g S_\theta \\ v_3\Omega_1 - v_1\Omega_3 - \frac{1}{m} v_2 C_{D_2} - g C_\theta S_\phi \\ v_1\Omega_2 - v_2\Omega_1 - \frac{1}{m} v_3 C_{D_3} - g C_\theta C_\phi \\ \frac{1}{J_1}[(J_2 - J_3)\Omega_2\Omega_3] \\ \frac{1}{J_2}[(J_3 - J_1)\Omega_1\Omega_3] \\ \frac{1}{J_3}[(J_1 - J_2)\Omega_1\Omega_2] \end{pmatrix}. \tag{5.75}$$

To simplify the notations, we will hereafter refer to K_r as α, $K_r d$ as β and K_d as γ and the control fields are given by:

$$F_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \frac{\alpha}{m} \\ -\frac{\beta}{J_1} \\ 0 \\ \frac{\gamma}{J_3} \end{pmatrix}, F_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \frac{\alpha}{m} \\ 0 \\ -\frac{\beta}{J_2} \\ -\frac{\gamma}{J_3} \end{pmatrix}, F_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \frac{\alpha}{m} \\ \frac{\beta}{J_1} \\ 0 \\ \frac{\gamma}{J_3} \end{pmatrix}, F_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \frac{\alpha}{m} \\ 0 \\ \frac{\beta}{J_2} \\ -\frac{\gamma}{J_3} \end{pmatrix}. \tag{5.76}$$

Proposition 6. *The quadcopter simplified model is an affine control system of the form:*

$$\dot{X} = F_0(X) + \sum_{i=1}^{4} F_i(X) u_i, \tag{5.77}$$

where $X = (b_1, b_2, b_3, \phi, \theta, \psi, v_1, v_2, v_2, \Omega_1, \Omega_2, \Omega_3)^t$, $u_i = \omega_i^2$ *and the vector fields* F_i *are given by (5.75) and (5.76). The domain of control for the quadcopters is taken as:*

$$\mathscr{F} = \{u \in \mathbb{R}^4 | 186^2 \leq u_i \leq 296^2 \text{ for } i = 1, \cdots, 4\}. \tag{5.78}$$

Remark 6. We note that the lower bound of each u_i for $i = 1, \ldots, 4$ is about $(196.1162 - 10)^2$ and the upper bound of each u_i is about $(196.1162 + 100)^2$. This is due to the fact that for the physical parameters given in Table 5.1 the angular velocity that the rotors need to be at for the quadcopter to be at hovering equilibrium is about $196.1162(rev/s)$.

5.4 Time Optimal Control for Quadcopters

In this section we analyze time optimal trajectories to steer a quadcopter from a starting configuration to an ending one. We use tools and concepts from geometric optimal control to analyze properties of the optimal trajectories with an emphasis on singular extremals. The maximum principle provides necessary conditions for a pair control-trajectory to be optimal. We follow the definition and construction methods outlined in [2, 10, 13]. We start by introducing some notations and stating the Maximum principle.

Consider a control system which can be expressed as a system of equations of the form:

$$\dot{x}(t) = f(x(t), u(t)), \tag{5.79}$$

where f is assumed to be $\mathbf{C}^1$ in x, and is continuous in u. We denote by $u : [t_0, t_f] \to \mathscr{F}$ the control function with $t_0 \leq t_f < \infty$, as well as $\mathscr{F} \subset \mathbb{R}^m$ as being the control domain and by $x : [t_0, t_f] \to \mathbb{R}^n$ the associated trajectory.

In the sequel, the control is said to be admissible if it is a measurable bounded function $u : [t_0, t_f] \to \mathscr{F}$ where the domain of control is assumed to be of the following form:

$$\mathscr{F} = \{u \in \mathbb{R}^m | \alpha_i^{min} \leq u_i \leq \alpha_i^{max}, \alpha_i^{min} < \alpha_i^{max} \text{ for } i = 1, \cdots, m\} \tag{5.80}$$

This is justified by the fact that our control for the quadcopters is given by the rotors angular accelerations: $u_i = \omega_i^2$'s which are individually bounded by a min and a max value.

For $x_0, x_1 \in \mathbb{R}^n$, fixed initial and final states, we assume that there exists an admissible control function $u : [t_0, t_f] \to \mathscr{F}$, and a corresponding absolutely continuous trajectory $x : [t_0, t_f] \to \mathbb{R}^n$ that satisfies equation (5.79) and $x(t_0) = x_0, x(t_f) = x_1$. Let us then consider a cost of the form:

$$J(u) = \int_{t_0}^{t_f} L(x(t), u(t))dt, \tag{5.81}$$

where $L : \mathbb{R} \times \mathbb{R}^n \times \mathscr{F} \to \mathbb{R}$ is the running cost which needs to be continuous with continuous first derivative over x. The goal is to minimize, for each pair of initial and final states for which the system is controllable, the cost $J(u)$ over the set of admissible controls steering the system from x_0 to x_1.

5.4.1 Maximum Principle

The maximum principle is a classical tool from optimal control theory, it provides necessary conditions for an admissible control u and its corresponding trajectory x to be optimal. The Maximum principle for a fixed end-state, which can be found in [10], is stated below.

Theorem 4 (Fixed Endpoint Maximum Principle). *Let $u^* : [t_0, t_f] \to \mathscr{F}$ be an admissible optimal control and let $x^* : [t_0, t_f] \to \mathbb{R}^n$ be the corresponding optimal state trajectory with initial condition $x^*(t_0) = x_0$, and final condition $x^*(t_f) = x_1$. Then there exists an absolutely continuous function $p^* : [t_0, t_f] \to \mathbb{R}^n$, called the adjoint vector, and a constant $p_0^* \leq 0$ satisfying $(p_0^*, p^*(t)) \neq (0, 0)$ for all $t \in [t_0, t_f]$ that have the following properties:*

1. *x^* and p^* satisfy the Hamiltonian equations almost everywhere on $[t_0, t_f]$:*

$$\begin{aligned} \dot{x}^*(t) &= H_p(x^*(t), u^*(t), p^*(t), p_0^*) \\ \dot{p}^*(t) &= -H_x(x^*(t), u^*(t), p^*(t), p_0^*) \end{aligned} \tag{5.82}$$

where the Hamiltonian function $H : \mathbb{R}^n \times \mathscr{F} \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ is defined as:

$$H(x, u, p, p_0) := \langle p, f(x, u) \rangle + p_0 L(x, u). \tag{5.83}$$

2. *For almost every $t \in [t_0, t_f]$, the function $u \mapsto H(x^*(t), u, p^*(t), p_0^*)$ has a global maximum at $u(t) = u^*(t)$, i.e.:*

$$H(x^*(t), u^*(t), p^*(t), p_0^*) \geq H(x^*(t), u, p^*(t), p_0^*) \qquad \forall u \in \mathscr{F}. \tag{5.84}$$

3. *$H(x^*(t), u^*(t), p^*(t), p_0^*) = 0$ for almost every $t \in [t_0, t_f]$.*

The Maximum Principle provides necessary conditions for an optimal control, and thus an optimal state trajectory, in other words it allows us to narrow down the set of admissible controls and corresponding trajectories that are candidates to optimality.

5.4.2 Time Minimization for Affine Control System

In this work we are interested in working with an affine control system, i.e. a system of the form:

$$\dot{x} = F_0(x) + \sum_{i=1}^{m} F_i(x) u_i, \tag{5.85}$$

and we consider the time minimization problem, where $J(u) = \int_{t_0}^{t_f} 1 dt$. The Hamiltonian function therefore takes the form:

$$H(x, u, p, p_0) = \langle p, F_0(x) \rangle + \sum_{i=1}^{m} \langle p, F_i(x) \rangle u_i + p_0, \tag{5.86}$$

where $p_0 \leq 0$.

The maximization condition of Theorem 4 states that an optimal control must maximize the expressions $\sum_{i=1}^{m} \langle p, F_i(x) \rangle u_i$ along optimal solutions. We deduce, using the assumption on the domain of control (5.80), that an optimal control u^* must satisfy for each component the following:

$$u_i^*(t) = \operatorname{sgn}(\langle p^*(t), F_i(x^*(t)) \rangle) := \begin{cases} \alpha_i^{max} & \text{if } \langle p^*(t), F_i(x^*(t)) \rangle > 0 \\ \alpha_i^{min} & \text{if } \langle p^*(t), F_i(x^*(t)) \rangle < 0 \\ ? & \text{if } \langle p^*(t), F_i(x^*(t)) \rangle = 0 \end{cases}. \tag{5.87}$$

We introduce the next few definitions [2, 7].

Definition 3. *We denote the i^{th} switching function for time minimization by:*

$$\epsilon_i(t) = \langle p(t), F_i(x(t)) \rangle \text{ for } i = 1, \ldots, m. \tag{5.88}$$

Definition 4. *An extremal, defined on an interval $I = [t_1, t_2]$, is a quadruplet (x, u, p, p_0) which satisfy the conditions of the Maximum principle.*

Remark 7. Since we are considering time minimization the condition $(p, p_0) \neq (0, 0)$ becomes simply $p \neq 0$. This is true because for time minimization the running cost, $L = 1$, is independent of u or x and thus all p_0 satisfy (5.82) and (5.84). Indeed, if $p(t) = 0$ then (5.83) and Condition 3 of the Maximum principle forces $p_0 = 0$.

Definition 5. *We say that a component u_i of the control is bang on a non-empty time interval if its corresponding function ϵ_i does not vanishes on that interval. It is said to be bang-bang on $[t_1, t_2]$ if $\epsilon_i(t) \neq 0$ for almost every $t \in [t_1, t_2]$.*

Definition 6. *If there is a interval of measure nonzero $[t_1, t_2]$ such that $\epsilon_i(t) = 0$ for all $t \in [t_1, t_2]$, the corresponding component of the control u_i is said to be singular on $[t_1, t_2]$. A singular component of the control is said to be strict if the other controls are bang.*

Definition 7. *A singular extremal is an extremal where a component of the control is singular, and if all components of the control are singular then it is called totally singular. A regular extremal is an extremal with no singular components of the control.*

5.4.3 Computation of Bang Controls

The maximum principle implies that a bang-bang component of the control takes the form:

$$u_i(t) = \begin{cases} \alpha_i^{max} & \epsilon_i(t) > 0 \\ \alpha_i^{min} & \epsilon_i(t) < 0. \end{cases} \tag{5.89}$$

5.4.4 Computation of Singular Controls

If a switching function $\epsilon_i(t)$ is equal to zero more investigation is needed by looking at its first and second derivative. The derivatives involve the notion of Lie brackets in $\mathbb{R}^n$. The following results are well-know [2].

Lemma 3. *The derivative of the switching function is given by:*

$$\dot{\epsilon_i}(t) = \langle p(t), [F_0, F_i](x(t)) \rangle + \sum_{j=1}^{m} \langle p(t), [F_j, F_i](x(t)) \rangle u_j(t). \tag{5.90}$$

Lemma 4. *Assuming the system satisfies $[F_j, F_i](x) \equiv 0$ for all i, j, then the second order derivative of the switching function exists and is given by:*

$$\ddot{\epsilon}_i(t) = \langle p(t), ad_{F_0}^2 F_i(x(t)) \rangle + \sum_{j=1}^{m} \langle p(t), [F_j, [F_0, F_i]](x(t)) \rangle u_j(t). \tag{5.91}$$

Proposition 7. *From Definition 6, along a singular control u_i on $t \in [t_1, t_2]$ and with the same assumptions as in Lemma 4, we must have $\epsilon_i(t) = \dot{\epsilon}_i(t) = \ddot{\epsilon}_i(t) = 0$ for almost all $t \in [t_1, t_2]$. Thus using Lemmas 3 and 4, the following conditions must be satisfied:*

$$\langle p(t), [F_0, F_i](x(t))\rangle = 0 \tag{5.92}$$

and:

$$\langle p(t), ad^2_{F_0} F_i(x(t))\rangle + \sum_{j=1}^{m} \langle p(t), [F_j, [F_0, F_i]](x(t))\rangle u_j(t) = 0. \tag{5.93}$$

We now apply the results above to the time-optimal control problem for a quadcopter.

5.4.5 Hamiltonian and Switching Functions

Applying the maximum principle to the quadcopter affine control system (5.77), from Equation (5.86) we know that for the time minimization problem the Hamiltonian function is given by:

$$H(X, u, p, p_0) = \langle p, F_0(X)\rangle + \sum_{i=1}^{4} \langle p, F_i(X)\rangle u_i - p_0. \tag{5.94}$$

Proposition 8. *The switching functions for the simplified model for quadcopters are given by:*

$$\epsilon_1 = \langle p, F_1\rangle = \frac{\alpha}{m} p_9 - \frac{\beta}{J_1} p_{10} + \frac{\gamma}{J_3} p_{12}, \tag{5.95}$$

$$\epsilon_2 = \langle p, F_2\rangle = \frac{\alpha}{m} p_9 - \frac{\beta}{J_2} p_{11} - \frac{\gamma}{J_3} p_{12}, \tag{5.96}$$

$$\epsilon_3 = \langle p, F_3\rangle = \frac{\alpha}{m} p_9 + \frac{\beta}{J_1} p_{10} + \frac{\gamma}{J_3} p_{12}, \tag{5.97}$$

$$\epsilon_4 = \langle p, F_4\rangle = \frac{\alpha}{m} p_9 + \frac{\beta}{J_2} p_{11} - \frac{\gamma}{J_3} p_{12}. \tag{5.98}$$

Proof. It is a direct calculation of the terms $\langle p, F_i(X)\rangle$ for $i = 1, \cdots, 4$.

5.4.6 Adjoint Equations

By the maximum principle trajectories for optimality are solutions of the Hamiltonian equations.

Proposition 9. *The adjoint vector for the Hamiltonian in (5.94) satisfies the following equations:*

$$\dot{p}(t) = -(F_0)_X(X)p. \tag{5.99}$$

where $(F_0)_X(X) = (\mathfrak{X}_{i,j}(X))_{1 \le i,j \le 12}$ is a 12×12 matrix whose non-zero entries are in Table 5.2.

$\mathfrak{X}_{4,1} = v_2(C_\psi S_\theta C_\phi + S_\psi S_\phi) + v_3(S_\psi C_\phi - C_\psi S_\theta S_\phi)$	
$\mathfrak{X}_{4,2} = v_2(S_\psi S_\theta S_\phi + C_\psi C_\phi) + v_3(S_\psi S_\theta C_\phi - C_\psi S_\phi)$	
$\mathfrak{X}_{4,3} = v_2 C_\theta C_\phi - v_3 C_\theta S_\phi$	$\mathfrak{X}_{4,4} = \Omega_2 C_\phi \frac{S_\theta}{C_\theta} - \Omega_3 S_\phi \frac{S_\theta}{C_\theta}$
$\mathfrak{X}_{4,5} = -\Omega_2 S_\phi - \Omega_3 C_\phi$	$\mathfrak{X}_{4,6} = \Omega_2 \frac{C_\phi}{C_\theta} - \Omega_3 \frac{S_\phi}{C_\theta}$
$\mathfrak{X}_{4,8} = -gC_\theta C_\phi$	$\mathfrak{X}_{4,9} = gC_\theta S_\phi$
$\mathfrak{X}_{5,1} = -v_1 C_\psi S_\theta + v_2 C_\psi C_\theta S_\phi + v_3 C_\psi C_\theta C_\phi$	$\mathfrak{X}_{5,2} = -v_1 S_\psi S_\theta + v_2 S_\psi C_\theta S_\phi + v_3 S_\psi C_\theta C_\phi$
$\mathfrak{X}_{5,3} = -v_1 C_\theta - v_2 S_\theta S_\phi - v_3 S_\theta C_\phi$	$\mathfrak{X}_{5,4} = \frac{1}{C_\theta^2}(\Omega_2 S_\phi + \Omega_3 C_\phi)$
$\mathfrak{X}_{5,6} = \frac{S_\theta}{C_\theta^2}(\Omega_2 S_\phi + \Omega_3 C_\phi)$	$\mathfrak{X}_{5,7} = gC_\theta$
$\mathfrak{X}_{5,8} = gS_\theta S_\phi$	$\mathfrak{X}_{5,9} = gS_\theta C_\phi$
$\mathfrak{X}_{6,1} = -v_1 S_\psi C_\theta - v_2(S_\psi S_\theta S_\phi + C_\psi C_\phi) + v_3(C_\psi S_\phi - S_\psi S_\theta C_\phi)$	
$\mathfrak{X}_{6,2} = v_1 C_\psi C_\theta + v_2(C_\psi S_\theta S_\phi - S_\psi C_\phi) + v_3(C_\psi S_\theta C_\phi + S_\psi S_\phi)$	
$\mathfrak{X}_{7,1} = C_\psi C_\theta$	$\mathfrak{X}_{7,2} = S_\psi C_\theta$
$\mathfrak{X}_{7,3} = -S_\theta$	$\mathfrak{X}_{7,7} = -\frac{1}{m} C_{D_1}$
$\mathfrak{X}_{7,8} = -\Omega_3$	$\mathfrak{X}_{7,9} = \Omega_2$
$\mathfrak{X}_{8,1} = C_\psi S_\theta S_\phi - S_\psi C_\phi$	$\mathfrak{X}_{8,2} = S_\psi S_\theta S_\phi + C_\psi C_\phi$
$\mathfrak{X}_{8,3} = C_\theta S_\phi$	$\mathfrak{X}_{8,7} = \Omega_3$
$\mathfrak{X}_{8,8} = -\frac{1}{m} C_{D_2}$	$\mathfrak{X}_{8,9} = -\Omega_1$
$\mathfrak{X}_{9,1} = C_\psi S_\theta C_\phi + S_\psi S_\phi$	$\mathfrak{X}_{9,2} = S_\psi S_\theta C_\phi - C_\psi S_\phi$
$\mathfrak{X}_{9,3} = C_\theta C_\phi$	$\mathfrak{X}_{9,7} = -\Omega_2$
$\mathfrak{X}_{9,8} = \Omega_1$	$\mathfrak{X}_{9,9} = -\frac{1}{m} C_{D_3}$
$\mathfrak{X}_{10,4} = 1$	$\mathfrak{X}_{10,8} = v_3$
$\mathfrak{X}_{10,9} = -v_2$	$\mathfrak{X}_{10,11} = \frac{1}{J_2}(J_3 - J_1)\Omega_3$
$\mathfrak{X}_{10,12} = \frac{1}{J_3}(J_1 - J_2)\Omega_2$	$\mathfrak{X}_{11,4} = S_\phi \frac{S_\theta}{C_\theta}$
$\mathfrak{X}_{11,5} = C_\phi$	$\mathfrak{X}_{11,6} = \frac{S_\phi}{C_\theta}$
$\mathfrak{X}_{11,7} = -v_3$	$\mathfrak{X}_{11,9} = v_1$
$\mathfrak{X}_{11,10} = \frac{1}{J_1}(J_2 - J_3)\Omega_3$	$\mathfrak{X}_{11,12} = \frac{1}{J_3}(J_1 - J_2)\Omega_1$
$\mathfrak{X}_{12,4} = C_\phi \frac{S_\theta}{C_\theta}$	$\mathfrak{X}_{12,5} = -S_\phi$
$\mathfrak{X}_{12,6} = \frac{C_\phi}{C_\theta}$	$\mathfrak{X}_{12,7} = v_2$
$\mathfrak{X}_{12,8} = -v_1$	$\mathfrak{X}_{12,10} = \frac{1}{J_1}(J_2 - J_3)\Omega_2$
$\mathfrak{X}_{12,11} = \frac{1}{J_2}(J_3 - J_1)\Omega_1$	

Table 5.2: Non-zero entries of the matrix $(F_0)_X(X)$ which when multiplied by $-p(t)$ gives $\dot{p}(t)$.

Explicitly written out we get the following equations:

$$\dot{p}_1 = 0 \tag{5.100}$$

$$\dot{p}_2 = 0 \tag{5.101}$$

$$\dot{p}_3 = 0 \tag{5.102}$$

$$\begin{aligned}\dot{p}_4 = &- p_1[v_2(C_\psi S_\theta C_\phi + S_\psi S_\phi) + v_3(S_\psi C_\phi - C_\psi S_\theta S_\phi)] - p_2[v_2(S_\psi S_\theta S_\phi \\ &+ C_\psi C_\phi) + v_3(S_\psi S_\theta C_\phi - C_\psi S_\phi)] - p_3(v_2 C_\theta C_\phi - v_3 C_\theta S_\phi) \\ &- p_4\Big(\Omega_2 C_\phi \frac{S_\theta}{C_\theta} - \Omega_3 S_\phi \frac{S_\theta}{C_\theta}\Big) + p_5(\Omega_2 S_\phi + \Omega_3 C_\phi) - p_6\Big(\Omega_2 \frac{C_\phi}{C_\theta} - \Omega_3 \frac{S_\phi}{C_\theta}\Big) \\ &+ p_8 g C_\theta C_\phi - p_9 g C_\theta S_\phi\end{aligned} \tag{5.103}$$

$$\dot{p}_5 = - p_1(v_2 C_\psi C_\theta S_\phi - v_1 C_\psi S_\theta + v_3 C_\psi C_\theta C_\phi) - p_2(v_2 S_\psi C_\theta S_\phi - v_1 S_\psi S_\theta$$

$$+ v_3 S_\psi C_\theta C_\phi) + p_3(v_1 C_\theta + v_2 S_\theta S_\phi + v_3 S_\theta C_\phi) - \frac{1}{C_\theta^2} p_4(\Omega_2 S_\phi + \Omega_3 C_\phi)$$
$$- \frac{S_\theta}{C_\theta^2} p_6(\Omega_2 S_\phi + \Omega_3 C_\phi) - p_7 g C_\theta - p_8 g S_\theta S_\phi - p_9 g S_\theta C_\phi \tag{5.104}$$

$$\dot{p}_6 = p_1[v_1 S_\psi C_\theta + v_2(S_\psi S_\theta S_\phi + C_\psi C_\phi) - v_3(C_\psi S_\phi - S_\psi S_\theta C_\phi)]$$
$$- p_2[v_1 C_\psi C_\theta + v_2(C_\psi S_\theta S_\phi - S_\psi C_\phi) + v_3(C_\psi S_\theta C_\phi + S_\psi S_\phi)] \tag{5.105}$$

$$\dot{p}_7 = - p_1 C_\psi C_\theta - p_2 S_\psi C_\theta + p_3 S_\theta + \frac{1}{m} p_7 C_{D_1} + p_8 \Omega_3 - p_9 \Omega_2 \tag{5.106}$$

$$\dot{p}_8 = - p_1(C_\psi S_\theta S_\phi - S_\psi C_\phi) - p_2(S_\psi S_\theta S_\phi + C_\psi C_\phi) - p_3 C_\theta S_\phi - p_7 \Omega_3$$
$$+ \frac{1}{m} p_8 C_{D_2} + p_9 \Omega_1 \tag{5.107}$$

$$\dot{p}_9 = - p_1(C_\psi S_\theta C_\phi + S_\psi S_\phi) - p_2(S_\psi S_\theta C_\phi - C_\psi S_\phi) - p_3 C_\theta C_\phi + p_7 \Omega_2$$
$$- p_8 \Omega_1 + \frac{1}{m} p_9 C_{D_3} \tag{5.108}$$

$$\dot{p}_{10} = - p_4 - p_8 v_3 + p_9 v_2 - \frac{1}{J_2}(J_3 - J_1) p_{11} \Omega_3 - \frac{1}{J_3}(J_1 - J_2) p_{12} \Omega_2 \tag{5.109}$$

$$\dot{p}_{11} = - p_4 S_\phi \frac{S_\theta}{C_\theta} - p_5 C_\phi - \frac{S_\phi}{C_\theta} p_6 + p_7 v_3 - p_9 v_1 - \frac{1}{J_1}(J_2 - J_3) p_{10} \Omega_3$$
$$- \frac{1}{J_3}(J_1 - J_2) p_{12} \Omega_1 \tag{5.110}$$

$$\dot{p}_{12} = - \frac{S_\theta}{C_\theta} p_4 C_\phi + p_5 S_\phi - \frac{C_\phi}{C_\theta} p_6 - p_7 v_2 + p_8 v_1 - \frac{1}{J_1}(J_2 - J_3) p_{10} \Omega_2$$
$$- \frac{1}{J_2}(J_3 - J_1) p_{11} \Omega_1 \tag{5.111}$$

Proof. By direct computation using $\dot{p}(t) = -H_X(X(t), u(t), p(t), p_0)$ from equation (5.82) with F_0 through F_4 in (5.94) being given by (5.75) and (5.76) we get (5.100) through (5.111) with (5.99) being another form of the same equations.

Remark 8. It can be observed that the first three variables of the adjoint vector are integrals of motion, they are constant throughout the solutions of the Maximum principle.

5.4.7 Regular Extremals

Regular extremals play an important role in the optimal synthesis. The goal of this section is to provide some numerical simulations to illustrate their structure. All simulations presented in this section start at a hovering equilibrium with $\psi = 0$ which we chose as the origin of the inertial frame in our plots. Figures 5.6 through 5.8 show a regular extremal corresponding to the initial value $p_0 = (0, 0, 0, 0, 0, 0, 0, 0, 100, 0, 0, 0)$. As can be seen in Figure 5.6 this corresponds to a motion going straight up with no yaw, pitch or roll. Thus the calculations using the maximum principle agrees with the common sense of how to optimize the motion of going straight up. The switching functions do not change sign and the controls are all equal and at their maximum value, see Figure 5.7.

The next regular extremal is also a vertical motion but going down and with a yaw. The initial value for $p(0)$ is given by $(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1)$. This regular extremal is represented in Figure 5.9. For that motion the pair of rotors are in opposite ends of the domain of control. A consequence is the yaw motion that can be observed from Figure 5.11.

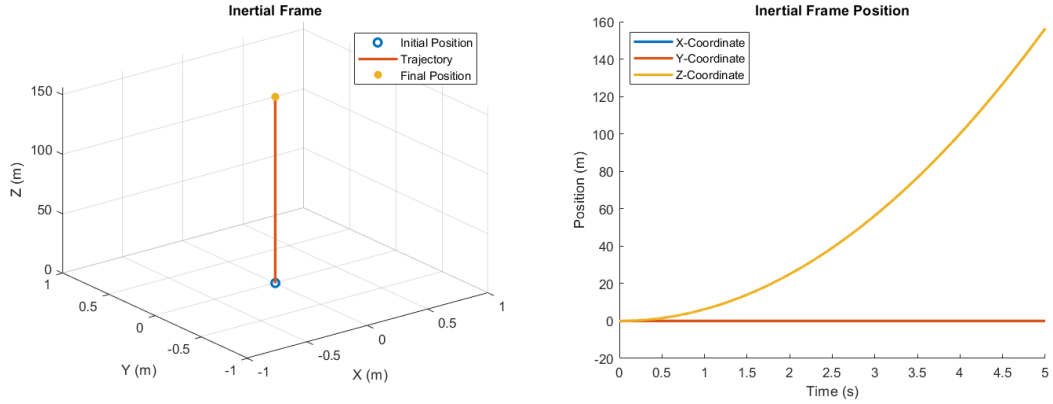


Fig. 5.6: Motion in the inertial frame when all components of p_0 are 0 except the ninth component which is 1000. Starts at the origin in a hovering equilibrium with $\psi = 0$ and goes for 5 seconds towards $\eta = (0, 0, 156.4251, 0, 0, 0)^t$, with a final velocity in the body frame of $(0, 0, 62.5369)^t$.

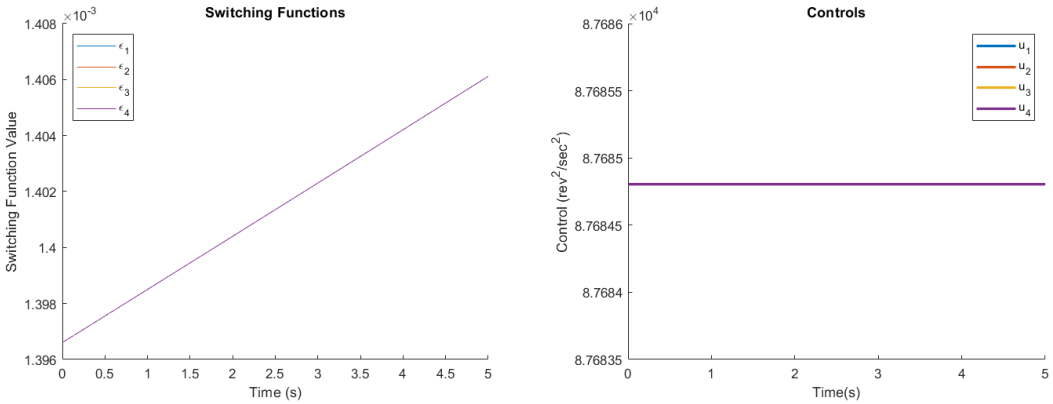


Fig. 5.7: Corresponding switching functions values and control for the regular extremal shown in Figure 5.6.

Figures 5.12 through 5.14 show a regular extremal corresponding to an initial value $p(0) = (0.6882, -0.3061, -4.8810, -1.6288, -3.3782, 2.9428, -1.8878, 0.2853, -3.3435, 1.0198, -2.3703, 1.5408)^t$. It can be observed that this corresponds to a motion which goes down in a spiral manner in the inertial frame, with the velocity in the body frame being sinusoidal with increasing magnitude and decreasing period. Each switching function has a single zero during the time between 0.5 seconds and 0.8 seconds, which changes the control.

Based on numerical simulations of regular extremals, there is evidence that chattering is to be expected for optimal trajectories. Chattering, or also known as the Fuller Phenomenon has been related to mechanical systems and studied in details, see [3] for a survey on the subject, and [14] for some examples of applications of chattering. For mechanical systems, this phenomenon typically arises when bang arcs connect with a singular arcs. As the switching function must transit to an

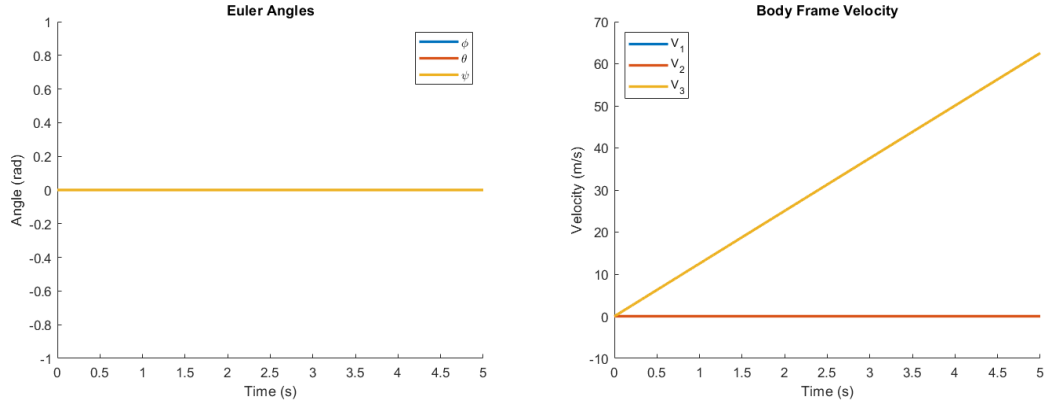


Fig. 5.8: Corresponding Euler angles and body frame velocities for regular extremal in Figure 5.6.

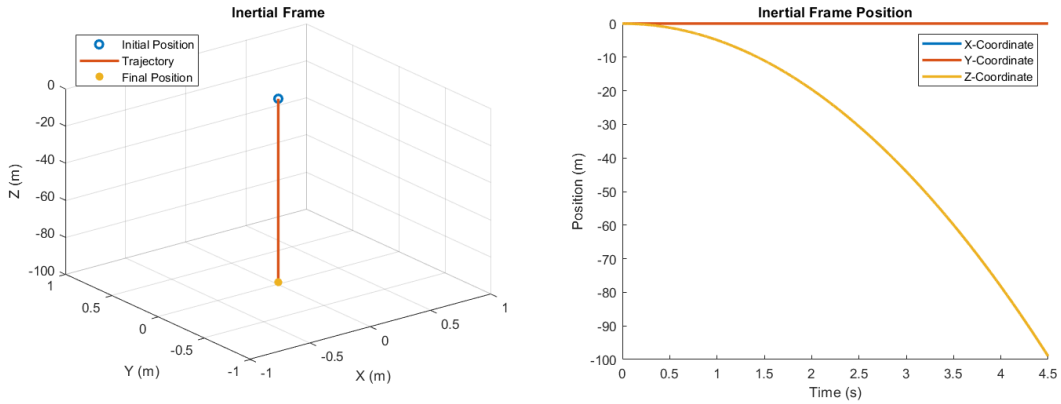


Fig. 5.9: Motion in the inertial frame when all components of p_0 are 0 except the twelfth component which is 1. Starts at the origin in a hovering equilibrium with $\psi = 0$ and goes for 4.5 seconds towards $\eta = (0, 0, -99.0584, 0, 0, 13.9046)^t$.

identically zero function it goes through an accumulation of zeroes (to guarantee that derivatives of all order to cancel at the transition point). While optimal trajectories might display this type of behavior, it is in practice not desired especially for mechanical systems. It also creates numerical difficulties which are not addressed here but would be an interesting direction for continuing this research.

Figure 5.15 shows the switching functions and the corresponding controls for an extremal starting with all components of p_0 set at 0 except $p_{10}(0) = 0.01$. Clearly there seems to be an accumulation point at the origin for control u_4. It could actually be the case that this control is singular from a little above 3 second before switching to a bang-bang structure, but our algorithm is not designed to handle the singular arcs, the calculations therefore breakdown in terms of integrating the system. The corresponding trajectory is shown on Figures 5.16 and 5.17 but we can observe that the variables "explode".

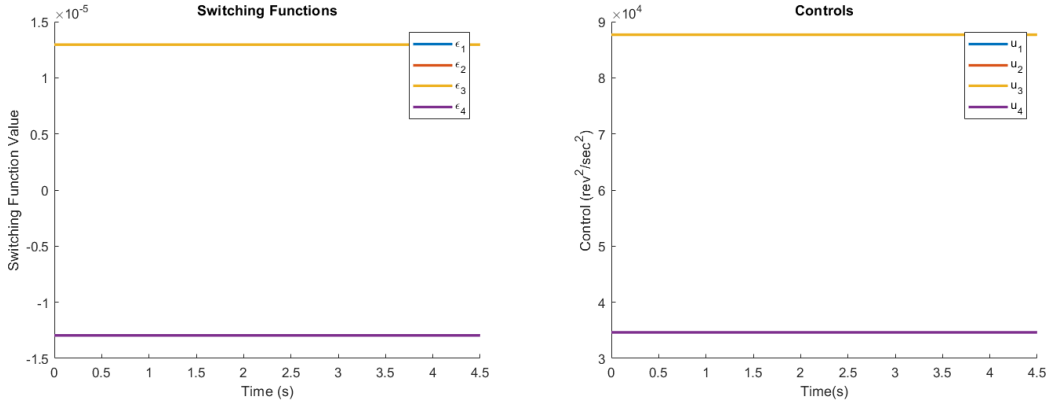


Fig. 5.10: Corresponding switching functions values and control for the regular extremal in Figure 5.9.

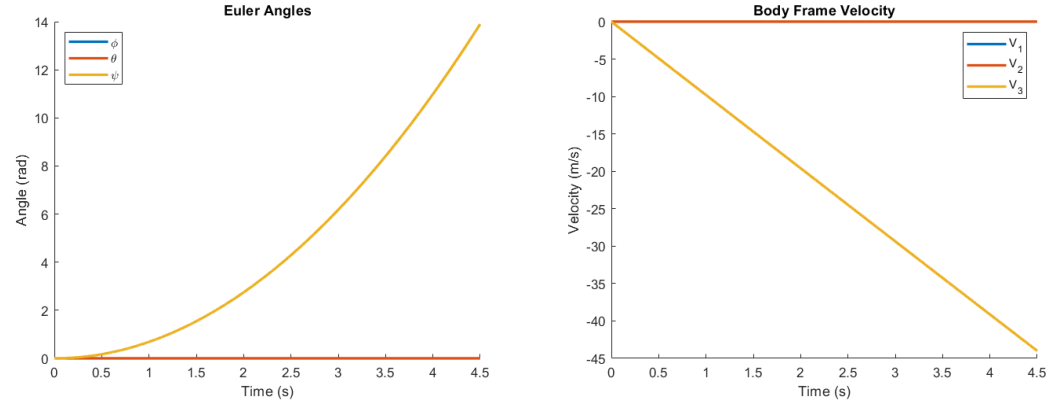


Fig. 5.11: Corresponding Euler Angles and Body Frame Velocities for motion in Figure 5.9.

The fact that we could detect chattering suggests that optimal trajectories contain singular arcs but more work is needed to show evidence of their optimality.

5.4.8 Singular Extremals

To determine the singular control we need to differentiate at least twice the switching functions. Provided the derivatives of switching functions given in Lemma 3 and 4 we then need to look at the Lie brackets for (5.75) and (5.76) of order two which include F_i.

For the Lie brackets of order one since the control vector fields F_i, $i = 1, \cdots, 4$ (5.76) are constant vector fields we have:

$$[F_i, F_j](q) = 0 \text{ for } i, j = 1 \dots 4. \tag{5.112}$$

Moreover, given any vector field X defined on our configuration space:

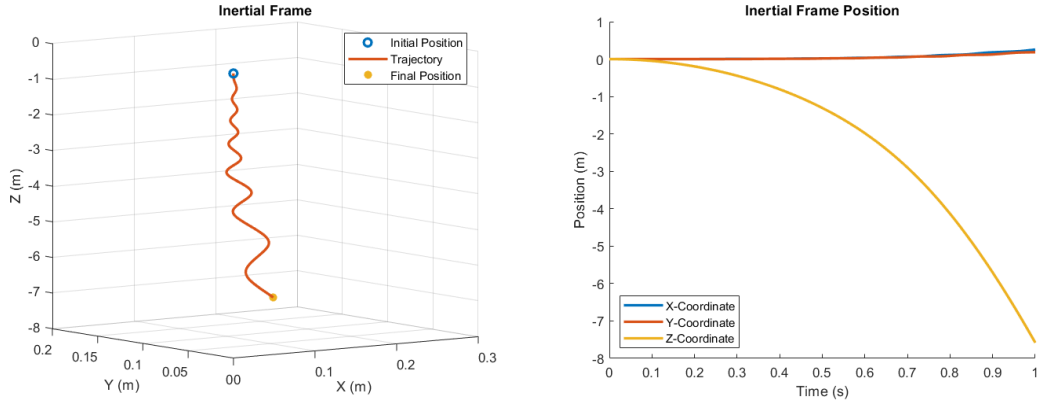


Fig. 5.12: Motion in the inertial frame when $p_0 = (0.6882, -0.3061, -4.8810, -1.6288, -3.3782, 2.9428, -1.8878, 0.2853, -3.3435, 1.0198, -2.3703, 1.5408)^t$. Starts at the origin in a hovering equilibrium with $\psi = 0$ and goes for 1 second towards $\eta = (0, 0, -99.0584, 0, 0, 13.9046)^t$.

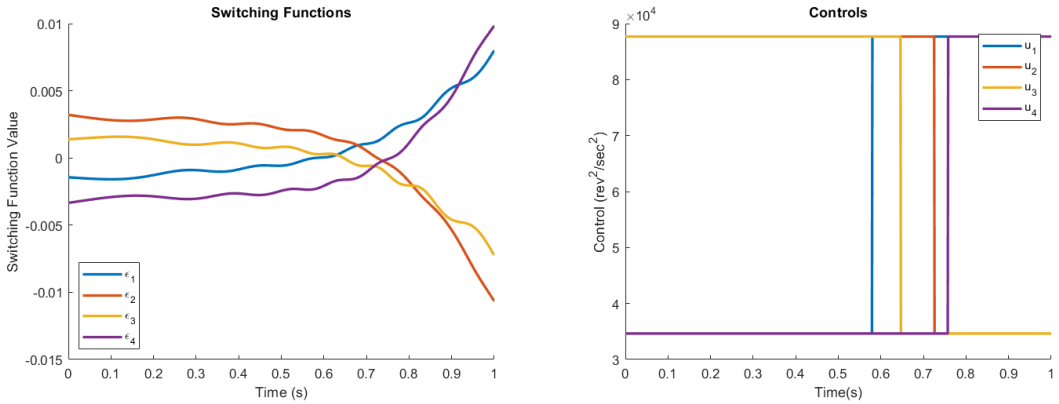


Fig. 5.13: Corresponding switching function value and control for motion in Figure 5.12.

$$[X, F_i](q) = -\frac{\partial X}{\partial q} F_i(q) \text{ for } i = 1 \dots 4. \tag{5.113}$$

Furthermore, since the first eight components of the control vector fields are $0s$, see Equation (5.76), we are only interested in the partial derivatives of X with respect to the variables $(v_3, \Omega_1, \Omega_2, \Omega_3)$ and thus we have:

$$[X, F_i](q) = -\frac{\partial X}{\partial v_3} F_i^9(q) - \sum_{j=1}^{3} \frac{\partial X}{\partial \Omega_j} F_i^{9+j}(q) \text{ for } i = 1 \dots 4, \tag{5.114}$$

where F_i^k denotes the k component of the vector F_i which actually does not depend on q but is constant.

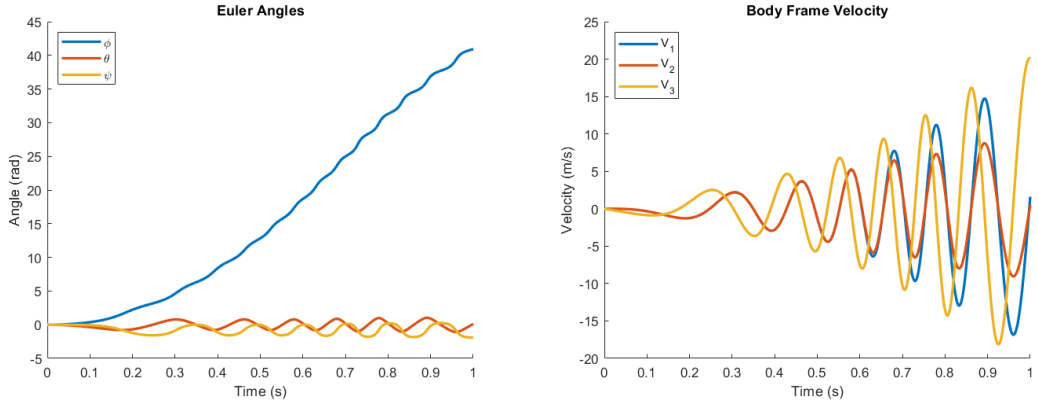


Fig. 5.14: Corresponding Euler Angles and Body Frame Velocities for motion in Figure 5.12.

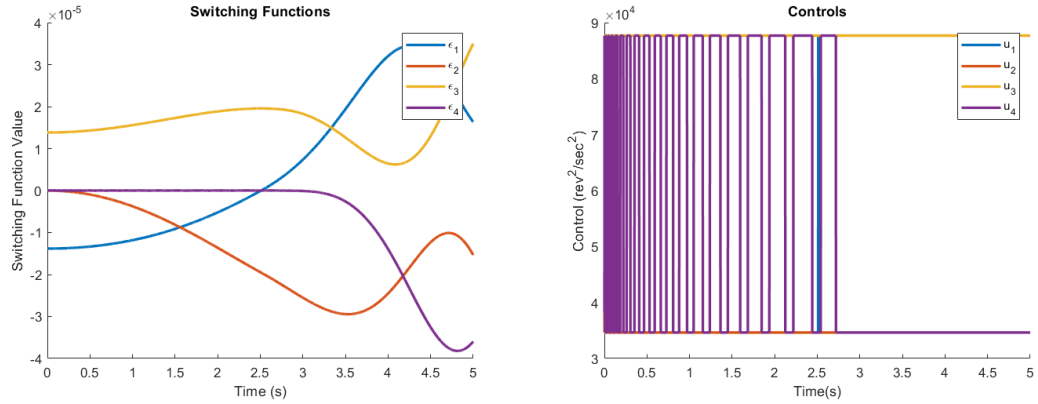


Fig. 5.15: Corresponding Switching function value and control for motion in Figure 5.16.

The following lemma provides the structure of the Lie bracket of the drift vector field F_0 with the control vectors field F_i, $i = 1, \cdots 4$.

Lemma 5. *Given F_0 and $F_i's$ from Equations (5.75) and (5.76), we have:*

$$[F_0, F_i] = -\frac{\alpha}{m}\frac{\partial F_0}{\partial v_3} - \sum_{j=1}^{3} \delta_i^j \frac{\partial F_0}{\partial \Omega_j}, \tag{5.115}$$

where:

$$\delta_1^1 = -\delta_3^1 = -\frac{\beta}{J_1} \tag{5.116}$$

$$\delta_2^2 = -\delta_4^2 = -\frac{\beta}{J_2} \tag{5.117}$$

$$\delta_1^2 = \delta_2^1 = \delta_3^2 = \delta_4^1 = 0 \tag{5.118}$$

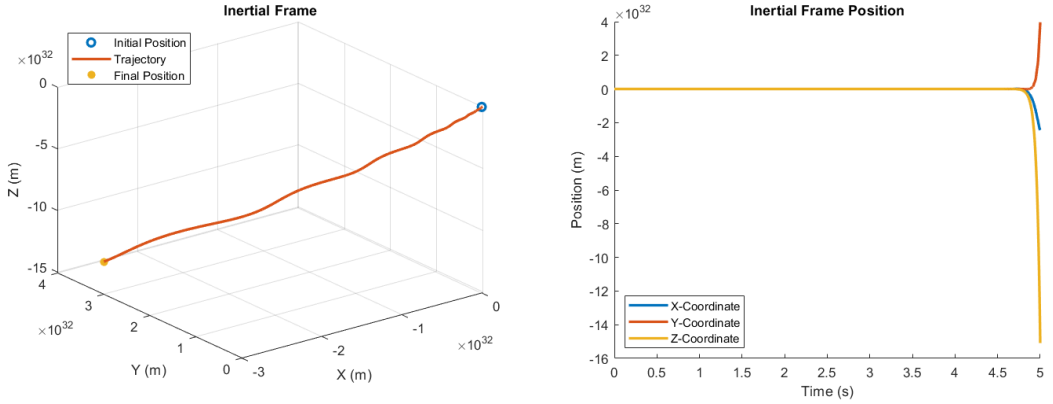


Fig. 5.16: Motion in the inertial frame when all components of p_0 are 0 except the tenth component which is 0.01. Starts at the origin in a hovering equilibrium with $\psi = 0$ and goes for 5 seconds towards $\eta = (-2.4383 \times 10^{32}, 3.9652 \times 10^{32}, -1.5087 \times 10^{33}, 60.6605, 0.7820, -152.6168)^t$.

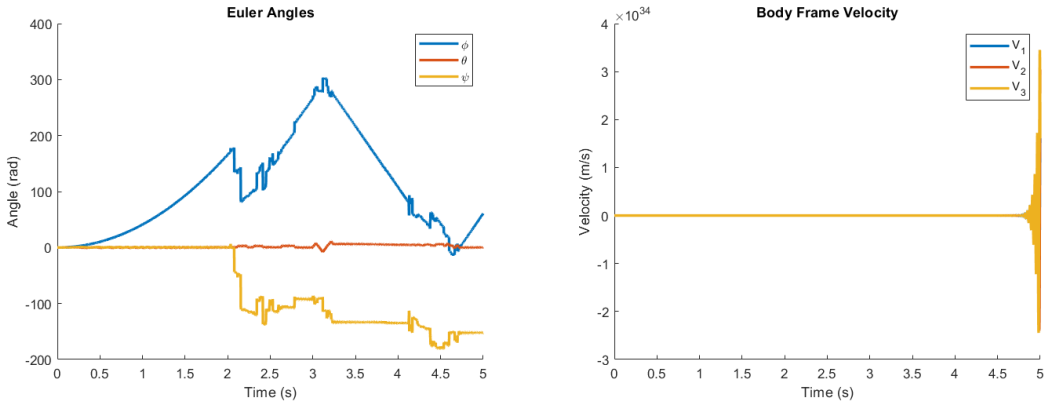


Fig. 5.17: Corresponding Euler Angles and Body Frame Velocities for motion in Figure 5.16.

$$\delta_1^3 = -\delta_2^3 = \delta_3^3 = -\delta_4^3 = \frac{\gamma}{J_3}. \tag{5.119}$$

In addition the following relations hold

$$[F_0, F_1] = [F_0, F_3] - 2\delta_1^1 \tfrac{\partial F_0}{\partial \Omega_1}, \qquad [F_0, F_2] = [F_0, F_4] - 2\delta_2^2 \tfrac{\partial F_0}{\partial \Omega_2}. \tag{5.120}$$

Proof. Introducing the δ_i^j as in equations (5.116) to (5.119), the control vector fields (5.76) take the form,

$$F_1 = \frac{\alpha}{m}\vec{e}_9 + \delta_1^1\vec{e}_{10} + \delta_1^3\vec{e}_{12} \tag{5.121}$$

$$F_2 = \frac{\alpha}{m}\vec{e}_9 + \delta_2^2\vec{e}_{11} + \delta_2^3\vec{e}_{12} \tag{5.122}$$

$$F_3 = \frac{\alpha}{m}\vec{e}_9 + \delta_3^1\vec{e}_{10} + \delta_3^3\vec{e}_{12} \tag{5.123}$$

$$F_4 = \frac{\alpha}{m}\vec{e}_9 + \delta_4^2\vec{e}_{11} + \delta_4^3\vec{e}_{12} \tag{5.124}$$

Substituting F_0 for X in (5.114) and replacing F_i^k with what the k component of F_i is we get:

$$[F_0, F_1] = -\frac{\alpha}{m}\frac{\partial F_0}{\partial v_3} - \delta_1^1\frac{\partial F_0}{\partial \Omega_1} - \delta_1^3\frac{\partial F_0}{\partial \Omega_3}, \tag{5.125}$$

$$[F_0, F_2] = -\frac{\alpha}{m}\frac{\partial F_0}{\partial v_3} - \delta_2^2\frac{\partial F_0}{\partial \Omega_2} - \delta_2^3\frac{\partial F_0}{\partial \Omega_3}, \tag{5.126}$$

$$[F_0, F_3] = -\frac{\alpha}{m}\frac{\partial F_0}{\partial v_3} - \delta_3^1\frac{\partial F_0}{\partial \Omega_1} - \delta_3^3\frac{\partial F_0}{\partial \Omega_3}, \tag{5.127}$$

$$[F_0, F_4] = -\frac{\alpha}{m}\frac{\partial F_0}{\partial v_3} - \delta_4^2\frac{\partial F_0}{\partial \Omega_2} - \delta_4^3\frac{\partial F_0}{\partial \Omega_3}. \tag{5.128}$$

The equations in (5.120) follows from subtracting (5.125) from (5.127) and (5.126) from (5.128) to get the last term in the two equations and then equating them using these terms.

Since yet again we have that (5.76) are constant vector fields, using (5.113) and Lemma 5 we obtain:

$$[F_j, [F_0, F_i]](q) = -\frac{\partial}{\partial q}\Big(\frac{\alpha}{m}\frac{\partial F_0}{\partial v_3} + \sum_{k=1}^{3}\delta_i^k\frac{\partial F_0}{\partial \Omega_k}\Big)F_j. \tag{5.129}$$

Lemma 6. *The Lie brackets of order two are:*

$$ad_{F_0}^2F_i(q) = -\frac{\partial}{\partial q}\Big(\frac{\alpha}{m}\frac{\partial F_0}{\partial v_3} + \sum_{k=1}^{3}\delta_i^k\frac{\partial F_0}{\partial \Omega_k}\Big)F_0 + \frac{\partial F_0}{\partial q}\Big(\frac{\alpha}{m}\frac{\partial F_0}{\partial v_3} + \sum_{k=1}^{3}\delta_i^k\frac{\partial F_0}{\partial \Omega_k}\Big), \tag{5.130}$$

and:

$$[F_j, [F_0, F_i]] = -\frac{\alpha}{m}\sum_{k=1}^{2}(\delta_i^k + \delta_j^k)\frac{\partial^2 F_0}{\partial v_3 \partial \Omega_k} - \sum_{k=1}^{3}\sum_{l=1}^{3}\delta_j^k\delta_i^l\frac{\partial^2 F_0}{\partial \Omega_k \partial \Omega_l}. \tag{5.131}$$

For (5.131) when $k = l$ the term in the double summation is zero. Moreover if δ_1^2, δ_2^1, δ_3^2 or δ_4^1 are in the double summation then that term is zero. In addition, if $k = 1$ or 2 while i and j are 2 and 4 or 1 and 3 then the term in the first summation is zero.

Proof. Expression (5.130) follows from direct calculation. Then we know from (5.75) that F_0 has no terms with a quadratic variable which means that the diagonal of the $\frac{\partial^2 F_0}{\partial q^2}$ in (5.129) would be zero. Also since F_0 does not contain any variations on the term $v_3\Omega_3$ in any of it's rows then:

$$\frac{\partial^2 F_0}{\partial v_3 \partial \Omega_3} = \frac{\partial^2 F_0}{\partial \Omega_3 \partial v_3} = 0. \tag{5.132}$$

Using similar symbology as (5.114) then we get by expanding (5.129):

$$[F_j, [F_0, F_i]](q) = -F_j^9\sum_{k=1}^{3}\delta_i^k\frac{\partial^2 F_0}{\partial v_3 \partial \Omega_k} - \frac{\alpha}{m}\sum_{k=1}^{3}F_j^{9+k}\frac{\partial^2 F_0}{\partial \Omega_k \partial v_3} \tag{5.133}$$

$$-\sum_{k=1}^{3}\sum_{l=1}^{3} F_j^{9+k}\delta_i^l \frac{\partial^2 F_0}{\partial\Omega_k\partial\Omega_l}, \tag{5.134}$$

with the double summation equaling zero if $k = l$. Since all the operations in F_0 have continuous second partial derivatives on $\mathbb{R}$ then order does not matter for the partial derivatives. Thus by just substituting what the corresponding component of F_j is and combining like terms in the first two summations we get (5.131). The conditions for the terms being zero then follows directly from (5.116) thru (5.119).

We note that if we hadn't neglected angular drag then the terms where $k = l$ in the second summation would not necessarily be zero and if we hadn't made the drag linear then (5.131) would have the added term $-\frac{\alpha^2}{m^2}\frac{\partial^2 F_0}{\partial v_3^2}$.

For the sake of simplicity we will refer to $\frac{\alpha\beta}{m}$ as κ and $\frac{\beta\gamma}{J_1J_2J_3}$ as λ which gives us:

$$[F_2,[F_0,F_1]] = -\frac{\kappa}{J_2}e_7 + \frac{\kappa}{J_1}e_8 + \lambda(J_2-J_3)e_{10} + \lambda(J_1-J_3)e_{11} - \frac{\beta\lambda(J_1-J_2)}{\gamma}e_{12} \tag{5.135}$$

$$[F_4,[F_0,F_1]] = \frac{\kappa}{J_2}e_7 + \frac{\kappa}{J_1}e_8 - \lambda(J_2-J_3)e_{10} + \lambda(J_1-J_3)e_{11} + \frac{\beta\lambda(J_1-J_2)}{\gamma}e_{12} \tag{5.136}$$

$$[F_2,[F_0,F_3]] = -\frac{\kappa}{J_2}e_7 - \frac{\kappa}{J_1}e_8 + \lambda(J_2-J_3)e_{10} - \lambda(J_1-J_3)e_{11} + \frac{\beta\lambda(J_1-J_2)}{\gamma}e_{12} \tag{5.137}$$

$$[F_4,[F_0,F_3]] = \frac{\kappa}{J_2}e_7 - \frac{\kappa}{J_1}e_8 - \lambda(J_2-J_3)e_{10} - \lambda(J_1-J_3)e_{11} - \frac{\beta\lambda(J_1-J_2)}{\gamma}e_{12} \tag{5.138}$$

Proposition 10.

$$[F_3,[F_0,F_1]] = [F_4,[F_0,F_2]] = 0 \tag{5.139}$$

$$[F_1,[F_0,F_1]] = -[F_3,[F_0,F_3]] = \frac{2\kappa}{J_1}e_8 - 2\lambda(J_1-J_3)e_{11} \tag{5.140}$$

$$[F_2,[F_0,F_2]] = -[F_4,[F_0,F_4]] = -\frac{2\kappa}{J_2}e_7 - 2\lambda(J_2-J_3)e_{10} \tag{5.141}$$

We also have the symmetric relations,

Proposition 11. *For $i, j = 1, \cdots, 4$ we have:*

$$[F_i,[F_0,F_j]] = [F_j,[F_0,F_i]], \tag{5.142}$$

as well as:

$$[F_4,[F_0,F_1]] + [F_4,[F_0,F_3]] + [F_2,[F_0,F_1]] + [F_2,[F_0,F_3]] = 0, \tag{5.143}$$

$$[F_4+F_2,[F_0,F_3+F_1]] = 0, \tag{5.144}$$

$$[F_3+F_1,[F_0,F_4+F_2]] = 0. \tag{5.145}$$

Proof. For (5.142) we can see in (5.131) that every operation and $\frac{\partial F_0}{\partial\Omega_k\partial\Omega_l}$, are all commutative, while (5.143) comes from direct calculations. Equation (5.144) comes from the fact that Lie Brackets are bilinear functions, and finally (5.145) comes from applying (5.142) to (5.143) before combining the Lie Brackets.

Proposition 7 provides conditions that have to be satisfied along singular arcs. Since the actuation of a quadcopter is based on four rotors then we have four controls, which results in complicated concatenations of bang and singular controls for extremals. Along an extremal, you can have only one control be singular or several of them. The more controls that are singular then the more conditions that exist, from Proposition 7.

To compute totally singular extremals we have that the four switching functions must be zeros and the conditions of Proposition 7 must be satisfied by all controls, $i = 1, \cdots 4$. Note that we can apply Proposition 7 because Equations (5.112) hold. Condition 5.93 can be stated explicitly as follows. Let us introduce $O_{ij} = \langle p, [F_j, [F_0, F_i]]\rangle$ and define by $\mathscr{O}$ the 4×4 matrix $\mathscr{O} = [O_{ij}]$. Note that by proposition 11, we have that $O_{ji} = O_{ij}$ which means that the matrix O is a symmetrix matrix. Condition 5.93 is now equivalent to:

$$\begin{bmatrix} O_{11} & O_{12} & O_{13} & O_{14} \\ O_{21} & O_{22} & O_{23} & O_{24} \\ O_{31} & O_{32} & O_{33} & O_{34} \\ O_{41} & O_{42} & O_{43} & O_{44} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} -\langle p, ad^2_{F_0} F_1(X)\rangle \\ -\langle p, ad^2_{F_0} F_2(X)\rangle \\ -\langle p, ad^2_{F_0} F_3(X)\rangle \\ -\langle p, ad^2_{F_0} F_4(X)\rangle \end{bmatrix} \tag{5.146}$$

We obtain that if the determinant of $\mathscr{O}$, provided the other conditions to be a singular control are satisfied, is non zero then the matrix is invertible and we can compute the singular controls.

Proposition 12. *A totally singular extremal, i.e. when the four controls are singular over the same non empty time interval, satisfies the following conditions:*

$$p_9 = p_{10} = p_{11} = p_{12} \equiv 0. \tag{5.147}$$

Moreover, the controls are a solution of (5.146) with the coefficients given by (5.150) and also satisfy:

$$\frac{p_8}{J_1}(u_1 - u_3) - \frac{p_7}{J_2}(u_2 - u_4) = -\sum_{i=1}^{4} \langle p, ad^2_{F_0} F_i(X))\rangle. \tag{5.148}$$

Proof. If the four controls are singular on a non empty interval we must have the four switching functions ϵ_i identically zero on that interval. Using Proposition 5.95 the adjoint vector must satisfy:

$$\begin{bmatrix} \frac{\alpha}{m} & -\frac{\beta}{J_1} & 0 & \frac{\gamma}{J_3} \\ \frac{\alpha}{m} & 0 & -\frac{\beta}{J_2} & -\frac{\gamma}{J_3} \\ \frac{\alpha}{m} & \frac{\beta}{J_1} & 0 & \frac{\gamma}{J_3} \\ \frac{\alpha}{m} & 0 & \frac{\beta}{J_2} & -\frac{\gamma}{J_3} \end{bmatrix} \begin{bmatrix} p_9 \\ p_{10} \\ p_{11} \\ p_{12} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{5.149}$$

Since this matrix is inevitable, the determinant is given by $-\frac{8\alpha\beta^2\gamma}{J_1 J_2 J_3 m}$, it implies that the last coordinates of the adjoint vector must be zeros along a complete singular extremal. The matrix $\mathscr{O}$ becomes:

$$\begin{bmatrix} \frac{2\kappa}{J_1} p_8 & -\frac{\kappa}{J_2} p_7 + \frac{\kappa}{J_1} p_8 & 0 & \frac{\kappa}{J_2} p_7 + \frac{\kappa}{J_1} p_8 \\ -\frac{\kappa}{J_2} p_7 + \frac{\kappa}{J_1} p_8 & -\frac{2\kappa}{J_2} p_7 & -\frac{\kappa}{J_2} p_7 - \frac{\kappa}{J_1} p_8 & 0 \\ 0 & -\frac{\kappa}{J_2} p_7 - \frac{\kappa}{J_1} p_8 & -\frac{2\kappa}{J_1} p_8 & \frac{\kappa}{J_2} p_7 - \frac{\kappa}{J_1} p_8 \\ \frac{\kappa}{J_2} p_7 + \frac{\kappa}{J_1} p_8 & 0 & \frac{\kappa}{J_2} p_7 - \frac{\kappa}{J_1} p_8 & \frac{2\kappa}{J_2} p_7 \end{bmatrix} \tag{5.150}$$

The determinant of this matrix is 0. This implies that there is a relation between the singular control along a totally singular extremal. Adding the four rows together, we find that the relation

between the controls can be represented by the following equation: $\frac{p_8}{J_1}(u_1 - u_3) - \frac{p_7}{J_2}(u_2 - u_4) = -\sum_{i=1}^{4}\langle p, ad_{F_0}^2 F_i(X))\rangle$.

Note that the adjoint vector also needs to satisfy Equations (5.92). Those provide 4 equations for 8 unknown (p_1 to p_8). The results above provide a nice algorithm to simulate numerically the totally singular extremals. It is harder to characterize non totally singular extremals. This comes from the fact that there are less conditions to be satisfied in those cases.

In [3] the author shows that chattering is closely related to the existence of singular extremals and their order. More specifically, concatenation between singular extremals of order two and bang-bang arcs leads to chattering. Indeed, there is a theorem (Kelley-Kopp-Moyer) that proves that the concatenation of a piecewise smooth nonsingular arc with a singular arc of even order is nonoptimal.

Definition 8. *The order of a singular control u_i is defined as the lowest integer n such that u_i appears explicitly in $\frac{d^{2n}}{dt^{2n}}\epsilon_i$.*

To characterize the order of singular extremals we need to understand the terms in front of the component of the singular control when we differentiate the corresponding switching function. Those are related to the Lie brackets $[F_i, [F_0, F_i]]$. Indeed, we have that for u_i to be a singular extremal of order 2 the following conditions need to be satisfied:

$$\langle p, F_i(X)\rangle = 0, \qquad \langle p, [F_0, F_i](X)\rangle = 0, \qquad \langle p, [F_i, [F_0, F_i]](X)\rangle = 0. \tag{5.151}$$

The first two conditions are related to the corresponding switching function and it's derivative to be zero and the last condition implies that the control cannot be retrieved from the second derivatives of the switching functions and we have to go to higher orders.

Proposition 13. *Assume u_i is singular. The terms in front of u_i in 5.91, $\langle p, [F_i, [F_0, F_i]](x)\rangle$, are given by:*

$$\frac{2\kappa}{J_1}p_8 - 2\lambda(J_1 - J_3)p_{11} \tag{5.152}$$

if $i = 1$ (with a - sign for $i = 3$) and

$$-\frac{2\kappa}{J_2}p_7 - 2\lambda(J_2 - J_3)p_{10} \tag{5.153}$$

if $i = 2$ (with a - sign for $i = 4$).

Proof. Use Proposition 10.

For our simulation seen in Figure 5.15, we conjecture that the phenomenon happening is the following. The component u_4 of the control is singular, which means that ϵ_4 (see Equation 5.98) must be zero:

$$\langle p, F_4(X)\rangle = \frac{\alpha}{m}p_9 + \frac{\beta}{J_2}p_{11} - \frac{\gamma}{J_3}p_{12} = 0. \tag{5.154}$$

The condition $\langle p, [F_4, [F_0, F_4]](X)\rangle = 0$ is given by:

$$\frac{2\kappa}{J_2}p_7 - 2\lambda(J_2 - J_3)p_{10} = 0. \tag{5.155}$$

Our conjecture is that the regular extremal shown in Figure 5.15 is a concatenation between a singular arc of order 2 and a bang arc for u_4 while the other components of the control are bang-bang.

5.5 Conclusions and Future Work

The core of this paper, beside deriving the equations of motion including a coordinate free formulation, is the application of the maximum principle to the time minimization problem for quadcopters. The time has been chosen as a criterion since for many scenarios, like rescue missions and short survey missions, a very rapid transit is usually the main consideration, it is however not the only criterion that could be considered. Another criterion of interest to minimize for quadcopters is the expanded energy during the mission. This is especially important since they rely on batteries and their flight time is still typically fairly short. In our case the energy could be first taken as the integral of the sum of the component of the controls since they represent the square of the angular velocities of the rotors: $\int \sum_{i=1}^{4} u_i(t)dt$. The maximization condition for $p_0 = -1$ (regular extremals) of the maximum principle would then provide $u_i = -\frac{1}{2}\langle p, F_i(x)\rangle$ and the saturation constraint that the control is admissible, i.e. $u(t) \in \mathscr{F}$. It would be interesting to simulate regular extremals corresponding to this situation. Singular extremals are intrinsic to the system and do not depend on the cost, they would therefore be the same. Their optimization status and role in the optimal synthesis might be different however depending on the cost.

For the time minimization problem addressed in this paper, an open and difficult question is to understand the structure of the optimal trajectories. Indeed, the optimal controls are formed by a concatenation of bang and singular arcs. The maximum principle does not however provide information about the number of switchings or how a transition between a singular arc and a bang arc happens. Note that pre-determined flights are typically made up of pure rotations and translations concatenated together. These are then broken up into a period where the quadcopter will change acceleration at a consistent rate and then hold that velocity steady. In these cases the number of switching is limited, and usually only occurs when the quadcopter transitions from one motion to another. Thus we would need to consider optimizing on a specific bang-bang structure with at most a fixed number of switchings.

References

1. A. Bloch, Nonholonomic Mechanics and Control, 2nd edition, Springer-Verlag, New York, 2015.
2. B. Bonnard and M. Chyba, The Role of Singular Trajectories in Control Theory, Springer, (2003).
3. V. Borisov, Fuller's Phenomenon: Review, *Journal of Mathematical Sciences*, Volume 100, No. 4 (2000).
4. H. Bouadi, M. Bouchoucha, and M. Tadjine, Sliding mode control based on backstepping approach for an UAV type-quadrotor, *International Journal of Mechanical and Mechatronics Engineering*, **1** (2007), 39–44.
5. H. Bouadi and M. Tadjine, Nonlinear observer design and sliding mode control of four rotors helicopter, *International Journal of Aerospace and Mechanical Engineering*, **1** (2007), 354–359.
6. R. Carney, M. Chyba, C. Gray, G. Wilkens, and C. Shanbrom, Multi-Agent Systems for Quadcoptors, *AIMS' Journals* Volume 14, Issue 1 (2022), 1–28. doi: 10.3934/jgm.2021005.
7. M. Chyba, T. Haberkorn, R. Smith, and G. Wilkens, A geometric analysis of trajectory design for underwater vehicles, *Discrete Contin. Dyn. Syst. Ser. B*, **11** (2009), 233–262.
8. D. Gandolfo, L. Salinas, A. Brandão and J. Toibero, Stable Path-Following Control for a Quadrotor Helicopter Considering Energy Consumption, *IEEE Transactions On Control Systems Technology*, VOL. 25, NO. 4, July 2017.
9. J. Keane and S. Carr, A Brief History of Early Unmanned Aircraft, *John Hopkins APL Technical Digest*, Volume 32, Number 3, (2013).

10. D. Liberzon, Calculus Of Variations And Optimal Control Theory: A Concise Introduction, Princeton University Press, New Jersey, 2012.
11. T. Luukkonen, Modelling and control of quadcopter, Independent research project, Aalto University in Espoo, Finland, (2011).
12. M. Mueller and T. D'Andrea, Stability and control of a quadrocopter despite the complete loss of one, two, or three propellers, *2014 IEEE international conference on robotics and automation (ICRA)*, (2014), 45–52.
13. Pontryagin VG et al., The Mathematical Theory of Optimal Processes, K. N. Tririgoff, Transl., L. W. Neustadt, Ed., Wiley, New York, 1962.
14. R. Robin, U. Boscain, M. Sigalotti, D. Sugny, Chattering Phenomenon in Quantum Optimal Control, *New Journal of Physics*, **24**, December 2022.
15. R. Smith, M. Chyba, G. Wilkens, and C. Catone, A geometrical approach to the motion planning problem for a submerged rigid body, *International Journal of Control*, **82** (2009), 1641–1656.
16. V. Stepanyan and K. Krishnakumar, Estimation, navigation and control of multi-rotor drones in an urban wind field, *AIAA Information Systems-AIAA Infotech @ Aerospace*, (2017).
17. Z. Zou, Trajectory tracking control design with command-filtered compensation for a quadrotor, *IET Control Theory & Applications*, **19** (2010), 2343–2355.

6

Hybrid Control, Morse Theory and Ivan Kupka.

Richard Montgomery[1] and Ricardo Sanfelice[2]

[1] Department of Mathematics, University of California Santa Cruz rmont@ucsc.edu
[2] Department of Electrical and Computer Engineering, University of California Santa Cruz ricardo@ucsc.edu

Summary. We begin with memories of Ivan Kupka. In the body of the paper we use Morse theory to construct a hybrid feedback law that robustly and globally asymptotically stabilizes the system to any desired point of any compact connected manifold. The method is a straightforward generalization of an example of performing this trick on the circle, found in a textbook by the second author. The logic variable part of "hybrid" is a single bit indicating whether or not to switch, with hysteresis, between two smooth vector fields on the manifold. One vector field is minus the gradient of a Morse function, to be constructed, whose global minimum is the desired point. The other vector field represents a steady breeze blowing by all the unstable equilibria of the gradient flow and pointing roughly parallel to their unstable manifolds. In order to motivate the use of hybrid control, we discuss how one might formulate the ideas of robustness, measurement, and measurement error to feedback systems on manifolds.

6.1 Ivan Kupka

6.1.1 Montgomery

Ivan Kupka and I became friends through mathematics. He remained close to my heart ever after our initial meetings.

We met through subRiemannian geometry and its interactions with control theory. Mike Enos[3], Ivan Kupka had been at a conference together. I had recently uncovered the phenomenon of topologically stable strictly abnormal geodesics in rank 2 subRiemannian geometries (see [14]). Enos explained my basic example to Ivan in the back row during a boring talk.

Ivan became intrigued and wrote several papers around the phenomenon and a survey of subRiemannian geometry. See [2], [11], and [1].

As a result of that introduction, Ivan and I had several visits. I particularly remember walking through Île Saint-Louis with Ivan in a downpour in early Spring. We ducked under the eaves of a cafe. He was grumpy about the prices and poshness of the island. He told me about growing up there when the central streets of the island were a slum. He grew up poor. I began to get a deep appreciation for the French education system whose notion of equality allowed a slum kid like Ivan to rise to the top. Later, Ivan took me on a tour of Versailles, not far from his home outside of Paris and afterwards we had a simple delicious dinner at his house with his wife.

[3] Enos was a retired gymnast who had switched into mathematical control theory, wanting to do "falling cat" type optimal control problems with the dream of designing new gymnastic moves.

As a young man, I had dropped out of society and lived in a tree house and made a living on rivers teaching people to kayak. Ivan had joined some version of the French merchant marine (the legends are various) and somehow ended up in Brazil where he reconnected with mathematics, and got his PhD under Peixoto. I felt like we were some type of alter egos - alternative selves. I loved his mathematical taste. He engaged in all areas of mathematics. I did not always understand him. His love and skill in genericity arguments and singularity theory as exemplified by the Kupka-Smale theorem (see [16] and references therein) combined in wonderful surprising ways with his deep appreciation and skill in hard down-to-earth explicit computations involving special functions. He had a particular love of elliptic functions which shined through in his work with Bonnard et al. I feel blessed for our friendship and the time we had together.

6.1.2 Sanfelice

I never met Ivan Kupka in person, however, I gained a deep appreciation for his work on observers, also known a state estimators, for the purpose of reconstructing the full state of a dynamical system from measurements of a (likely nonlinear and noninvertible) function on state space. I became aware of this work during my short stint at the Ecole de Mines de Paris in Fall 2008, working with Laurent Praly on high gain observers using adaptive gains.

Laurent introduced me to Ivan's book on observers coauthored with J-P. Gauthier [7]. This book presents, in a deep and concise manner, a general theory for analysis and design of observers. It gives a much detailed presentation of their general approach then their seminal 1994 SIAM Journal on Control and Optimization article "Observability and observers for nonlinear systems," which, currently, has more than 600 citations.

Being shortly after I finished my PhD (in 2007), I was really thirsty for knowledge on state estimation, as my PhD focused mostly on control theory for the solution of state feedback problems. Infused with Laurent's courage (and great espresso), I carefully read the formulation, results, and proofs in Ivan's book. His work is mathematically deep and rigorous, arguably, among the most impactful ones on the topic. The generality of the mathematical development is also unique – a particular feature of it is that, unlike much of the work in the literature, his results do not assume that (maximal) solutions exist for all time. Many of our recent articles on observers for dynamical systems propose solutions that are inspired from the constructions in his book. It is evident that his work has made a long lasting impact on the field.

6.1.3 Acknowledgement

We thank AFOSR for the financial support through Grant no. FA9550-23-1-0313 which made our collaboration possible. In addition, research by R. G. Sanfelice was partially supported by NSF Grants no. CNS-2039054 and CNS-2111688, by AFOSR Grants nos. FA9550-19-1-0169, FA9550-20-1-0238, and FA9550-23-1-0145, by AFRL Grant nos. FA8651-22-1-0017 and FA8651-23-1-0004, by ARO Grant no. W911NF-20-1-0253, and by DoD Grant no. W911NF-23-1-0158. The authors would like to thank Akhil Datla and Piyush Jirwankar for helping digitize the drawings included in the figures.

6.2 Introduction and Setup

Many control problems are difficult to solve due to topological obstructions intrinsic to the system being controlled. Such obstructions emerge in most autonomous vehicles problems. We focus here on the problem of stabilizing a system on a manifold to a single fixed point using feedback. If the point is stable for a vector field then its basin of attraction is contractible. The flow itself yields the contraction to the the stable equilibrium point! But compact boundaryless manifolds are not contractible. It follows that finding a global Lipshitz feedback law for a smooth system on such a manifold is impossible. See [3] and [12] for more concerning topological obstructions to Lipshitz feedback stabilization.

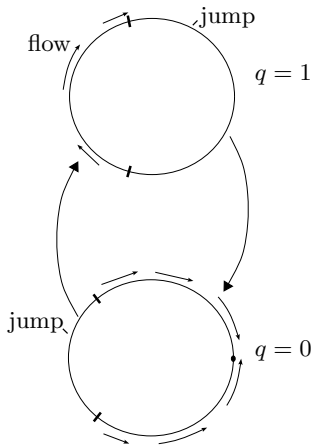


Fig. 6.1: Turning a gradient flow on the circle into a hybrid system with a single global attractor. The main trick is that the flow set for $q = 1$, the subset of the circle where you see the arrows, contains the jump set for $q = 0$.

Consider the problem of achieving *robust* global asymptotic stability of a desired point for the attitude of a planar rigid body. The goal is to render the desired point stable – trajectories starting nearby the point stay nearby – and globally attractive – every trajectory limits to the desired point as time approaches positive infinity – and, perhaps most importantly, to achieve these goals with robustness to perturbations such as noise in the measurements telling us our current attitude. The state space of the planar rigid body is the group $SO(2)$ of rotations of the plane, a group which is diffeomorphic to the circle $\S^1$ in the standard way:

$$R(\theta) = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \in SO(2)$$

being parameterized by the single angle θ. The circle is not contractible so we cannot design a smooth feedback system driving us to our goal, the identity, which corresponds to $\theta = 0$. Nevertheless, let us try. Introduce the control system

$$\dot{\theta} = u. \tag{6.1}$$

The feedback law

$$u = -\sin(\theta)$$

yields the negative gradient flow $\dot{\theta} = -\sin(\theta)$ for the function $\phi(\theta) = -\cos(\theta)$. It has the origin $\theta = 0$ as stable equilibrium. The basin of attraction of $\theta = 0$ is all of the circle minus the single point $\theta = \pi$ antipodal to $\theta = 0$. The point $\theta = \pi$ is an unstable equilibrium so our feedback law leaves it fixed where it is. We have not achieved global asymptotic stability. Almost - we missed by a hair: one point, $\theta = \pi$ just sits there forever, all the others limit to $\theta = 0$. We have failed to achieve global asymptotic stability of $\theta = 0$, in line with the basic fact from topology that the circle is not contractible. See also [4].

We can regain global asymptotic stability by using a discontinuous feedback control law. Any feedback law which interpolates in a convex manner between $u = -\text{sgn}(\pi - \theta)$ near $\theta = \pi$ and $u = -\sin(\theta)$ near $\theta = 0$ will do the trick. (Here $\text{sgn}(x)$ is the sign function, so that $\text{sgn}(x) = +1, x > 0$ and $\text{sgn}(x) = -1, x \leq 0$.)

However, introducing this discontinuity to our feedback law destroys robustness to measurement error. Suppose that m represents measurement error in the angle θ. Near $\theta = \pi$, the actual recieved feedack by the system would then be $u = -\text{sgn}(\pi - \theta + m)$. An arbitrarily small oscillatory measurement noise m can render the previously unstable equilbrium point $\theta = \pi$ into a stable equilibrium! [4]

The notion of robustness and measurement error are central to this paper. Hermes, in [9] brought the importance of measurement error its potentially devastating effects when feedback laws are discontinuous, and its beautiful connections to the Fillipov lemma to the attention of the control community. In Section 6.3.3 below we touch on his paper and define robustness to measurement noise so as to make sense on manifolds.

We can achieve global robust asymptotic stability by moving into the world of hybrid systems where we mix analog and digital. See [5] and [17]. Introduce a logic variable, or simply, a single bit $q \in \{0, 1\}$ which we carry around with us and monitor as we travel about the circle. The bit acts as a state-dependent switch to between two vector fields, say[5] $-\mathrm{d}\theta$ for $q = 1$ and $-\sin(\theta)\mathrm{d}\theta$ for $q = 0$, switching depending both on where we are on the circle and what the current state of the bit is. See figure 6.1. This is a basic example in the subject of hybrid feedback controllers. See p. 21, Section 1.2.1 of [17]

The point of this note is to show how we can use Morse theory to generalize the circle example so as to work on any compact manifold M. We "hybridize" M in the same way as we did the circle, by introducing a single bit $q \in \{0, 1\}$. The hybrid feedback law allows us to carry on with two interpenetrating vector fields which we can switch between depending on where we are and the value of our bit and in this way achieve a global robust asymptotic feedback stablizer on M. See theorem 9.27 at the end of the next-to-last section of this article.

6.2.1 Setup, Goal, and Strategy

Let M be a compact connected manifold and $m_0 \in M$ be our target. Our goal is to design a control system having a robust global feedback law with m_0 as its global attractor. As described above, for

[4] The basic phenomenon of stabilizing an unstable fixed point by imposing small amplitude high frequency oscillations earned Paul the Nobel prize in 1989 for the Paul Trap. See [10]. R. M. is grateful to Mark Levi for pointing out this connection.

[5] We use the standard notation of differential topology. The vector field $f(\theta)\mathrm{d}\theta$ implements the differential equation $\dot{\theta} = f(\theta)$.

topological reasons this is impossible within the standard framework of smooth feedback systems on M. We can however, using a bit of Morse theory, make m_0 into an "almost global attractor" for the gradient vector field of a function ϕ to be designed below:

$$\dot{z} = -\nabla\phi(z) \qquad z \in M. \tag{6.2}$$

When we say "almost global attractor" we mean that the basin of attraction for m_0 is an open dense subset of M.

We take $\phi : M \to \mathbb{R}$ to be a Morse function whose only local minimum is m_0. Consequently m_0 is the global minimizer of ϕ. [6] The hybrid strategy, following the circle example, is to understand where and how the gradient flow gets hung up and misses limiting to m_0. We then use another vector field Y – called a "breeze" below– to nudge the system away from these bad sticking points. The sticking points are exactly the other critical points of ϕ. Finally, we use the idea of hybrid feedback to switch back and forth between these two vector fields at judicious locations of the manifold with the help of an auxiliary bit $q \in \{0, 1\}$ which allows the introduction of memory in the feedback control algorithm.

6.2.2 Morse theory

We recall the relevant definitions and basic properties around Morse functions. A critical point of a smooth function $\phi : M \to \mathbb{R}$ is a point p where the differential $d\phi(p) = \sum \mathrm{d}\phi x^i|_p dx^i$ of the function ϕ vanishes. Here, the $x^i, i = 1, \ldots, n$, are coordinates near p and n is the dimension of M. At a critical point p, we can form the Hessian of ϕ:

$$\mathrm{Hess}(\phi) = \sum \frac{\partial^2 \phi}{\partial x_i \partial x_j} dx^i dx^j$$

which is understood as a bilinear symmetric form on the tangent space. The Hessian is independent of coordinates, but, unlike the Euclidean space setting, the Hessian is undefined if p is not a critical point. (The associated quadratic form obtained by using the formula at a non-critical point is coordinate dependent, its value changing as we change coordinates.)

Definition 1. *A critical point p of a function ϕ is called non-degenerate if the Hessian of ϕ is non-degenerate (i.e. the matrix of the Hessian is invertible) at p.*

Definition 2. *A smooth function is called a Morse function if all its critical points are non-degenerate.*

Lemma 1 (Morse lemma). *If p is a non-degenerate critical point of the smooth function ϕ on the n-dimensional manifold M then there exists a smooth coordinate system $x_1, \ldots, x_c, y_1, \ldots y_k$ on M centered at p such that, in these coordinates,*

$$\phi(x_1, \ldots x_c, y_1, \ldots, y_k) = \phi(p) + \sum_{a=1}^{c} x_a^2 - \sum_{b=1}^{k} y_b^2. \tag{6.3}$$

Here $k + c = n$ and k is the index of the critical point p.

[6] In order to define the gradient we need an auxiliary Riemannian metric on M. In tensor notation $-\nabla\phi(z) = \sum g^{ij}(z)\mathrm{d}\phi x^i \mathrm{d}x^j$ where the metric is $\sum g_{ij}(z) dx^i dx^j$.

We have employed two standard definitions:

Definition 3. *A coordinate system* [7] $x : M \dashrightarrow \mathbb{R}^n$ *is centered at* p *if* $x(p) = 0$.

Definition 4. *The index* k *of a nondegenerate critical point* p *for* ϕ *is the index of its Hessian: the largest possible dimension of a subspace* $S \subset T_pM$ *such that the restriction of* $\mathrm{Hess}(\phi)_p$ *to* S *is negative definite.*

Lemma 2 (Sard-Morse). *Every manifold admits Morse functions. Moreover, the space of Morse functions is open and dense within the space of all smooth functions on* M *endowed with the Whitney* C^2*-topology*

We refer the reader to Guillemin-Pollack [8], or Milnor [15] for proofs of the Morse lemma and the Sard-Morse theorem.

We need a special case of the "handle slide procedure" in order to guarantee only one local minimum for ϕ.

Proposition 1. *If* ϕ_0 *is a Morse function on the connected manifold* M *having* $m_0 \in M$ *as a local minimum, then we can deform* ϕ_0 *into another Morse function* ϕ_1 *which has* m_0 *as its only local minimum and is such that the critical values* c_i *of* ϕ_1 *are all distinct.*

This deformation is a homotopy, i.e., a path $\phi_t, 0 \leq t \leq 1$ of smooth functions all of which have m_0 as a local minimum. Except for a finite number of times t, each ϕ_t is Morse. The critical points of all the ϕ_t can be taken to be isolated. For a proof see [18, p. 143, Proposition 5.4.1] and the discussion in the paragraph preceding this proposition.

6.2.3 Hang Ups

For the same reasons that the gradient descent method works in Euclidean space, ϕ decreases strictly monotonically along any non-equilibrium trajectory to the gradient flow (6.2). It follows that almost all trajectories converge to m_0, it being the only minimum of ϕ. Some trajectories will get hung up on saddle points, that is to say, limit to an unstable equilibrium of the gradient flow. The equilibria of our gradient flow are exactly the critical points of ϕ. All trajectories which are not equilibria converge to one of these critical points.

By the Morse lemma the critical points are isolated, and hence finite in number. *We write* $N + 1$ *for this number, with* m_0 *counted amongst the critical points.* Consequently, there are N critical points, which are saddles or local maxima. We write the non-minimal critical points as $p_i, i = 1, \ldots, N$, and their critical values as $c_i = \phi(p_i)$.

Recall that the stable manifold of an equilibrium point p_i is the set of initial conditions z for which the trajectory of (6.2) through z converges to p_i in the limit as time approaches infinity. We denote this manifold by $W^+(p_i)$. It is a smooth embedded [8] manifold passing through p_i and whose dimension is the *co-index* $c = n - k$ of the critical point p_i.

[7] The broken arrow notation here is used to simply denote that the domain of x is an open subset of M and not all of M.

[8] Stable manifolds for general smooth vector fields are immersed, not embedded submanifolds. To wit: heteroclinic tangles and Hamiltonian chaos. However the stable manifolds of gradient flows are embedded submanifolds. See for example Corollary 7.4.1 in [Jost, Riemannian geometry and geometric analysis].

By assumption, the only critical point of index 0 is our target point m_0. See Proposition 1. Take the union of all the stable manifolds *except* for m_0's: That is, consider

$$\Omega = \bigcup_{i=1}^{N} W^+(p_i).$$

We have that $W^+(m_0) = M \setminus \Omega$: the basin of attraction of m_0 equals the complement of Ω. Away from Ω all trajectories of the gradient flow in (6.2) converge to m_0. Note that Ω has measure zero in M, being the finite union of submanifolds all of which have codimension at least 1. Consequently, the basin of attraction for m_0 is an open dense set of full measure – a non-linear counterpart of the complement of a finite collection of proper linear subspaces in a Euclidean space.

6.2.4 A Steady Breeze

One strategy for finding a hybrid feedback stabilizer to bring all points to m_0 would be to find a nonzero vector field Y transverse to each stratum $W^+(p_i)$ of Ω. Think of Y as a "strong wind, blowing past Ω." When we get close to Ω turn off the gradient flow and "let this wind blow." The flow of Y, being transverse to Ω, will push us back into the basin of attraction of m_0.

Finding such a Y is hard. It requires global knowledge of the stable manifolds $W^+(p_i)$ of our unstable critical points. *We can make due instead with the local knowledge provided by the Morse lemma* and, in essence, construct a collection of local winds or "breezes" Y_i , one for each unstable critical point p_i. The flow of Y_i will push all points sufficiently near p_i into a region collecting points p' such that $\phi(p') < \phi(p_i) - K$, where $K > 0$ is a constant. Once in this region we revert to the gradient vector field whose flow decreases ϕ, pushing points way from p_i and further decreasing ϕ either all the way down to its global minimum at m_0 or, with bad luck, near another unstable critical point p_j, one with $\phi(p_j) < \phi(p_i)$. Once near to this p_j, we can repeat the process, invoking the local breeze Y_j. Cycles between neighborhoods of different critical points cannot occur since we will insist that these neighborhoods do not intersect and between them ϕ strictly decreases since we use the gradient flow.

Consider a vector field $Y : M \to TM$ satisfying the property that

$$\mathrm{Hess}(\phi)_{p_i}(Y(p_i), Y(p_i)) = -2 \tag{6.4}$$

The existence of such a Y is straightforward. Since the Hessian has negative directions $y_1, \ldots, y_k$ at each p_i finding a vector $v_i \in T_{p_i}M$ with $\mathrm{Hess}(\phi)_{p_i}(v_i, v_i) = -2$ is easily done. Indeed, $v_i = \mathrm{d}y_1$ works, where (x, y) are Morse coordinates. Now all smooth manifolds M share a number of basic extension properties, one which is as follows. Given a vector $v \in T_pM$ at a point p, we can always find a vector field $Y : M \to TM$ with $Y(p) = v$. This extension property holds for any finite number $v_1, \ldots, v_N$ of vectors attached at distinct points of M. Consequently we have the existence of our Y.

Lemma 3 (steady breeze lemma). *[See Figure 6.2.] Associated to our vector field Y there are neighborhoods V_i of each non-minimal critical point p_i of our Morse function ϕ, and positive constants k_1, k_2 with the following property. Any trajectory for Y crossing into or starting in V_i leaves V_i within a time k_1, exiting at a point p of ∂V_i with $\phi(p) < \phi(p_i) - k_2$.*

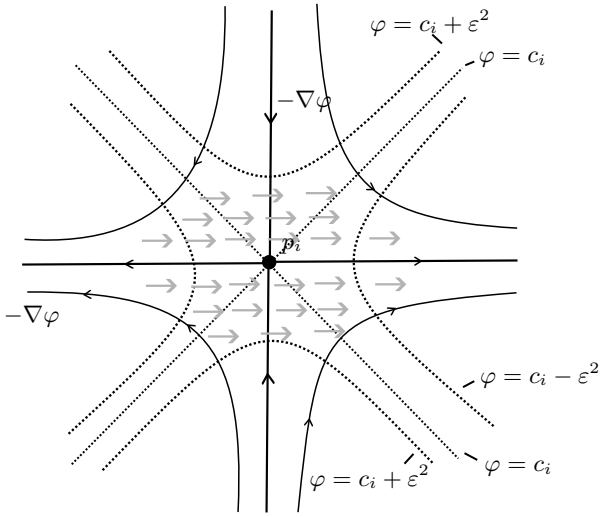


Fig. 6.2: The level sets of the Morse function ϕ near the critical point p_i are dashed. The vector field $\nabla\phi$ for gradient flow is indicated by its solid trajectories. The vector field Y for the breeze that blows past p_i is indicated by the short solid (brown) horizontal arrows.

Proof.

To begin with, take Y to be the constant vector fields $\mathrm{d}y_1$ in the coordinates of the Morse Lemma, (Lemma 1 above). The flow Ψ_t of Y in our Morse coordinates is the translational flow $(x, y) \mapsto (x, y + te_1) = \Psi_t(x, y)$, where e_1 denotes the vector in $\mathbb{R}^k$ (of $\mathbb{R}^n = \mathbb{R}^c \times \mathbb{R}^k$) whose only 1 corresponds to the choice of index $a = 1$, i.e., e_1 is the coordinate representation of $\mathrm{d}y_1$. Rewrite the Morse normal form as

$$\phi(x, y) - c_i = |x|^2 - |y|^2, \text{ where } \phi(p_i) = c_i$$

where the norms are the standard coordinate norms on the corresponding x and y coordinate spaces. Then

$$\phi(\Psi_t(x, y)) - c_i = (\phi(x, y) - c_i) - 2ty_1 - t^2.$$

View this as a quadratic expression in t. Imposing the conditions that $|x|^2 + |y|^2$ and hence y_1 are very small, the constant term $(\phi(x, y) - c_i)$ and the coefficient of the linear term $-2ty_1$ can be made arbitrarily small, so that the quadratic term eventually beats them. We view the conditions on $|x|^2 + |y|^2$ and y_1 as initial conditions for solving for the flow of Y. For V_i we can take a flow box of the form $|y_1| < A$, $|x|^2 + \sum_{a>1} y_a^2 < \delta$. The lemma follows immediately for this case.

There are at least two routes in to the general case. For one of these routes, use the symmetry group $SO(n - k, k)$ of the quadratic form $|x|^2 - |y|^2$ to "rotate" coordinates so that $Y(p_i) = \mathrm{d}y_1$. Then, argue that $Y(p)$ does not deviate far from $Y(p_i)$ as long as we stay in a small enough neighborhood of p_i. For the other route, use the 2nd order Taylor series with error estimates for the trajectory $\gamma_*(t)$ of Y passing through p_i to get that $\phi(\gamma_*(t)) < c_i - \frac{3}{4}t^2$ for all sufficiently small t, and then argue by uniform convergence of the flow $\Psi_t(p)$ of Y that the "far side" of the Taylor estimates, $k_2/2 < t < k_2$, hold for k_2 small and p close to p_i. We leave the details up to the reader.

QED

Remark. There are points p' arbitrarily close to p_i for which $\phi(p') > c_i$. Since $\Psi_0(p') = p'$ the inequality $\phi(\Psi_t(p')) - c_i < -\frac{1}{2}t^2$ must fail for t in an interval about 0 for these p'. For such p' we need to flow a non-zero amount of time before $\phi - c_i$ begins to become negative and then for a bit longer until our inequality holds.

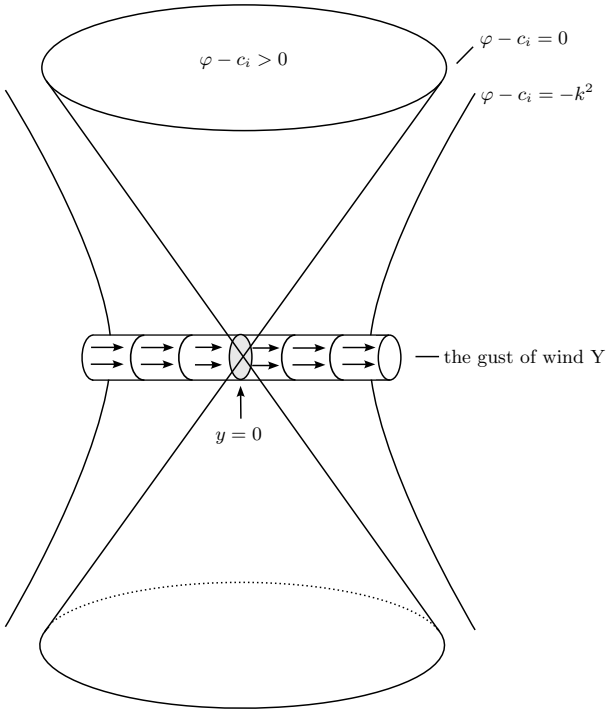


Fig. 6.3: A flowtube for the breeze flow of Y and its relation to the level sets of ϕ.

6.2.5 Topology

Morse theory relates critical points and their indices to the topology of the manifold. A basic topological invariant of a manifold is its "Betti numbers" $b_k = b_k(M)$, $k = 0, 1, \ldots$, which are popularly described as the "number of k-dimensional holes" in M. We have $b_j = 0, j > n$. Stated more carefully, for each choice of field F there are integers $b_k(M, F)$ that are equal to $\dim_F H_k(M, F)$, where $H_k(M, F)$ is the k-th homology group of M with coefficients in the field F. The Betti number we are talking about is the maximum over all fields of the $b_k(M, F)$.

Write m_k for the number of index k critical points of our Morse function ϕ. Then

$$m_k \geq b_k.$$

In particular

$$N + 1 \geq \sum b_k.$$

since $\sum m_k = N + 1$ where $N + 1$ is the number of critical points $m_0, p_1, \ldots, p_N$.

Example 1. Take $M = SO(3)$. It is well-known that $SO(3) = \mathbb{RP}^3$. Its Betti numbers are $b_0 = b_1 = b_2 = b_3 = 1$. If we work over the field F of rational or real numbers we will find that $b_1(M, F) = b_2(M, F) = 0$. However, over the field $F = \mathbb{Z}_2$ of two elements we have that $b_1(M, \mathbb{Z}_2) = b_2(M, \mathbb{Z}_2) = 1$. For all fields $b_0(M, F) = b_3(M, F) = 1$. From the Morse inequalities it follows that any Morse function on $SO(3)$ has at least 4 critical points.

6.3 Errors, Robustness, Hybridization

In this section, we propose a simple way to switch between Y and $-\nabla\phi$ so as to arrive to a feedback law that globally asymptotically stabilizes m_0. Then, we shoot down this law on grounds of robustness. Measurement errors can make discontinuous feedback laws induce undesired behavior, for example, it can "stabilize" the system to one of the unstable fixed points p_i of the gradient flow. Through the study of robustness (or lack of) to measurement noise of such a feedback law, a hybrid control feedback law is discovered. The intuition is that if we carry a bit $q \in \{0, 1\}$ in our pocket (it does not have to be a qubit!) as we travel around M, taking measurements of ϕ and $\|\nabla\phi\|$ as we travel, and switching bits appropriately, we can build a robustly globally stabilizing hybrid feedback law.

6.3.1 A Discontinuous Stabilizer

Let us return to the circle example in Section 6.2. Modify our feedback law near $\theta = \pi$ using a discontinuous control law having a discontinuity at $\theta = \pi$. One way to achieve this is to add to $u = \sin(\theta)$ any function of the form $g(\theta) := \beta(\theta)\text{sgn}(\theta - \pi)$, where $\text{sgn}(x)$ is the sign function $\text{sgn}(x) = -1$ if $x < 0, \text{sgn}(x) = +1$ if $x > 0$, and where $\beta(\theta)$ is a bump function supported in a small neighborhood of $\theta = \pi$ and such that $\beta(\pi) = 1$. Choose either -1 or $+1$ for the value of $\beta(\theta)\text{sgn}(\pi - \theta)$ at $\theta = \pi$. Thus, we are investigating the flow of the discontinuous vector field $\dot{\theta} = \sin(\theta) + g(\theta)$. Declare a solution to be an absolutely continuous curve $t \mapsto \theta(t)$ that satisfies $\dot{\theta} = g(\theta)$ almost everywhere. Then, for every initial condition $\theta_0 \in M$ passes a unique solution and this solution converges to m_0, which we recall is the point $\theta = 0$.

We can copy this example onto our manifold. Recall the neighborhoods V_i of the steady breeze lemma (Lemma 3). They can be taken to be balls or tubes with smooth boundaries, and so that $-\nabla\phi$ and Y are transverse to ∂V_i at all but a finite number of points. Set

$$V = \bigcup_{i=1}^{N} V_i \tag{6.5}$$

Define the discontinuous vector field

$$F(z) := \begin{cases} -\nabla\phi(z) & \text{if } z \notin V \\ Y(z) & \text{if } z \in V \end{cases}$$

Under this discontinuous vector field, every (maximally defined) trajectory of $\dot{z} = F(z)$ converges to m_0.

6.3.2 ...Gets Ruined in the Presence of Measurement Noise

Hermes [9] made a basic observation linking noise and uncertainty to the Fillipov lemma. As a consequence we can establish (or design) arbitrarily small noise or measurement uncertainties, applied near the discontionuities, which stabilize the system there!

Imagine we are working in a single Morse chart and that Y is straightened out so as to be the constant vector field $e_1 = \mathrm{d}y_1$. Suppose that we are near the discontinuity of F at ∂V as defined above. Rewrite our system as a control system

$$\dot{z} = -u\nabla\phi(z) + (1-u)Y(z) \tag{6.6}$$

where u is only allowed to be 0 or 1. We have been choosing the possibility of 0 or 1 depending on whether or not we are in V or outside of V. We suppose that the measurements of z are not exact, namely, we do not know the value of z with infinite precision. Imagine, for example, imposing one possibility or another depending on some very noisy and highly oscillatory imprecision as to where the boundary of V_i lies.

Now recall the Fillipov lemma. [9] The lemma asserts that the accessible set for a control system with only binary off-on ("bang-bang") controls as above, agrees with the accessible set for the convex hull of the two vector fields. In particular, at points z_* where Y and $\nabla\phi$ are linearly dependent and pointing in the same direction, we can choose a system of controls which turns this z_* into a fixed point. Now model this control with uncertainty on the measurements of z. A bit more work turns the new fixed point z_* into a stable fixed point under the effect of such uncertainty.

Are there really points z_* so that we can write $0 = -u\nabla\phi(z_*) + (1-u)Y(z_*)$ for some u, $0 \leq u \leq 1$? The degree of $-\nabla\phi$ at p_i is $(-1)^{k_i}$ and, in particular, is nonzero. It follows that $\nabla\phi/\|\nabla\phi\|$ sweeps out all possible points $e \in \S^{n-1}$ of the unit sphere as z varies over a small sphere surrounding p_i. By elementary topology (∂V_i is homologous to the boundary of this small sphere) the same is true as z varies over ∂V_i. In particular, there will be points $z \in \partial V_i$ where $Y = e_1$ and $\nabla\phi$ point in the same direction [10] and we can then use u to scale accordingly. We have our u and our point z_*.

We have indicated how arbitrarily small uncertainty, such as measurement noise, can render our previously unstable fixed point p_i for the gradient flow into a stable fixed point if we try to implement our above discontinuous alteration of gradient flow. This is not a good solution if we want to achieve our goal.

6.3.3 Robustness and Measurement

Measurements come with uncertainties. So do control forces or torques. The environment in which our controlled object moves has noise, wind, uneven terrain, etc. And our physical analog model of our system, the way we encode it as an ODE, will be imprecise. It turns out that measurement noise can wreak havoc with discontinuous vector fields, rendering previously unstable locations stable and inadvertently hanging us up indefinitely near one of the unstable equilibria p_i. The goal of robust

[9] We do not mean to trivialize the result or Hermes discussion of it. There is a deep and non-trivial discussion of what is meant by a "Fillipov solution" and the consequent measure theory around it in [9] and in the subsequent literature.

[10] If, as in the figure, our ∂V_i has corners, use the usual subdifferential style tangent space at the corners a la Clarke and this argument still works.

control is to guarantee that we arrive within a pre-specified window of our desired goal m_0 in the presence of appropriately bounded uncertainty.

Suppose our state space is a real vector space, say $\mathbb{R}^n$, and on it we have an expected or "nominal" [11] vector field $z \mapsto F(z)$. We imagine this vector field arriving to us after *implementing some feedback control loop*. So we expect that the system evolves according to

$$\dot{z} = F(z) \qquad z \in \mathbb{R}^d. \tag{6.7}$$

Measurement uncertainty corresponds to not knowing exactly where we are. So replace the state variable z at which we evaluate the vector field F by $z + \eta_m(z,t)$ where $(z,t) \mapsto \eta_m(z,t)$ represents measurement noise. We allow η_m to depend on time since measurement noise could be time dependent. We want to compare the end results of our nominal ODE in (6.7) with that of its "nearby cousins"

$$\dot{z} = F(z + \eta_m(z,t)). \tag{6.8}$$

Suppose that the nominal system has the origin as a global attractor. Do the cousins continue to have the origin as global attractor? This is too much to hope for, since it would require that the noise vanish at the origin.

Definition 5 (Robustness to measurement error). *Suppose the nominal vector field (6.7) – imagined to arise from a feedback stabilization control scheme – has the origin as a global attractor. Then, we say this control scheme (or its vector field) is "robust" to measurement errors if, given any $\delta > 0$ sufficiently small we can find $\epsilon > 0$ such that all trajectories to all the noisy cousins (6.8) to the nominal control with $\|\eta_m\| < \epsilon$ converge to a δ-ball of the origin as time tends to infinity.*

Remark 1. Of course the norm used to measure $\|\eta_m\|$ will matter! We use the sup norm.

Measurement Noise on Manifolds

We are in a decidedly vector space setting in this formulation of robustness since we cannot add points on manifolds! If $F : M \to TM$ is a vector field on a manifold the expression $F(z + \eta_m(z,t))$ does not make sense! We cannot add points on a manifold. Even if we could, $F(z+\eta)$ would be a vector in the tangent space to M at $z+\eta$, not to the tangent space of M at z, so it would not represent a vector field. Rather than follow these lines to try to make sense of measurement noise and robustness on a manifold, we return to the control theory drawing board and look into where F comes from. Notably, we introduce the control-theoretic idea of a "measurement" in addition to "control" and "feedback" Rewrite our original system in the traditional form

$$\dot{z} = f(z,u) \qquad z \in M, u \in \mathbb{R}^m$$

where the controls u take values in a convex subset of $\mathbb{R}^m$. Naturally,

$$f : M \times \mathbb{R}^m \to TM$$

with

[11] Dictionaries give multiple definitions of "nominal." Here, by "nominal" we mean that the system is operating without perturbations, namely, the system under study has state z that is precisely governed by (6.7).

$$f(z,u) \in T_z M \qquad \forall z \in M$$

uniformly on u. See also Brockett [6] who takes the u's to vary within an auxiliary vector bundle over M. We want to implement a feedback law $u = \kappa(z)$ in a way which allows the modeling of measurement noise. To do this we introduce the *intermediary of a measurement.*

Definition 6. *A measurement on M is a vector valued map*

$$h : M \to \mathbb{R}^\ell, \qquad z \mapsto y = h(z)$$

meant to model the sampling and recording of partial information regarding the state $z \in M$.

We insist that our feedback laws depend only on what we measure, that is,

$$u = \kappa(h(z)),$$

where, now

$$\kappa : \mathbb{R}^\ell \to \mathbb{R}^m$$

represents our feedback law. Since h takes values in a vector space, we can simply add time-dependent measurement uncertainty $\eta_m : M \times \mathbb{R} \to \mathbb{R}^\ell$ to our measurements by

$$h \mapsto h + \eta_m; \qquad \eta_m : M \times \mathbb{R} \to \mathbb{R}^\ell$$

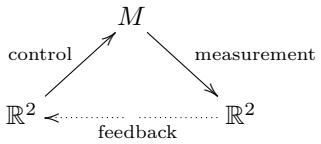
thus replacing the feedback law $z \mapsto \kappa(h(z))$ by its nearby noisy cousins given by

$$\kappa(h(z) + \eta_m(z,t)).$$

We have set things up now so that we can define "robustness to measurement error" in essentially a way identical to our earlier definition. We merely replace the feedback law in $F(z, \kappa(h(z)))$ by its perturbation $F(z, \kappa(h(z) + \eta_m(z,t))$.

Remark 2. Modeling environmental noise, control noise, and uncertainty in the model are all straightforward on a manifold. They correspond to the perturbations $F(z,u) + \eta_{env}$, $F(z,u) \to F(z, u + \delta u)$, and $F(z,u) \to F(z,u) + (\delta F)(z,u)$, respectively.

We can summarize what we have done using a commutative diagram where the dotted arrow of "feedback" closes the loop. In the case of our example of stabilizing to $m_0 \in M$, we will see next that we need two feedback control laws and two measurements, so $k = \ell = 2$.



6.3.4 Our New Setup

To put our "gradient flow / breeze system" into this framework introduce two controls u_1, u_2 so as to encode our system as the control system:

$$\dot{z} = -u_1 \nabla\phi(z) + u_2 Y(z). \tag{6.9}$$

If $u_1 = 1$ and $u_2 = 0$ we have pure gradient flow. If $u_1 = 0$ and $u_2 = 1$ we have pure steady breeze. Introduce measurements $y : M \to \mathbb{R}^2$ where $y = (y_1, y_2)$ is the function

$$y_1 = \|\nabla\phi(z)\|, \qquad y_2 = \phi(z) \tag{6.10}$$

We will be continuously monitoring y_1. Whenever y_1 is sufficiently small, we are in a Morse neighborhood U_i of one of the p_i or perhaps of m_0. We can decide which point p_i or m_0 z is closest to (and closed to which neighborhood U_i), by measuring y_2 and comparing it to the possible (known) critical values of ϕ.

Preparing the Morse Function for Hybridization

Recall that our goal point m_0 is the globaly minimum of ϕ and its only local minimum. Translating $\phi \mapsto \phi - \phi(m_0)$ insures that $\phi(m_0) = 0$ so that $\phi(z) > 0$ for each $z \neq m_0$. We have also assumed (by a wiggling of ϕ) that the critical values $\phi(p_i)$ of ϕ are all distinct. (See Proposition 1 above.) Scaling ϕ by a (possibly large) scalar $K > 0$, we can separate the critical values so they are all at least a unit apart

$$p_i \neq p_i \implies \phi(p_i) - \phi(p_j) \geq 1$$

and

$$\phi(p_i) \geq 1.$$

This scaling of ϕ can be used to ensure that the rescaled ϕ also enjoys the property that $\{z : \|\nabla\phi(z)\| < 1\}$ consists of $N+1$ topological (open) balls $W_1, W_2, \ldots, W_N$, one for each critical point p_i, and one, say W_0 for m_0, and that each of these balls is contained in a Morse neighborhood U_i of the critical point. This scaling and translating of ϕ does not change the location of the critical points p_i or their index.

Note that scaling ϕ by K rescales both the Morse coordinates y, x and the breeze Y by $1/\sqrt{K}$.

For each $i = 1, \ldots, N$, we may take the breeze neighborhoods V_i on which the flow of Y is well controlled and brings us to $\phi < c_i - k_2$ so that $V_i \subset W_i$. Note that the intersections of W_i with $\{z : \|\nabla\phi(z)\| < r\}$ form a family of nested balls converging to p_i as $r \to 0$. Now choose k_3 small enough so that

$$B_i := \{z : \|\nabla\phi(z)\| < k_3\} \cap W_i \subset V_i.$$

and that the boundary of B_i and of V_i are disjoint. See Figure 6.4. Since $\|\nabla\phi\|$ acts as a measure of distance from p_i, we have that $k(i) > k_3$ for each i, where

$$k(i) = \min_{p \in \partial V_i} \|\nabla\phi(p)\|.$$

Set

$$k_V = \min_i k(i).$$

Our "margin of robustness" – the measurement tolerance we need to guarantee for $y_1 := \|\nabla\phi\|$ to ensure that our control scheme will stabilize z to m_0 – is some fraction of the minimum of $k_V - k_3$ and k_3. With such a measurement area we can be sure to distinguish between being inside B_i and leaving V_i,

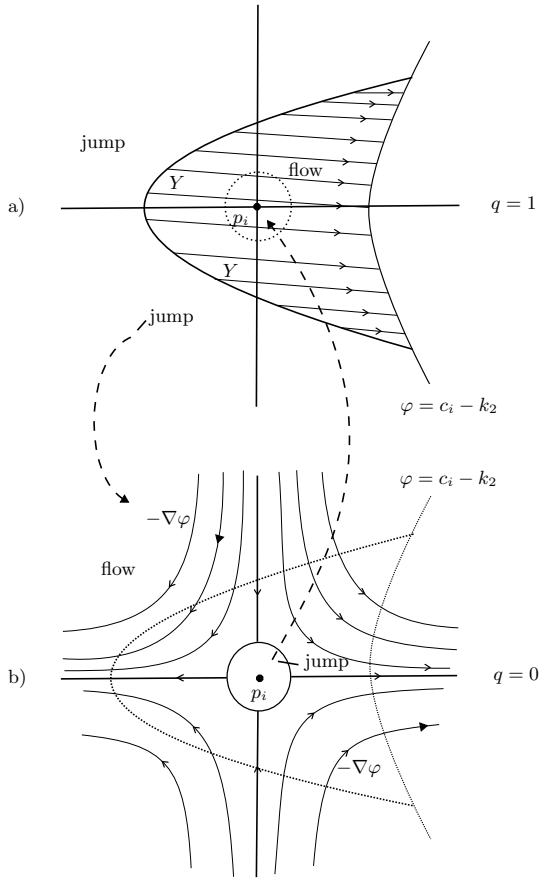


Fig. 6.4: The disc B_i centered about the unstable equilibrium p_i of the gradient flow forms a connected component of the jump set for $q = 0$. The set B_i is contained in the parabolic V_i which is a component of the flow set for $q = 1$, whose flow is that of Y. The complement of the union of the V_i forms the jump set for $q = 1$. The exterior of the union of the B_i is the flow set for $q = 0$ for the gradient flow. The $q = 1$ flow lines in V_i terminate when $\phi \leq c_i - k_2$. Sample jumps from $q = 0$ to $q = 1$, and vice versa, are marked with dashed arrows.

6.3.5 Hybridizing

Let us introduce the discrete variable

$$q \in \{0, 1\}$$

which will toggled *on* or *off* to define a hybrid feedback control law depending on the measurements. The role of q is select whether $-\nabla\phi$ or Y should update z during flows when the state is in the so-called *flow set*, which we denote by C. The toggles of q occur when the state is in the so-called *jump set*, which we denote as D. Specifically, we define the state of the closed-loop system with the hybrid feedback controller as (z, q), whose goal is to globally and robustly asymptotically stabilize

$$M \times \{0\}.$$

The flow set C is defined as the union of the sets $C_0 \times \{0\}$ and $C_1 \times \{1\}$, and the jump set D as the union of the sets $D_0 \times \{0\}$ and $D_1 \times \{1\}$, where the sets C_0, C_1, D_0, and D_1 are defined next. Set

$$D_0 := \bigcup_{i=1}^{N} B_i, \qquad C_1 := \bigcup_{i=1}^{N} V_i,$$

the index of the disjoint union running from 1 to N, the labels of the nonminimal critical points p_i. Since $B_i \subset V_i$, we have that $D_0 \subset C_1$ – in fact, C_1 contains a neighborhood of D_0. Use these sets to define two partitions of M, namely

$$C_0 := M \setminus D_0, \qquad D_1 := M \setminus C_1.$$

To properly selects the among the two feedback laws, we define the *jump map* as the map that keeps z constant and toggles q from 0 to 1 or from 1 to 0 when the state z is in the jump set. More precisely, we denote the jump map as

$$G : M \times \{0,1\} \to M \times \{0,1\}$$

and define it as

$$G(z,1) := (z,0), \qquad G(z,0) := (z,1).$$

The state z is updated continuously according to the *flow map* obtained from (9.31), which is

$$F(z,u) := (-u_1 \nabla\phi(z) + u_2 Y(z), 0)$$

where $u = (u_1, u_2)$ and, conveniently, we apply the feedback law

$$\kappa(z,0) := (1,0), \qquad \kappa(z,1) := (0,1).$$

Then, during flow – that is, when $(z,q) \in C$, the state (z,q) is governed by

$$(\dot{z}, \dot{q}) = F(z,q) = (-\kappa(z,q)\nabla\phi(z) + \kappa(z,q)Y(z), 0)$$

while at jumps, which occur when $(z,q) \in D$, the state (z,q) is updated by

$$(z^+, q^+) = G(z,q) = (z, 1-q)$$

The flow and jump dynamics described above lead to the hybrid closed-loop system given as

$$\mathscr{H} \;:\; \begin{cases} (\dot{z}, \dot{q}) = F(z,q) & (z,q) \in C \\ (z^+, q^+) = G(z,q) & (z,q) \in D \end{cases} \tag{6.11}$$

Our hybrid stabilization scheme operates as follows:

- If $q = 0$ and $z \in C_0$, z flows according to the first component of $F(z,0)$, namely, $-\nabla\phi(z)$. As we do so, the controller measures $y_1 = \|\nabla\phi(z)\|$ and $y_2 = \phi$. If y_1 ever crosses below the threshold value k_3 while $y_2 = \phi$ is greater than c_1, the smallest nonzero critical value of ϕ, then z entered $D_0 = \bigcup_i B_i$. If $z \in D_0$, then the jump map is applied to reset q to 1 – note that z remains at the same point in B_i. Since $B_i \subset V_i$, z can flow with $q = 1$.

- If $q = 1$ and $z \in C_1$, z flows according to the first component of $F(z, 1)$, namely, $Y(z)$ while measuring $y_2 = \phi(z)$. The value of y_2 will be close to some critical value c_i. Eventually, y_2 crosses below $c_i - k_2$, which means that z leaves C_1 and enters D_1. If $z \in D_1$ with $q = 1$, then the jump map is applied to reset q to zero, and z remains unchanged. Since $B_i \subset V_i$, z is outside B_i and so in the flow regime for $q = 0$.

Since $C_0 \cup D_0 = M$ and $C_1 \cup D_1 = M$, the above rules cover all possibilities for points $(z, q) \in M \times \{0, 1\}$. Teaders can convince themselves that this scheme provides a global feedback stabilization law to m_0.

Robustness of the scheme follows from the strict containment $C_0 \subset D_1$. Specifically, the scheme we just described is robust to measurement errors in y_1 provided these errors are small enough to allow us to distinguish between being inside a B_i and leaving a V_i. We can quantify the error bounds by recalling that $y_1 := \|\nabla\phi(z)\|$, $y_1 = k_3$ on ∂B_i and $k(i) = \min_{p\in\partial V_i} \|\nabla\phi(p)\| > k_3$. Set $k_V = \min_i k(i)$ and $k_* = \frac{1}{2}\min\{k_3, k_V - k_3\}$. If our error bars on measuring y_1 are less than k_* then by evaluating y_1 we can guarantee whether or not we are in B_i or have left V_i with sufficient accuracy as to know whether we should be flowing or jumping. We call k_* the "margin of robustness" for this scheme.

Theorem 1 (Theorem). *On any compact connected n-dimensional manifold M, and for any chosen point m_0 of that manifold, we can design a hybrid control system whose logic part consists of a single bit $q \in \{0, 1\}$ as in* (6.11) *and which has $\{m_0\} \times \{0\}$, as a global attracting and stable set, this property being robust with respect to measurement and all other errors in the system.*

Why the parabolic-shaped V_i?

In Figure 6.4, we have made V_i so as to have a parabolic boundary capped by a level set of ϕ. We did this to guarantee that the vector field Y is transverse to the boundary ∂V_i everywhere except at the points where the cap joins the parabola. Being transverse is "robust" (unchanged by perturbations) whereas tangency is easily destroyed by perturbations. This is why we prefer the parabolic profile for the boundary.

Here is how to make such a parabolic neighborhood. Begin with a standard flox-box for Y. In flow-box coordinates, the flow-box is a cylinder of the form tube has the form $I \times B$, where $I = [-T, T] \subset \mathbb{R}$ is in the $Y = \mathrm{d}y_1$ direction and B is a solid unit ball in $\mathbb{R}^{n-1}$. For simplicity of notation, label the coordinates of $\mathbb{R}^{n-1}$ as x_a instead of the old $(x_a, y_b), b > 1$. Then, the flow tube can be expresses as $\rho \leq 1, -T \leq y_1 \leq T$, where

$$\rho = \sqrt{\sum_a x_a^2}.$$

Now take any smooth strictly monotonic increasing function $g : [-T, T] \to [0, 1]$, $g = g(y_1)$, which starts out either with $g(-T) = 0$ and increases strictly monotonically to $g(T) = 1$. (For a standard parabola take $g(y) = \frac{1}{4T^2}(y + T)^2$.) Our neighborhood is given by $\{(y_1, \rho) : \rho \leq g(y_1)\}$. This parabolic neighborhood has the property that all trajectories of Y enter into it through the parabolic bottom and leave it along the cap with $\phi = c_i - k_2$. Since transversality cannot be changed by small perturbations, a perturbed $\tilde{Y} = Y + w$ will continue to have these nice entrance and exit properties.

6.3.6 Solutions to Hybrid systems

Some words are in order regarding what we mean by a "solution" to this system which is a combination of continuous flow and discrete jump. The instantaneous state of our hybrid system is a $(z, q) \in M \times \{0, 1\}$, where the index q indicates that we should think of $z \in M_q$.

HYBRID TIME In the hybrid literature one keeps track of jumps by introducing a discrete integer time $j \in \mathbb{N}$ as well as the continuous time. Solutions are parameterized by "stair steps" $E \subset \mathbb{R} \times \mathbb{N}$. These stair steps are graphs of piecewise constant monotone functions taking integer values with jumps of 1. In other words, $E = \bigcup_{j=1,N} I_j \times \{j\}$, where, for the particular construction in (6.11) $I_j \subset \mathbb{R}$ are the intervals whose endpoints are where the jumps in $q \to \bar{q}$ occur. So, in this case, the right endpoint of I_j equals the left endpoint of I_{j+1}. In the open part of each interval, (z, q) flows according to F. The continuous variable z flows on the flat part of each step, i.e., on the interior of the I_j's. At jumps, $j \to j+1$ from one step to the next, (z, q) is reset by the jump map, which keeps z constant and flips q. Using this language, one expresses solutions as maps $x : S \to M \times \{0, 1\}$ by writing $x(t, j) = (z(t, j), q(t, j))$. For (6.11), the discrete variable is constant on each open interval $int(I_j) \times \{j\}$. It makes a jump at the transition from one interval (step) to the next. (In the general case, solutions may be such that z exhibits jumps: $z(t, j) \to z(t, j+1) = G(z(t, j), q(t, j))$ according to some pre-specified collection of maps $G(\cdot, q)$ whose domains and ranges may depend on q.)

HOW MANY JUMPS? If a solution to (6.11) starts with $q = 0$ then typically we expect that there will be no jump at all. The initial z would lie in the basin of attraction of m_0 and its trajectory would avoid all of the B_i, so the gradient flow would take it all the way down to m_0. Similarly, if a solution starts with $q = 1$ and in C_1, we expect that there will be a single jump, followed by a gradient flow all the way to m_0.

In the worst case, if the solution starts with $q = 0$ with z sitting at the global maximum for ϕ, there could be as many as $2N$ jumps, with two jumps per critical point until z enters a ball about m_0. There are two jumps per close encounter with a critical point p_j, one upon entering B_j from 0 to 1 to turn on the breeze, and then one upon leaving V_j from 1 to 0 to turn back on the gradient flow. We can insure fewer jumps if the gradient flow is *Morse-Smale*. Let $\beta \leq n + 1$ be the number of indices k such that the kth Betti number $b_k(M)$ is nonzero. (Here n is the dimension of M.) To be Morse-Smale [12] means that the stable and unstable manifolds of all critical points intersect transversally and implies that whenever a trajectory connects one critical point p_i to another one p_j then the index of p_i is larger than that of p_j. If the balls B_j are sufficiently small and $-\nabla\phi$ is Morse-Smale, then trajectories of the gradient flow will only enter at most β balls as they travel down to m_0. We do not need to count the final ball about m_0 since solutions do not jump upon entering it. In this way, we get the worst-case scenario count of $2(\beta - 1) \leq 2n$ jumps total.

References

1. A. Agrachev, El Alaoui, El-Houcine Chakir, J-P. Gauthier, I. Kupka, Generic singularities of sub-Riemannian metrics on $\mathbb{R}^3$, *Comptes Rendus Acad. Sci.* v. 322, Serie I, 377–384 (1996).
2. A. Agrachev, B. Bonnard, M. Chyba and I. Kupka, SubRiemannian Sphere in Martinet Flat Case, *ESAIM: Control, Optimisation and Calculus of Variations*, v. 2, pp. 377-448 (1997).
3. Y. Baryshnikov, Topological Perplexity of Feedback Stabilization, *J. Appl. Comput. Topol.* 7(1): 75-87 (2023).

[12] Being Morse-Smale is a generic condition. Small perturbations of the function ϕ or Riemannian metric will insure that the flow is Morse-Smale

4. S. P. Bhat and D. S. Bernstein, A topological obstruction to continuous global stabilization of rotational motion and the unwinding phenomenon, *Systems and Control Letters*, v. 39, No. 1, 63-70 (2000).
5. M. Branicky, Introduction to Hybrid Systems, a chapter in the book *Handbook of Networked and Embedded Control Systems*, editor, D. Hristu-Varsakelis, and W. S. Levine, pp. 91-116, Birkhäuser Boston, Boston, MA, (2006), https://doi.org/10.1007/0-8176-4404-0_5
6. R.W. Brockett, *Nonlinear Control Theory and Differential Geometry*, International Congress of Mathematicians, Warsaw, (1983).
7. J-P Gauthier and Ivan Kupka, *Deterministic observation theory and applications*, Cambridge University Press, Cambridge, (2001).
8. V. Guillemin and A. Pollack, *Differential topology*, Reprint of the 1974 original, AMS Chelsea Publishing, Providence, RI, (2010).
9. H. Hermes, Discontinuous vector fields and feedback, pp 155-165 in the book *Differential Equations and Dynamical Systems*, Proc. of an Internat. Sympos., Mayaguez, P.R., (1965).
10. M. Levi, Geometrical aspects of rapid vibrations and rotations, *Phil. Trans. R. Soc. A* 377: 20190014. http://dx.doi.org/10.1098/rsta.2019.0014 (2019).
11. Ivan Kupka, Géomtrie sous-riemannienne, *Séminaire Bourbaki*, v. 38, pp. 351-380, (1995-1996) https://eudml.org/doc/110220
12. R. Mahony, V. Kumar, and P. Corke, Multirotor aerial vehicles: Modeling, estimation, and control of quadrotor, *IEEE Robotics and Automation Magazine*, v. 19, No. 3, 20-32 (2012).
13. M. Malisoff, M. Krichman, and E. Sontag, Global Stabilization for Systems Evolving on Manifolds, *Journal of Dynamical and Control Systems*, v. 12, No. 2, , 161-184, DOI: 10.1007/s10450?006?0379?x (2006).
14. R. Montgomery, (1994), Abnormal Minimizers, *SIAM J. Control and Optimization*, v. 32, no. 6, pp. 1605-1620.
15. J. Milnor, *Morse theory*, Annals of Mathematics Studies, No. 51, Princeton University Press, Princeton, NJ, (1963).
16. M. Peixoto, On an Approximation Theorem of Kupka and Smale *J. Diff. Eq.*, v. 3, 214-227 (1966).
17. R. Sanfelice, *Hybrid Feedback Control*, Princeton U. Press, Princeton, (2021).
18. C. T. C. Wall, *Differential Topology*, in *Cambridge Studies in Advanced Mathematics*, Cambridge University Press, Cambridge, (2016).

7

Optimisation of Functional Determinants on the Circle

J.-B. Caillau[1], Y. Chitour[2], P. Freitas[3], and Y. Privat[4]

[1] Université Côte d'Azur, CNRS, Inria, LJAD jean-baptiste.caillau@univ-cotedazur.fr
[2] Université Paris-Saclay, CNRS, CentraleSupélec, L2S yacine.chitour@centralesupelec.fr
[3] Universidade de Lisboa, Departamento de Matemática, Instituto Superior Técnico pedrodefreitas@tecnico.ulisboa.pt
[4] Université de Lorraine, CNRS, Inria, IECL yannick.privat@univ-lorraine.fr
Institut Universitaire de France

This project is partially supported by the FMJH Program PGMO and EDF-Thales-Orange (extdet PGMO grant), by the iCODE Institute, research project of the IDEX Paris-Saclay, by the Hadamard Mathematics LabEx (LMH) through the grant number ANR-11-LABX-0056-LMH in the "Programme des Investissements d'Avenir", and by the Fundação para a Ciência e a Tecnologia (Portugal) through project UIDB/00208/2020.

Summary. The functional determinant of elliptic differential operators on the circle was introduced in [3]. In the present paper, optimisation of this determinant over essentially bounded functions is studied as an optimal control problem on the special linear group of real matrices. In the one dimensional case, existence and uniqueness of maximisers and minimisers is proved.

7.1 Statement of the Problem

Following [3] we consider the determinant of a differential operator

$$A := \sum_{k=0}^{p} A_k D^k$$

defined on $\mathbb{R}^N$-valued functions, N a positive integer, where $D = -i\mathrm{d}/\mathrm{d}x$ is the complex valued derivation operator for such functions ($i^2 = -1$) and where the $A_k : \mathbb{S}^1 \to \mathrm{M}(N, \mathbb{R})$, $0 \leq k \leq p$, are matrix-valued (square matrices of order N) functions defined on the circle.[5] We are interested in addressing optimisation issues for such determinants under suitable restrictions on the potentials involved. For the rest of the paper, we identify $\mathbb{S}^1$ with $\mathbb{R}/\mathbb{Z}$ and functions on $\mathbb{S}^1$ with one-periodic functions. For $Q \in \mathrm{M}(N, \mathbb{R})$ we use the Frobenius norm $\|Q\| = \mathrm{tr}(Q^T Q)^{1/2}$ and recall it derives from the inner product on $\mathrm{M}(N, \mathbb{R})$ given by

[5] The fundamental reference for spectral problems on the circle $\mathbb{S}^1$ (geometrisation of the periodic boundary conditions) is [3], more general than [4]. The latter reference, however, provides much more elementary arguments enabling one to establish links with the discrete setting.

$$\langle Q_1, Q_2 \rangle = \mathrm{tr}(Q_1^T Q_2), \quad Q_1, Q_2 \text{ in } \mathrm{M}(N, \mathbb{R}). \tag{7.1}$$

We will assume that

$$A = -\,\mathrm{Id}_N \frac{\mathrm{d}^2}{\mathrm{d}x^2} + V(x), \tag{7.2}$$

i.e., the maximal order of differentiation p is equal to two, and the operator is in normal form with $A_2 = \mathrm{Id}_N$ (the identity matrix of order N), $A_1 = 0$ and $A_0 = V$ a Hill potential. Ray and Singer [7] define the *functional determinant* of such an operator as

$$\det A := e^{-\zeta_A'(0)} \tag{7.3}$$

where

$$\zeta_A(s) := \sum_{\lambda_j \neq 0} \frac{1}{\lambda_j^s}$$

the sum being taken over nonzero eigenvalues of A. The function ζ_A is well defined for s with a large enough real part (depending on the eigenvalues asymptotics), and has a meromorphic extension to the plane that is regular at $s = 0$ [6, 8]. While (7.3) clearly equals the product of eigenvalues when there are only finitely many of them, the expression provides a regularisation of the otherwise divergent product. It is proven in [3] that

$$\det A = (-1)^N \det(\mathrm{Id}_{2N} - R(A)) \tag{7.4}$$

with $R(A)$ the monodromy operator. More precisely, $R(A)$ is equal to the fundamental matrix at time 1 of the linear time-varying system on $\mathrm{M}(2N, \mathbb{R})$

$$\begin{cases} \dot{R}(x) = \mathscr{A}_{V(x)} R(x), \\ R(0) = \mathrm{Id}_{2N}, \end{cases} \tag{7.5}$$

where one sets

$$\mathscr{A}_Q := \begin{bmatrix} 0 & \mathrm{Id}_N \\ Q & 0 \end{bmatrix}, \text{ for every } Q \in \mathrm{M}(N, \mathbb{R}).$$

Remark 1. In [3], the potential V appears as $-V$ in (7.5) and we have changed notations in order to stick with previous optimisation literature [1].

Since its trace is zero, the matrix $\mathscr{A}_V$ belongs to the lie algebra $\mathfrak{sl}(2N, \mathbb{R})$ and (7.5) defines a dynamics on the special linear group $\mathrm{SL}(2N, \mathbb{R})$, a Lie group of dimension $4N^2 - 1$. This dynamics is bilinear in R and V. We are now in position to properly define the optimisation problems discussed in the present paper.

For every positive M, the set $\mathscr{V}_M$ of admissible Hill potentials is given by the measurable functions V so that

$$\mathscr{V}_M = \{V : [0,1] \to \mathrm{M}(N, \mathbb{R}) \mid \operatorname*{ess\,sup}_{x \in [0,1]} \|V(x)\| \leq M^2\}, \tag{7.6}$$

and we say that a potential V satisfies an L^∞-constraint if it belongs to some $\mathscr{V}_M$.

Remark 2. Note that $\mathscr{V}_M$ is convex and invariant by transposition and conjugation by orthogonal matrices, i.e. $V_{U(\cdot)} = U^T(\cdot)V(\cdot)U(\cdot)$ belongs to $\mathscr{V}_M$ if and only V does, for any measurable $\mathrm{SO}(N)$-valued $U(\cdot)$ defined on $[0,1]$. One could have defined equivalently $\mathscr{V}_M$ with potentials $V : \mathbb{R} \to \mathrm{M}(N, \mathbb{R})$ periodic of period 1 and satisfying the same L^∞ bound. In that case, $\mathscr{V}_M$ is clearly invariant by translation of $x_0 \in \mathbb{R}$, *i.e.* $V_{x_0}(\cdot) = V(\cdot + x_0)$ belongs to $\mathscr{V}_M$ if and only V does.

Remark 3. For $q \in [1, \infty)$, one could replace the L^∞ constraint by the integral condition

$$\int_0^1 \|V(x)\|^q \, \mathrm{d}x \leq M^{2q}$$

which is referred to as an L^q-constraint.

The cost function associated to a potential V is from now on denoted $\mathscr{C}(V)$ and is given by

$$\mathscr{C}(V) = (-1)^N \det\big(\mathrm{Id}_{2N} - R(1)\big), \tag{7.7}$$

where R is defined in (7.5). We will study the following optimisation questions: for every $M > 0$,

$$\mathbf{Max-Det(M)}: \quad \max_{V \in \mathscr{V}_M} \mathscr{C}(V) \text{ subject to (7.5)}, \tag{7.8}$$

$$\mathbf{Min-Det(M)}: \quad \min_{V \in \mathscr{V}_M} \mathscr{C}(V) \text{ subject to (7.5)}. \tag{7.9}$$

To derive common statements for both optimisation problems, we use $\mathscr{C}_\varepsilon$ to denote $\varepsilon\mathscr{C}$ where $\varepsilon = \pm 1$ and in that way **Max-Det** becomes the minimisation of $\mathscr{C}_-$ while **Min-Det** is simply the minimisation of $\mathscr{C}_+$. That is we study, for a given $M > 0$,

$$\mathbf{Ext-Det}_\varepsilon\mathbf{(M)}: \quad \min_{V \in \mathscr{V}_M} \mathscr{C}_\varepsilon(V) \text{ subject to (7.5)}. \tag{7.10}$$

This problem is a Mayer optimal control problem with state R in $\mathrm{SL}(2N, \mathbb{R})$, potential V (control) valued in a Euclidean ball of $\mathrm{M}(N, \mathbb{R})$, and bilinear dynamics. Control problems on Lie groups were intensively studied by Ivan Kupka and his collaborators [2], and were foundational for what has ever since emerged as *Geometric control theory*.

We begin our analysis in Section 7.2 by stating the necessary condition satisfied by optimisers (existence is clear). The problem can be formulated as an optimal control problem over the set of matrices with a matrix valued control, so the Pontryagin maximum principle provides the appropriate information. We also obtain some additional properties of optimisers Section 7.3. In Section 7.4 we focus on the one-dimensional case. We prove existence and uniqueness of maximisers and minimisers for the determinant over a bounded set in $L^\infty(\mathbb{S}^1)$.

7.2 Optimality Conditions

In this section, we will derive the equations verified by the minimisers of $\mathbf{Ext-Det}_\varepsilon$ as well as their first properties. From now on, M is an arbitrary positive number and $\varepsilon \in \{-1, 1\}$. First of all, since $\mathscr{V}_M$ is non empty and, for any $R \in \mathrm{M}(2N, \mathbb{R})$, the set $\{\mathscr{A}_V \mid V \in \mathrm{M}(N, \mathbb{R}),\ \|V\| \leq M^2\}$ is compact and convex, then $\mathbf{Ext-Det}_\varepsilon\mathbf{(M)}$ admits minimisers according to Filippov theorem. According to the Pontryagin maximum principle (PMP), a solution R of $\mathbf{Ext-Det}_\varepsilon\mathbf{(M)}$ with minimising potential V is necessarily the projection of an *extremal*, *i.e.*, an integral curve $\lambda = (R, P) \in \mathrm{M}(2N, \mathbb{R})^2$ of a Hamiltonian vector field satisfying certain additional conditions. We hereby present a definition of extremal adapted to our setting. The fact that this is equivalent to the standard definition of normal extremal is the subject of Proposition 1 given below.

Definition 1. *A curve* $\lambda : [0, T] \to \mathrm{M}(2N, \mathbb{R})^2$ *is called* extremal *with respect to the control* $V \in \mathscr{V}_M$ *if:*

(i) Letting $\lambda = (R, P)$, it satisfies

$$\dot{R}(x) = \mathscr{A}_{V(x)} R(x), \tag{7.11}$$

$$\dot{P}(x) = -\mathscr{A}^T_{V(x)} P(x). \tag{7.12}$$

(ii) It holds that $R(0) = \mathrm{Id}_{2N}$ and the following transversality condition holds true[6]

$$P(1) = (-1)^N \varepsilon \ \mathrm{Com}\big(\mathrm{Id}_{2N} - R(1)\big). \tag{7.13}$$

(iii) Assume moreover that there exists $h \in \mathbb{R}$ such that a.e. on $[0,1]$

$$h = H\big(R(x), P(x), V(x)\big) = \max_{\|W\| \leq M^2} H\big(R(x), P(x), W\big), \tag{7.14}$$

where H is the Hamiltonian function defined on $\mathrm{M}(2N,\mathbb{R})^2 \times \mathrm{M}(N,\mathbb{R})$ by

$$H\big(R, P, W\big) = \langle P, \mathscr{A}_W R\rangle = \langle \mathscr{A}^T_W P, R\rangle. \tag{7.15}$$

such an extremal is called strong extremal.

Remark 4. Note that every potential V admits a unique extremal (which is possibly strong).

We then get the following.

Proposition 1. *Let $R : [0,T] \to \mathrm{M}(2N,\mathbb{R})$ be an optimal trajectory of $\mathbf{Ext} - \mathbf{Det}_\varepsilon(\mathbf{M})$ with minimising potential V. Then R is the projection on $\mathrm{M}(2N,\mathbb{R})$ of a unique strong extremal $\lambda = (R, P) : [0,T] \to \mathrm{M}(2N,\mathbb{R})^2$.*

Proof. Let V be a minimising potential of $\mathbf{Ext} - \mathbf{Det}_\varepsilon(\mathbf{M})$ and R the associated trajectory by (7.5). Pontryagin maximum principle implies that there exists a nontrivial pair (p^0, P) where the cost multiplier p^0 is a nonpositive real number and the covector $P : [0,1] \to \mathrm{M}(2N,\mathbb{R})$ is a Lipschitz function so that

1. $(R(x), P(x)) \in \mathrm{M}(2N,\mathbb{R})^2$ satisfy on $[0,1]$ the adjoint equations:

$$\dot{R} = \nabla_P H, \tag{7.16}$$

$$\dot{P} = -\nabla_R H; \tag{7.17}$$

2. we have the maximality condition given by (7.14);
3. the following transversality condition holds true: $P(1) = p^0 \nabla \mathscr{C}_\varepsilon(V)$.

In addition, since H does not depend on time, its value in (7.14) does not depend on time and is denoted by the constant real number h. As

$$\nabla_P H = \mathscr{A}_W R, \quad \nabla_R H = \mathscr{A}^T_W P, \quad \nabla \det(\mathrm{Id}_{2N} - R) = -\,\mathrm{Com}(\mathrm{Id}_{2N} - R),$$

the items of Proposition 1 follow at once, except the facts that p^0 can be taken equal to -1 and λ is unique. To establish the first fact, it is enough to show that p^0 cannot be null. To show that, we argue by contradiction and, in that case, it follows that $P(1) = 0$. Since (7.19) is linear in P,

[6] We denote $\mathrm{Com}(M)$ the comatrix of a square matrix M.

one gets that P is identically equal to zero on $[0,1]$. This contradicts the non triviality of the pair (p^0, P) and hence $p^0 \neq 0$. Regarding the uniqueness of λ, note first that, given $M > 0$, trajectories of (7.5) are in one to one correspondence with potentials in $\mathscr{V}_M$, since to each such trajectory, there is a unique potential $V \in \mathscr{V}_M$ necessarily defined as the lower left $N \times N$ block of $\mathscr{A}_V = \dot{R}R^{-1}$ (recall that R is absolutely continuous). By Item 3., $P(1)$ is determined by $R(1)$ and hence P is computed from (7.17). □

To take advantage of the maximisation condition (7.14), after defining $q = PR^T$, we rewrite Proposition 1 only using q and we deduce at once that

Proposition 2. *Assume that a trajectory R of $\mathbf{Ext} - \mathbf{Det}_\varepsilon(\mathbf{M})$ with potential V is the projection of an extremal trajectory $\lambda = (R, P)$. Define*

$$q = PR^T = \begin{bmatrix} Z_1 & \psi \\ \varphi & Z_2 \end{bmatrix}, \tag{7.18}$$

where the various blocs are $N \times N$ matrices. Then the dynamics of q is given, a.e. on $[0,1]$, by[7]

$$\dot{q}(x) = \left[q(x), \mathscr{A}^T_{V(x)}\right], \quad q(1) = (-1)^N \varepsilon \ \mathrm{Com}\left(\mathrm{Id}_{2N} - R(1)\right) R^T(1), \tag{7.19}$$

which yields, for a.e. $x \in [0,1]$,

$$\dot{Z}_1 = \psi - V^T\varphi, \tag{7.20}$$
$$\dot{\varphi} = Z_2 - Z_1, \tag{7.21}$$
$$\dot{\psi} = Z_1 V^T - V^T Z_2, \tag{7.22}$$
$$\dot{Z}_2 = \varphi V^T - \psi. \tag{7.23}$$

The Hamiltonian function H defined in (7.15) is equal to

$$H\big(R, P, W\big) = \langle q, \mathscr{A}^T_W \rangle = \mathrm{tr}(\psi) + \langle \varphi, W \rangle. \tag{7.24}$$

Moreover, it holds

$$q^T(x) = R(x) q^T(1) R^{-1}(x), \ \textit{for every } x \in [0,1], \tag{7.25}$$
$$\ddot{\varphi} = -2\psi + V^T\varphi + \varphi V^T \ \textit{for a.e. } x \in [0,1], \tag{7.26}$$

and in particular $q(\cdot)$ is periodic of period one.

Assume moreover $\lambda = (R, P)$ is a strong extremal. If $\varphi(x) \neq 0$, then $V(x) = M^2 \frac{\varphi(x)}{\|\varphi(x)\|}$ and, for every $x \in [0,1]$, it holds

$$h = \mathrm{tr}(\psi) + M^2 \|\varphi(x)\|, \tag{7.27}$$
$$\mathrm{tr}(\ddot{\varphi}) = -2h + 4M^2 \|\varphi(x)\|. \tag{7.28}$$

Proof. Most the above is immediate except (7.25). The latter follows from the fact that, for every $x \in [0,1]$,

$$q^T(x) = R(x) R^{-1}(1) q^T(1) R(1) R^{-1}(x).$$

The above equation then yields (7.25) after noticing that $R(1)$ and $q^T(1)$ commute. □

From now on, we indifferently call extremal either the pair (R, P) or the pair (R, q).

[7] We denote $[Q_1, Q_2] = Q_1Q_2 - Q_2Q_1$ the commutator of matrices.

Remark 5. In the light of Item (*iii*) of the above proposition, one can see that the potential V is not (immediately) defined at a zero of φ. In the sequel, the latter function φ is refereed to as the *switching function* and we single out a particular instance of zero of φ, namely that of *switching time* defining such a point $x_* \in (0,1)$ for which $\varphi(x_*) = 0$ and there exist two sequences $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ of two by two distinct points, both converging to x_* such that $\langle \varphi(x_n), \varphi(y_n)\rangle < 0$ for $n \in \mathbb{N}$. Clearly, a zero of φ in $(0,1)$ which is not a zero of $\dot\varphi$ is a switching time.

Remark 6. At every $R \in \mathrm{SL}(2N,\mathbb{R})$, the tangent space is

$$T_R\,\mathrm{SL}(2N,\mathbb{R}) = \{rR \mid r \in \mathrm{M}(2N,\mathbb{R}) \text{ such that } \mathrm{tr}(r) = 0\}. \tag{7.29}$$

Using now the inner product introduced in (7.1), one can identify the cotangent space $T_R^*\,\mathrm{SL}(2N,\mathbb{R})$ as

$$T_R^*\,\mathrm{SL}(2N,\mathbb{R}) = \{q(R^{-1})^T \mid q \in \mathrm{M}(2N,\mathbb{R}) \text{ such that } \mathrm{tr}(q) = 0\}. \tag{7.30}$$

We next notice that the flow associated with (7.19) is isospectral (*cf.* for instance [5]), in particular the trace of q is constant on $[0,1]$ equal to $\mathrm{tr}(q(1))$. Define indeed

$$\tilde q(x) = q(x) - \frac{\mathrm{tr}(q(1)}{2N}\,\mathrm{Id}_{2N}, \quad \tilde P(x) = \tilde q(x)\big(R^T(x)\big)^{-1}, \text{ for } x \in [0,1].$$

Clearly the curve $\tilde\lambda = (R, \tilde P)$ takes values in $T^*\,\mathrm{SL}(2N,\mathbb{R})$ and is an integral curve of the Hamiltonian vector field $\vec H$ associated with H. Finally, when applying the PMP to R, we claim that $\tilde\lambda$ turns out to be the required extremal with R as projection onto $\mathrm{SL}(2N,\mathbb{R})$: the dynamics of $\tilde\lambda$ has been described just previously, *i.e.*, $\dot{\tilde\lambda} = \vec H(\tilde\lambda)$, the maximality condition is exactly (7.14) and the tranversality condition (7.13) now says that $P(1) - p^0\nabla\mathscr{C}_\varepsilon(V)$ belongs to the normal cone at $T^*_{R(1)}\,\mathrm{SL}(2N,\mathbb{R})$, where the gradient is projected on $T^*_{R(1)}\,\mathrm{SL}(2N,\mathbb{R})$. Since that normal cone is equal to $\mathbb{R}(R^T(1))^{-1}$ and since one easily shows that $p^0 = -1$, one gets the claim regarding $\tilde\lambda$.

7.3 Invariance and Symmetries

We begin by providing the following property regarding translated potentials ensuring that the problem is well posed for controls defined on $\mathbb{S}^1 \simeq \mathbb{R}/$
mathbbZ. In particular, the uniqueness results of Section 7.4 are stated for controls in $L^\infty(\mathbb{S}^1)$.

Lemma 1. *Let R be a trajectory of* $\mathbf{Ext} - \mathbf{Det}_\varepsilon(\mathbf{M})$ *associated with potential V and corresponding extremal (R,q). For $x_0 \in \mathbb{R}$, consider the potential V_{x_0} translated from V according to Remark 2. Then V_{x_0} has same cost as V with corresponding extremal (R_{x_0}, q_{x_0}) and one gets that*

$$q_{x_0}(x) = q(x + x_0), \quad \varphi_{x_0}(x) = \varphi(x + x_0), \quad \forall x \in \mathbb{R}. \tag{7.31}$$

where φ (φ_{x_0}, respectively) denotes the switching function associated with V (V_{x_0}, respectively).

Proof. It is immediate to derive that the trajectory R_{x_0} of (7.5) associated with V_{x_0} is given by

$$R_{x_0}(x) = R(x + x_0)R(x_0)^{-1}, \quad \forall x \in \mathbb{R}, \tag{7.32}$$

and, by periodicity of V, it follows that

$$R_{x_0}(1) = R(x_0)R(1)R(x_0)^{-1}. \tag{7.33}$$

Using the above equation, one gets that

$$\mathscr{C}_\varepsilon(V_{x_0}) = (-1)^N \varepsilon \det(\mathrm{Id}_{2N} - R_{x_0}(1)) = \mathscr{C}_\varepsilon(V),$$

and hence has same cost as V. Let $\lambda_{x_0} = (R_{x_0}, P_{x_0})$ be the unique extremal associated with R_{x_0}. Then, from (7.25), it holds

$$q^T_{x_0}(x) = R_{x_0}(x)\big(R_{x_0}(1)\big)^{-1} q^T_{x_0}(1) R_{x_0}(1)\big(R_{x_0}(x)\big)^{-1}, \quad \forall x \in [0,1],$$

and, from (7.19), one has

$$\begin{aligned} q_{x_0}(1) &= (-1)^N \varepsilon \ \mathrm{Com}\,\big(\mathrm{Id}_{2N} - R_{x_0}(1)\big) R^T_{x_0}(1) \\ &= (-1)^N \varepsilon \ \mathrm{Com}\,\Big(\mathrm{Id}_{2N} - R(x_0)R(1)R(x_0)^{-1}\Big) R^T_{x_0}(1) \\ &= (-1)^N \varepsilon \ \big(R(x_0)^T\big)^{-1} \mathrm{Com}\,\big(\mathrm{Id}_{2N} - R(1)\big) R(x_0)^T R(x_0)^{-1} \\ &= \big(R(x_0)^T\big)^{-1} q(1) R(x_0)^T. \end{aligned}$$

Using the above equation, (7.32) and (7.33), one gets (7.31). □

We then prove that there always exists potentials V with negative costs, implying that minimal values for $\mathbf{Ext} - \mathbf{Det}_\varepsilon(\mathbf{M})$ are always negative, which in particular, exclude the zero potential from optimality.

Lemma 2. *The cost $\mathscr{C}_\varepsilon(0)$ associated with the zero potential is equal to zero. For every $N \times N$ diagonal matrix $D = \mathrm{diag}(\varepsilon_1 d_1^2, \cdots, \varepsilon_N d_N^2)$, where $\varepsilon_i^2 = 1$ and $d_i > 0$ for $1 \le i \le N$, let $\mathscr{C}_\varepsilon(D)$ be the cost associated with the constant potential equal to D. Then*

$$\mathscr{C}_\varepsilon(D) = (-2)^N \varepsilon \Pi_{i=1}^N \big(1 - c_{\varepsilon_i}(d_i)\big). \tag{7.34}$$

Moreover, for every $M > 0$, $D \in \mathscr{V}_M$ if $\sum_{i=1}^N d_i^2 \le M^2$ and then $\mathscr{C}_\varepsilon(D) < 0$ if one chooses $\varepsilon_1 \varepsilon = -1$, $\varepsilon_i = 1$ for $2 \le i \le N$ and d_1 not a multiple of 2π if $\varepsilon_1 = -1$.

Proof. One clearly has that the trajectory R_0 of (7.5) associated with the zero potential is equal to

$$R_0(x) = \begin{bmatrix} \mathrm{Id}_N & x\,\mathrm{Id}_N \\ 0 & \mathrm{Id}_N \end{bmatrix} \text{ for } x \in [0,1].$$

The conclusion follows at once. Using (7.43), one easily deduces the value resolvent matrix R_D associated with D at $x = 1$,

$$R_D(1) = \begin{bmatrix} \mathrm{diag}(c_{\varepsilon_1}(d_1), \cdots, c_{\varepsilon_N}(d_N)) & \mathrm{diag}(\frac{s_{\varepsilon_1}(d_1)}{d_1}, \cdots, \frac{s_{\varepsilon_N}(d_N)}{d_N}) \\ \mathrm{diag}(\varepsilon_1 d_1 s_{\varepsilon_1}(d_1), \cdots, \varepsilon_N d_N s_{\varepsilon_N}(d_N)) & \mathrm{diag}(c_{\varepsilon_1}(d_1), \cdots, c_{\varepsilon_N}(d_N)) \end{bmatrix}. \tag{7.35}$$

An elementary computation yields (7.34) and the lemma follows. □

We now derive basic facts on optimal trajectories.

Lemma 3. *Assume that R is an optimal trajectory associated with a minimising cost V and let h be the constant value of the Hamiltonian defined in* (7.14)*. Then the following facts hold true.*

(*a*) *The cost $\mathscr{C}_\varepsilon(V)$ is negative and hence $\mathrm{Id}_{2N} - R(1)$ is invertible. Moreover, the switching function φ is of class C^2, the matrix $q = RP^T$ defined in* (7.18) *is periodic of period one, it holds that*

$$q^T(1) = \mathscr{C}_\varepsilon(V)\big(\mathrm{Id}_{2N} - R(1)\big)^{-1} R(1) \text{ and } q^T(x) = R(x)q^T(1)R^{-1}(x) \tag{7.36}$$

for every $x \in [0,1]$, and the following relation holds true

$$h = 2M^2 \int_0^1 \|\varphi(x)\| \, dx. \tag{7.37}$$

(*b*) *If $h = 0$ then there exists an invertible $Z_* \in \mathrm{M}(N, \mathbb{R})$ such that*

$$q \equiv \begin{bmatrix} Z_* & 0 \\ 0 & Z_* \end{bmatrix}, \tag{7.38}$$

and $(-1)^N \varepsilon$ is negative.

(*c*) *If $h > 0$, then φ has a finite number of zeroes in $[0,1]$ at which either $\dot\varphi$ does not vanish or $\ddot\varphi$ is well defined and does not vanish.*

Proof. From (7.26) and the expression of V at points where φ does not vanish, one deduces that φ is of class C^2 on $[0,1]$. The one periodicity of q is an immediate consequence of Lemma 2. In that case, one can simplify (7.19) and (7.25) to get (7.36). The latter equation implies that $R(1)$ and $q^T(1)$ commute, which implies by using (7.36) that $q(0) = q(1)$. Since q is solution of a Cauchy problem (the ODE $\dot q = \big[q, \mathscr{A}_V^T\big]$ together with an initial condition), it follows that q is periodic of period one. Finally, integrating (7.28) between $x = 0$ and $x = 1$ and using the periodicity of $\mathrm{tr}(\dot\varphi)$, one gets (7.37). Assume $h = 0$. From (7.37), it follows that $\varphi \equiv 0$ and then (7.26) implies that $\psi \equiv 0$ as well. The rest of the dynamics of q clearly yields that q is constant on $[0,1]$, verifying (7.38). By using the latter fact after taking the determinant in (7.36) it follows that

$$(\det Z_*)^2 = (-1)^N \varepsilon \left[\det\big(\mathrm{Id}_{2N} - R(1)\big)\right]^{2N-1} = \mathscr{C}_\varepsilon(V)^{2N-1},$$

and the last part of Item (*b*) follows. We provide next an argument for Item (*c*). Arguing by contradiction, it would follow that there exists a sequence $(x_k)_{k\in\mathbb{N}}$ of two by two distinct times in $[0,1]$ so that $\lim_{k\to\infty} x_k = \bar x$ and $\varphi(x_k) = 0$ for $k \geq 0$. Since φ is of class C^1, one has that $\varphi(\bar x) = 0$ by continuity of φ and then

$$0 = \lim_{k\to\infty} \frac{\varphi(x_k) - \varphi(\bar x)}{x_k - \bar x} = \dot\varphi(\bar x).$$

Since V is bounded, one deduces from (7.26) that $\ddot\varphi$ is twice differentiable at $x = \bar x$. Moreover, $\ddot\varphi(\bar x)$ is not zero since, from (7.26), it holds

$$\mathrm{tr}(\ddot\varphi(\bar x)) = -2h < 0.$$

By a Taylor expansion at order two, one obtains that there exists an open interval I centered at $\bar x$ so that $\varphi(x) = 0$ for $x \in I$ only if $x = \bar x$. That contradicts the existence of the sequence $(x_k)_{k\in\mathbb{N}}$.□

We end the section by providing preliminary symmetry properties for a minimising potential. For that purpose we define the following matrices of $\mathrm{M}(2N,\mathbb{R})$:

$$J = \mathscr{A}_{\mathrm{Id}_{2N}} \text{ i.e. } J = \begin{bmatrix} 0 & \mathrm{Id}_N \\ \mathrm{Id}_N & 0 \end{bmatrix}, \quad A = \mathscr{A}_{-\mathrm{Id}_{2N}} \text{ i.e. } A = \begin{bmatrix} 0 & \mathrm{Id}_N \\ -\mathrm{Id}_N & 0 \end{bmatrix},$$

$$\mathscr{U} = \begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix}, \text{ for every } U \in \mathrm{SO}(N), \quad \mathscr{B}_Q = \mathscr{A}_{Q^T}^T, \text{ for every } Q \in \mathrm{M}(N,\mathbb{R}).$$

Note that $J^2 = A^T A = \mathrm{Id}_{2N}$.

Proposition 3. *Let $M > 0$, $V \in \mathscr{V}_M$ and R the trajectory of (7.5) associated with V. The following items are equivalent:*

(1.) *V is a minimising potential for $\mathbf{Ext} - \mathbf{Det}_\varepsilon(M)$ along (7.5);*
(2.) *for every $U \in \mathrm{SO}(N)$, $V_{\mathscr{U}} = \mathscr{U}^T V \mathscr{U}$ is a minimising potential for $\mathbf{Ext} - \mathbf{Det}_\varepsilon(M)$ along (7.5) with $\mathscr{U}^T R \mathscr{U}$ as associated trajectory;*
(3.) *V is a minimising potential for $\mathbf{Ext} - \mathbf{Det}_\varepsilon(M)$ along trajectories of each of the following four dynamical systems*

$$\begin{cases} \dot{S}(x) = \mathscr{B}_{V(x)} S(x), \\ S(0) = \mathrm{Id}_N, \end{cases} \quad \begin{cases} \dot{S}(x) = -\mathscr{B}_{V(x)} S(x), \\ S(0) = \mathrm{Id}_N, \end{cases}$$

$$\begin{cases} \dot{S}(x) = S(x)\mathscr{B}_{V^T(x)}, \\ S(0) = \mathrm{Id}_N, \end{cases} \quad \begin{cases} \dot{S}(x) = -S(x)\mathscr{A}_{V(x)}, \\ S(0) = \mathrm{Id}_N, \end{cases}$$

with JRJ, $A^T R A$, R^T and R^{-1} as associated optimal trajectories respectively and same value of the cost;
(4.) *V^T is a minimising potential for $\mathbf{Ext} - \mathbf{Det}_\varepsilon(M)$ along (7.5) with associated trajectory $A^T (R^T)^{-1} A$.*

Proof. Showing the several items is immediate once one notices that

$$J\mathscr{A}_Q J = \mathscr{B}_Q, \quad A^T \mathscr{A}_Q A = -\mathscr{B}_Q, \text{ for every } Q \in M_N(\mathbb{R}).$$

As for the equality of the costs, we just check the following

$$\det(\mathrm{Id}_{2N} - R^{-1}(1)) = \det\Big((R(1) - \mathrm{Id}_{2N}) R^{-1}(1)\Big) = \det(\mathrm{Id}_{2N} - R(1)).$$

□

7.4 One-Dimensional Case

From now on $N = 1$, M is still a positive number and

$$\mathscr{V}_M = \{V : [0,1] \to \mathbb{R} \mid V \text{ measurable and } \operatorname*{ess\,sup}_{x\in[0,1]} |V(x)| \le M^2\}. \tag{7.39}$$

From Item (*iii*) of Proposition 2, it holds that $V(x) \in \{-M^2, M^2\}$ as soon as $\varphi(x) \neq 0$ and this motivates the following definition.

Definition 2. *Let R be a trajectory of* (7.5) *associated to some $V \in \mathscr{V}_M$. A* bang arc $\gamma : I \to \mathrm{M}(2,\mathbb{R})$ *is a piece of R defined on some non empty subinterval $I \subset [0,1]$ such that $V = \nu M^2$ is constant on I, with $\nu \in \{-1,1\}$. A trajectory R of* (7.5) *is said to be* bang *if it is made of a unique bang arc and* bang-bang *if it is the concatenation of bang arcs.*

We first examine the **Max-Det** problem. In dimension $N = 1$, the cost to maximise is

$$\begin{aligned}\mathscr{C}_V &= -\det(I_2 - R(1)),\\ &= -(1 - \operatorname{tr} R(1) + \det R(1)),\\ &= \operatorname{tr} R(1) - 2\end{aligned}$$

since the monodromy $R(1)$ belongs to $\mathrm{SL}(2,\mathbb{R})$. Maximising $\mathscr{C}_V$ is so equivalent to maximising the trace of the monodromy

$$\operatorname{tr} R(1) = z(1) + y'(1),$$

where z and y satisfy $-w'' + V(x)w = 0$ with respective initial conditions $(z(0), z'(0)) = (1,0)$ and $(y(0), y'(0)) = (0,1)$.

Proposition 4. *Let V_1 and V_2 be two potentials in $L^1_{\mathrm{loc}}(\mathbb{R}_+)$, $V_1 \geq |V_2|$ a.e., and let y_1 and y_2 satisfy $-y_i'' + V_i(x)y_i = 0$, $i = 1,2$. If $y_1(0) \geq |y_2(0)|$ and $y_1'(0) \geq |y_2'(0)|$, then $y_1(x) \geq |y_2(x)|$ and $y_1'(x) \geq |y_2'(x)|$ for all $x \geq 0$.*

Proof. (i) First assume V_1 and V_2 constant, $V_1 \equiv A$ and $V_2 \equiv B$ with A and B two reals such that $A \geq |B|$. One has

$$y_1(x) = y_1(0)\cosh(\alpha x) + x y_1'(0)\operatorname{sinhc}(\alpha x)$$

where $\alpha = \sqrt{A}$, and where we denote

$$\operatorname{sinhc}(x) = \begin{cases} \sinh(x)/x & \text{if } x \neq 0,\\ 1 & \text{if } x = 0.\end{cases}$$

If B is nonnegative, let $\beta := \sqrt{B} \leq \alpha$; one has

$$\begin{aligned}|y_2(x)| &= |y_2(0)\cosh(\beta x) + x y_2'(0)\operatorname{sinhc}(\beta x)|\\ &\leq |y_2(0)|\cosh(\beta x) + x|y_2'(0)|\operatorname{sinhc}(\beta x)\\ &\leq y_1(0)\cosh(\alpha x) + x y_1'(0)\operatorname{sinhc}(\alpha x) = y_1(x)\end{aligned}$$

for $x \geq 0$ since both cosh and sinhc are nondecreasing functions on $\mathbb{R}_+$ (and $\beta \leq \alpha$). Similarly, for $x \geq 0$,

$$\begin{aligned}|y_2'(x)| &= |\beta y_2(0)\sinh(\beta x) + y_2'(0)\cosh(\beta x)|\\ &\leq \alpha y_1(0)\sinh(\alpha x) + y_1'(0)\cosh(\alpha x) = y_1'(x).\end{aligned}$$

If B is negative, let $\beta := \sqrt{-B} \leq \alpha$; one has (denoting $\operatorname{sinc}(x) = \sin(x)/x$ if $x \neq 0$, $\operatorname{sinc}(0) = 1$)

$$\begin{aligned}|y_2(x)| &= |y_2(0)\cos(\beta x) + x y_2'(0)\operatorname{sinc}(\beta x)|\\ &\leq |y_2(0)|\cosh(\beta x) + x|y_2'(0)|\operatorname{sinhc}(\beta x)\\ &\leq y_1(x)\end{aligned}$$

for $x \geq 0$ since $|\cos| \leq \cosh$ and $|\operatorname{sinc}| \leq \operatorname{sinhc}$ on $\mathbb{R}_+$. Similarly, for $x \geq 0$,

$$\begin{aligned} |y_2'(x)| &= |-\beta y_2(0)\sin(\beta x) + y_2'(0)\cos(\beta x)| \\ &\leq \alpha y_1(0)\sinh(\alpha x) + y_1'(0)\cosh(\alpha x). \end{aligned}$$

(ii) Take now some positive x, and assume V_1 and V_2 are piecewise constant on $[0, x]$; there exists a common subdivision $0 = x_0 < x_1 < ... < x_N = x$, $N \geq 1$, such that on every $[x_i, x_{i+1}[$ both V_1 and V_2 are constant, with $V_1 \geq |V_2|$. A simple recurrence using step (i) allows to conclude that $y_1(x) \geq |y_2(x)|$ and $y_1'(x) \geq |y_2'(x)|$.

(iii) Consider eventually V_1 and V_2 locally integrable on $\mathbb{R}_+$, and fix $x > 0$. There exist two sequences $(V_{1,n})_n$, $(V_{2,n})_n$ of piecewise constant functions converging in $L^1(0, x)$ to V_1 and V_2, respectively. These sequences can be chosen such that $V_{1,n} \geq |V_{2,n}|$, $n \in \mathbb{N}$. Then according to point (ii), for all $n \in \mathbb{N}$, $y_{1,n}(x) \geq |y_{2,n}(x)|$ and $y_{1,n}'(x) \geq |y_{2,n}'(x)|$, where $y_{i,n}$ denotes the solution associated with $V_{i,n}$ and fixed initial conditions $(y_i(0), y_i'(0))$, $i = 1, 2$. Since, for any given initial condition (y_0, y_0'), the mapping $V \mapsto (y(x), y'(x))$ (where y is the solution of $-y'' + Vy = 0$, $y(0) = y_0$, $y'(0) = y_0'$) is continuous from $L^1(0, x)$ to $\mathbb{R}^2$ (see, *e.g.*, Proposition 7 in [1]), passing to the limit one obtains that $y_1(x) \geq |y_2(x)|$ and $y_1'(x) \geq |y_2'(x)|$. As x is arbitrary, the desired conclusion holds. □

Corollary 1. *For V in $L^\infty(0, 1)$, let y and z denote the solutions of*

$$-y'' + V(x)y = 0, \quad y(0) = 0, \quad y'(0) = 1,$$

$$-z'' + V(x)z = 0, \quad z(0) = 1, \quad z'(0) = 0.$$

Then, for any positive bound M, the constant potential $V \equiv M^2$ is the unique function maximising both $y(1)$, $y'(1)$, $z(1)$ and $z'(1)$ over essentially bounded potentials such that $\|V\|_\infty \leq M^2$.

Theorem 1. *The unique solution of the* **Max-Det(M)** *problem in the periodic case is the constant potential equal to M^2.*

Proof. It is clear from the previous corollary that the constant potential $V \equiv M^2$ maximises $z(1) + y'(1)$ among essentially bounded potentials such that $\|V\|_\infty \leq M^2$. Let V be a measurable function satisfying the same bound and such that V is strictly inferior to M^2 on a positive measure subset of $[0, 1]$; a direct estimation allows to prove that the associated values of both $z(1)$ and $y'(1)$ (hence of their sum) are strictly smaller than the values obtained for the constant potential $V \equiv M^2$. □

We eventually handle **Min-Det**. In particular, we immediately derive the following result after Lemmas 3 and 2.

Lemma 4. *Assume that R is an optimal trajectory associated with a potential V minimising $\mathscr{C}_1$. Then the following cases may occur.*

(*i*) *If $h = 0$, then V is equal to the constant potential $V_0 \equiv -M^2$ and φ never vanishes on on $[0, 1]$. In that case, the minimal cost is equal to $\mathscr{C}_1(V_0) = -2\big(1 - c_-(M)\big)$;*

(*ii*) *if $h \neq 0$, then φ has a finite number of zeroes in $[0, 1]$ and $V(x) = M^2\operatorname{sgn}(\varphi(x))$ outside a finite set made of the zeroes of φ.*

Hence, either R is the bang trajectory R_0 associated with V_0 or it is a bang-bang trajectory with a finite number of bang arcs.

Proof. From Lemma 2, we know that the minimal value of $\mathscr{C}_1$ is negative, and then, Item (a) of Lemma 3 only leaves the possibility of φ never vanishing on $[0,1]$. Hence V is constant equal M or $-M$. Since $C_1(M) > 0$, Item (i) follows at once. Item (ii) is essentially a rewriting of Item (b) of Lemma 3 together with Item (iii) of Proposition 2. □

In the one-dimensional case, we can actually give a more elementary proof that minimising potentials are bang-bang with finitely many switchings using the structure of $\mathfrak{sl}(2,\mathbb{R})$. Our minimisation problem is a Mayer problem with linear cost, $\operatorname{tr} R(1) \to \min$, and bilinear dynamics

$$\dot{R}(x) = F_0 R(x) + V(x) F_1 R(x)$$

with a single input control such that, a.e., $|V(x)| \leq M^2$, and matrices (linear vector fields)

$$F_0 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad F_1 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}.$$

Together with their commutator[8]

$$F_{01} := [F_0, F_1] = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},$$

these matrices form an $\mathfrak{sl}_2$-*triple* of the dimension three Lie algebra. In particular, one has

$$F_{001} = [F_0, F_{01}] = -2F_0, \quad F_{101} = [F_1, F_{01}] = 2F_1. \tag{7.40}$$

Denoting $H_i := \langle P, F_i R\rangle$, for $i = 0, 1$, the Hamiltonian lifts of F_0 and F_1, the Hamiltonian is $H = H_0 + VH_1$. To analyse the structure of the set of zeroes of H_1 along an extremal, one can compute (with the same notation as before)

$$\dot{H}_1 = H_{01}, \quad \dot{H}_{01} = H_{001} + VH_{101}.$$

Because of (7.40), $\ddot{H}_1 = 2(VH_1 - H_0)$ so H_1 is $\mathscr{C}^2$ (since V is bounded, VH_1 vanishes whenever H_1 does) and there are two cases at a switching time: either H_{01} is not zero, or H_{01} is zero and H_{001} is not (P would otherwise vanish, which is forbidden, since F_0, F_1 and F_{01} form a basis of the Lie algebra). In both cases, the switching time must be isolated.

We focus now on strong extremals associated with $h \neq 0$, and introduce the following notations: if $\nu^2 = 1$, we use $c_\nu(t)$ (respectively $s_\nu(t)$) to denote $\cosh(t)$ if $\nu = 1$ and $\cos(t)$ if $\nu = -1$ (respectively $\sinh(t)$ if $\nu = 1$ and $\sin(t)$ if $\nu = -1$). With these conventions, one also has for every $x \in \mathbb{R}$ that

$$c_\nu^2(x) - \nu s_\nu^2(x) = 1, \ \dot{c}_\nu(x) = \nu s_\nu(x), \ \dot{s}_\nu(x) = c_\nu(x), \tag{7.41}$$

$$c_\nu(2x) = 1 + 2\nu s_\nu^2(x), \ s_\nu(2x) = 2\nu s_\nu(x) c_\nu(x). \tag{7.42}$$

As a consequence, if d is a positive real number, the solution of the linear second order equation $\ddot{y} = \nu d y$ is given by

$$y(t) = c_\nu(dt) y(0) + \frac{1}{d} s_\nu(dt) \dot{y}(0), \quad t \in \mathbb{R}. \tag{7.43}$$

We have the following two intermediate results.

[8] Note that we use the matrix commutator whose sign is opposite to the Lie bracket of the associated linear vector fields.

Lemma 5. *Let (R, q) be a strong extremal projecting on an optimal trajectory R which is associated with a potential V minimising $\mathscr{C}_1$ with corresponding $h \neq 0$. Assume furthermore that*

1. *V is not identically equal to $-M^2$;*
2. *$x_0 < x_1$ are two consecutive zeroes of φ in $[0,1]$,* i.e., *$|\varphi| > 0$ on (x_0, x_1).*

Set $T := x_1 - x_0 > 0$ and $\nu = \operatorname{sgn}(\varphi)$ on (x_0, x_1). Then both $c_\nu(MT)$ and $s_\nu(MT)$ are non zero and the following holds:

$$\varphi(x) = \frac{h}{M^2 c_\nu(MT)} s_\nu(M(x - x_0)) s_\nu(M(x_1 - x)), \text{ for } x \in [x_0, x_1]. \tag{7.44}$$

In particular,

$$\dot\varphi(x_0) = -\dot\varphi(x_1) = h\frac{s_\nu(MT)}{c_\nu(MT)} \neq 0. \tag{7.45}$$

Proof. In the case $N = 1$ and using the notations of the lemma, one can rewrite (7.26) as

$$\ddot\varphi = 4\nu M^2(\varphi - \frac{\nu h}{2M^2}) \text{ for } x \in [x_0, x_1]. \tag{7.46}$$

Integrating (7.46) yields that

$$\varphi(x) = \frac{\nu h}{2M^2}\big(1 - c_\nu(2M(x - x_0))\big) + B s_\nu(2M(x - x_0)), \tag{7.47}$$

$$\dot\varphi(x) = 2M^2 B c_\nu(2M(x - x_0)) - h s_\nu(2M(x - x_0)), \tag{7.48}$$

where B is a constant satisfying

$$-\frac{\nu h}{2M^2}(1 - c_\nu(2MT)) = B s_\nu(2MT). \tag{7.49}$$

From (7.48), one deduces that

$$\dot\varphi(x_0) = 2M^2 B, \quad \dot\varphi(x_1) = 2M^2 B c_\nu(2MT) - h s_\nu(2MT). \tag{7.50}$$

We prove next that $s_\nu(MT) \neq 0$. Arguing by contradiction, it would first imply that $\nu = -1$ and then $V = -M^2$, $c_\nu(2MT) = 1$, $s_\nu(2MT) = 0$ and, from (7.50), that $\dot\varphi(x_0) = \dot\varphi(x_1) = 2M^2 B$. If $B \neq 0$, then $\operatorname{sgn}(B)\dot\varphi$ is positive in a right neighborhood of x_0 while it is negative in a left neighborhood of x_1, implying that φ must vanish inside (x_0, x_1). This contradicts Item 2., and therefore one deduces that $B = 0$ and then $\ddot\varphi(x_0) = \ddot\varphi(x_1) = -2h$, yielding that $h > 0$ and x_0 and x_1 are not switching times. We claim that every zero of φ is not a switching time and that $V \equiv -M^2$. Indeed, recall that a zero of φ is isolated and there are a finite number of them. Consider then x_2 distinct from x_0 and x_1. Assume that it is consecutive to x_1, *i.e.* $|\varphi| > 0$ on (x_1, x_2). Reproducing the reasoning done on $[x_0, x_1]$ with x_1 (respectively x_2) replacing x_0 (respectively x_1), we conclude from (7.50) that the corresponding B is equal to zero and from (7.47) that $c_{\nu'}(2M(x_2 - x_1)) = 1$, *i.e.*, $\nu' = -1$ and $s_{\nu'}(M(x_2 - x_1)) = 0$. Being back to the previous situation, one deduces that $\dot\varphi(x_2) = 0$. Proceeding in that way step by step, one gets the claim. This contradicts Item 1. and finally one has proved that $s_\nu(MT) \neq 0$. From (7.49) and (7.42), one gets that

$$B = \frac{h}{2M^2}\frac{c_\nu(MT)}{s_\nu(MT)},$$

and direct computations finally yield (7.44) and (7.46). □

To state our subsequent results, one needs to define, for every positive real number M the function $F_M : [0,1] \to \mathbb{R}_+$ by

$$F_M(x) = x + \frac{\pi - \arctan\big(\tanh(Mx)\big)}{M}. \tag{7.51}$$

The basic facts on this function are the following:

$$F_M(0) = \frac{\pi}{M},\ F_M(1) = 1 + \frac{\pi - \arctan\big(\tanh(M)\big)}{M}, F'_M(x) = \frac{2\tanh^2(Mx)}{1+\tanh^2(Mx)}, \tag{7.52}$$

for all $x \in [0,1]$. Hence F_M is a C^1, strictly increasing bijection from $[0,1]$ to $[\frac{\pi}{M}, F_M(1)]$ and $F_M(1) > 1$. Our second intermediate result goes as follows.

Lemma 6. *Let (R,q) be a strong extremal projecting on an optimal trajectory R which is associated with a potential V minimising $\mathscr{C}_1$ with corresponding $h \neq 0$. Assume furthermore that R is not a bang trajectory. Then, up to a translation, V is periodic of period T_1+T_2 so that $V = M^2$ on $[0,T_1]$ and $V = -M^2$ on $[T_1, T_1+T_2]$ where $T_1, T_2 \in (0,1)$ so that they satisfy*

$$T_2 = \frac{\pi - \arctan\big(\tanh(MT_1)\big)}{M}, \tag{7.53}$$

and there exists a positive integer l such that

$$F_M(T_1) = 1/l. \tag{7.54}$$

Proof. Notice that R must have at least two distinct bang arcs and then at least two switching points. Moreover, all the zeroes of φ must be switching times according to (7.45). Thanks to Lemma 1, we can assume, up to translating the potential V, that 0 is a switching time and $\varphi > 0$ in a right neighborhood of zero (since both signs are taken on $[0,1]$). Since $\dot\varphi(0) \neq 0$, it must be positive and (7.45) yields that both h and ν are positive. We first claim that $x = 1$ must be a switching time. For otherwise, $\varphi(1) \neq 0$ and hence V has a constant sign in a left neighborhood of 1. If $V = M^2$ there, then for $a > 0$ small enough one has that $\varphi_{-a}(a) = \varphi(0) = 0$ and $\dot\varphi_{-a}(a) = \dot\varphi(0) \neq 0$, *i.e.*, a is a switching time for V_{-a}. This is in contradiction with the fact that $V_{-a} = M$ in an open neighborhood of a. If now $V = -M^2$ in a left neighborhood of 1, let $x_r < 1$ be the largest zero of φ in $[0,1]$. It turns out that V_{x_r} changes sign at $x = 1 - x_r$ but this is in contradiction with the fact that $\varphi_{x_r}(1-x_r) = \varphi(1) \neq 0$. We have proved the claim. Now we show that the last bang must correspond to $V = -M^2$. Indeed if it were not the case, then $V_a = M^2$ in an open neighborhood of some $a > 0$ small enough with $\varphi_a(a) = 0$, which is not possible. It means that R is the concatenation of an even number of bang arcs, γ_i, $1 \leq i \leq 2l$, where on the γ_{2j-1}'s, $1 \leq j \leq l$, one has $V = M^2$ and on the γ_{2j}'s, $1 \leq j \leq l$, one has $V = -M^2$. Let $T_i > 0$ be the duration of each bang arc γ_i, for $1 \leq i \leq 2l$, and clearly

$$\sum_{i=1}^{2l} T_i = 1. \tag{7.55}$$

We next prove that $T_2 = F(T_1)$. Indeed, consider (7.45) written for $(x_0, x_1) = (0, T_1)$ and then $(x_0, x_1) = (T_1, T_1 + T_2)$. One deduces that

$$h\tanh(MT_1) = \dot\varphi(0) = -\dot\varphi(T_1), \quad h\tan(MT_2) = \dot\varphi(T_1) = -\dot\varphi(T_1+T_2). \tag{7.56}$$

It follows at once that

$$\tanh(MT_1) = -\tan(MT_2) \in (0,1).$$

It follows that $MT_2 - k\pi \in (\frac{3\pi}{4}, \pi)$ for some non negative integer k. Then $k = 0$ otherwise, using (7.44), φ would have another zero in (T_1, T_1+T_2), which is not possible. One deduces (7.53). We finally prove that

$$T_{2j-1} = T_1, \quad T_{2j} = T_2, \text{ for } 1 \le j \le l. \tag{7.57}$$

We only provide an argument for $T_3 = T_1$ since the other equalities are deduced in an identical manner. For that purpose, consider (7.45) written for $(x_0, x_1) = (T_1+T_2, T_1+T_2+T_3)$. One deduces that

$$h\tanh(MT_3) = \dot\varphi(T_1+T_2) = -\dot\varphi(T_1+T_2+T_3).$$

Using (7.56), one gets that

$$\tanh(MT_3) = \frac{\dot\varphi(T_1+T_2)}{h} = -\tan(MT_2) = \tanh(MT_1),$$

yielding that $T_1 = T_3$ and V is (T_1+T_2)-periodic. One deduces (7.54) from (7.55), which concludes the proof of Lemma 6. □

We are able to state the proposition providing a complete solution to **Min-Det** in the case $N = 1$.

Theorem 2. *For every positive M, the optimal control problem* **Min-Det(M)** *admits a unique minimising potential V_{min} in $L^\infty(\mathbb{S}^1)$ defined as follows.*

(*a*) *If $M \in (0,\pi]$, $V_{min} = V_0 \equiv -M^2$ and the minimal value for* **Min-Det(M)** *is equal to $\mathscr{C}_1(V_0) = -2\big(1 - c_-(M)\big)$;*

(*b*) *If $M > \pi$, V_{min} is the potential V_1 equal to M^2 on $[0,t_1]$ and $-M^2$ on $[t_1,1]$, with $F_M(t_1) = 1$ and the minimal value for* **Min-Det(M)** *is equal to $\mathscr{C}_1(V_1) = -2\big(1 - c_-(M(1-t_1))c_+(t_1)\big)$.*

Proof. If $M \le \pi$, then $F_M(x) > 1$ for every $x \in (0,1]$ and one deduces from (7.54) that there is no $T_1 \in (0,1)$ satisfying the properties required for the existence of a an optimal trajectory R which is not a bang trajectory. Therefore, the only candidate left as minimising potential by Lemma 4 is $V = V_0$, *i.e.* Item (*a*) holds true. Assume now that $M > \pi$. Define the positive integer $L := E(\frac{M}{\pi})$ (where $E(x)$ stands for the integer part of the real x), and the $2L$ times

$$t_l = F_M^{-1}(1/l), \ s_l = 1/l - t_l, \quad 1 \le l \le L. \tag{7.58}$$

According to Lemma 6, there exists a bang-bang trajectory R_l with $2l$ bang arcs and associated with the periodic potential V_l of period $1/l$ so that $V_l = M^2$ on $[0,t_l]$ and $V_l = -M^2$ on $[t_l, t_l+s_l]$. Recall that R_0 is the trajectory of (7.5) associated with V_0. Then, one gets from Lemmas 4 and 6 that a minimising potential V_{min} must be equal to V_l for some integer $0 \le l \le L$. In order to conclude, one is left with the computation of the costs $\mathscr{C}_1(V_l)$, for positive integers $1 \le l \le L$. A lengthy but straightforward computation yields that

$$R_l(1/l) = \begin{bmatrix} c_-(Ms_l) & \frac{s_-(Ms_l)}{M} \\ -Ms_-(Ms_l) & c_-(Ms_l) \end{bmatrix} \begin{bmatrix} c_+(Mt_l) & \frac{s_+(Mt_l)}{M} \\ Ms_+(Mt_l) & c_+(Mt_l) \end{bmatrix}$$

$$= \begin{bmatrix} c_-(Ms_l)c_+(Mt_l) + s_-(Ms_l)s_+(Mt_l) & \frac{c_-(Ms_l)s_+(Mt_l)+s_-(Ms_l)c_+(Mt_l)}{M} \\ M(s_-(Ms_l)c_+(Mt_l) + c_-(Ms_l)s_+(Mt_l)) & c_-(Ms_l)c_+(Mt_l) - s_-(Ms_l)s_+(Mt_l) \end{bmatrix}, \tag{7.59}$$

and one has that $\det(R_l(1/l)) = 1$ and

$$\alpha_l = -\frac{\mathrm{tr}(R_l(1/l))}{2} = -c_-(Ms_l)c_+(Mt_l),\ 1 \leq l \leq L. \tag{7.60}$$

We use $r_l, \frac{1}{r_l}$ in $\mathbb{C}$ to denote the eigenvalues of $R_l(1/l)$. Since V_l is $1/l$-periodic, one gets that $R_l(1) = R_l^l(1/l)$ and hence

$$\mathscr{C}_1(V_l) = -\det\big(\mathrm{Id}_2 - R_l^l(1/l)\big) = (-2)(1 - \frac{r_l^l + r_l^{-l}}{2}),\ 1 \leq l \leq L. \tag{7.61}$$

Recall that $Ms_l \in (\frac{3\pi}{4}, \pi)$ and hence, it holds, for $1 \leq l \leq L$ that

$$\begin{aligned} -c_-(Ms_l) &= -c_-\Big(\pi - \arctan\big(\tanh(Mt_l)\big)\Big) = c_-\Big(\arctan\big(\tanh(Mt_l)\big)\Big) \\ &= \frac{1}{\sqrt{1+\tanh^2(Mt_l)}} = \frac{c_+(Mt_l)}{\sqrt{c_+^2(Mt_l) + s_+^2(MT_1)}}, \end{aligned}$$

and then

$$\alpha_l = \frac{c_+^2(Mt_l)}{\sqrt{2c_+^2(Mt_l) - 1}} > 1. \tag{7.62}$$

Let $\xi_l > 0$ such that $\alpha_l = c_+(\xi_l)$. Since r_l and $\frac{1}{r_l}$ are the roots of the degree two polynomial $X^2 + 2c_+(\xi_l)X + 1$, one gets that $r_l = -e^{\xi_l}$ and finally it holds

$$\mathscr{C}_1(V_l) = (-2)\Big(1 - (-1)^l c_+(l\xi_l)\Big).$$

For even l's, the cost is non negative, implying that V_l cannot be minimising. For odd l's, the cost is smaller than -4 and then smaller than $\mathscr{C}_1(V_0)$. It remains to show that $\mathscr{C}_1(V_l)$ reaches its minimal value for $l = 1$. For that, it is enough to prove that the mapping $G : l \mapsto l\xi_l$ is strictly decreasing for $l \in [1, L]$. Computing, one gets

$$G'(l) = Mt_l\Big(\frac{\xi_l}{Mt_l} - \frac{c_+(Mt_l)}{s_+(Mt_l)}\frac{F_M(Mt_l)}{t_l}\Big), \quad l \in [1, L].$$

Since $F_M(Mt_l) > t_l$, one would have that $G'(l) < 0$ if one shows that $\xi_l < Mt_l$. In turn, that last inequality is itself equivalent $\alpha_l < c_+(Mt_l)$, inequality which does hold true by (7.62). This concludes the proof of Theorem 2. □

References

1. Aldana, C. L.; Caillau, J.-B.; Freitas, P. Maximal determinants of Schrödinger operators on bounded intervals. *J. Éc. polytech. Math.*, 7:803–829, 2020.

2. Bonnard, B.; Jurdjevic, V.; Kupka, I.; Sallet, G. Systèmes de champs de vecteurs transitifs sur les groupes de lie semi-simples et leurs espaces homogènes. *Astérisque*, 75-76:19–45, 1980.
3. Burghelea, D.; Friedlander, L.; Kappeler, T. On the determinant of elliptic differential and finite difference operators in vector bundles over S^1. *Comm. Math. Phys.*, 138(1):1–18, 1991.
4. Forman, R. Determinants, finite-difference operators and boundary value problems. *Comm. Math. Phys.*, 147(3):485–526, 1992.
5. Helmke, U. Isospectral flows on symmetric matrices and the riccati equation. *Systems Control Lett.*, 16:159–165, 1991.
6. Minakshisundaram, S.; Pleijel, A. Some properties of the eigenfucntions of the laplace operator on riemannian manifolds. *Canad. J. Math.*, 1:242–256, 1949.
7. Ray, D. B.; Singer, I. M. R-torsion and the laplacian on riemannian manifolds. *Adv. Math.*, 7:145–210, 1971.
8. Seeley, R. T. The resolvent of an elliptic boundary value problem. *Amer. J. Math.*, 91:889–928, 1969.

8

A Note on Reversible Mappings and Folds: A Local Approach

Otávio M. L. Gomide[1] and Marco A. Teixeira[2]

[1] Departamento de Matemática, Instituto de Matemática e Estatística (IME), Universidade Federal de Goiás (UFG), 74690-900, Goiânia, GO, Brazil. otaviomarcal@ufg.br

[2] Departamento de Matemática, Instituto de Matemática, Estatística e Computação Científica (IMECC), Universidade Estadual de Campinas (UNICAMP), Rua Sérgio Buarque de Holanda, 651, Cidade Universitária Zeferino Vaz, 13083-859, Campinas, SP, Brazil. teixeira@ime.unicamp.br

Summary. Sometimes, the best tool to analyze the qualitative behavior of a vector field in $\mathbb{R}^n$ is to consider the Poincaré return map to an $(n-1)$-dimensional section, $n \geq 2$. In this article we develop a local qualitative analysis of a first return transformation when it is a reversible mapping characterized by the composition of two involutions, each of them having an $(n-1)$-dimensional fixed-point set. Some structural stability results are provided as well as applications to nonsmooth dynamical systems and diagram of mappings. In this way, stability conditions of a fold-fold singularity in nonsmooth dynamical systems in dimension greater than 3 are discussed. It is worth to say that such subject is still poorly understood in higher dimension.

8.1 Introduction

8.1.1 Motivation

Some problems in control theory, economics and nonlinear oscillations lead to consideration of differential equations whose right-hand sides are defined by smooth different mappings giving rise to nonsmooth systems. Many industrial applications, for instance in mechanical and electromechanical systems are reported, see for instance [4], [9], [20], [19] and [26]. It is worthwhile to cite the work of Ekeland (see [10]) and Klok (see [15]), where the main problem in the classical calculus of variations was carried out to study discontinuous Hamiltonian vector fields.

Singularities of Nonsmooth Dynamical Systems (in short NSDS) theory was mainly well developed for dimensions 2 and 3, in establishing persistence results. In this field the most interesting persistent phenomenon, lies in the fold-fold singularity, with focus on the so called T-singularity. The most important tool in this setting is the behavior of a reversible mapping that works as a first return map associated to the system. Persistence results were firstly developed in [22], and complemented in [6, 7, 8, 12, 24], where is also provided a discussion on current directions of research involving typical singularities of $3D$ NSDS. We could extend this technique to many other settings, like diagram of mappings in higher dimensions than 2.

A 3-dimensional fold-fold singularity is an intriguing phenomenon that has no counterparts in smooth systems, and the complete characterization of the local structural stability of a $3D$-nonsmooth system around an elliptic fold-fold singularity has been an open problem over the last 30 years (see for instance [6, 8, 12, 22, 24]. The methods employed in [12, 24] lead us to the completely

mathematical understanding of the dynamics around a 3-dimensional T-singularity. This approach leads itself to applications to generic bifurcation theory. It is worth to say that such subject is still poorly understood in higher dimension. In this paper it is presented procedures to deal with stability problems in higher dimensions, involving the T-singularity and diagram of mappings. In short, the T-singularity provides a central phenomenon in the NSDS theory.

Our main goal is to present local stability results of n-dimensional systems around a T-singularity with $n > 3$. In addition we will briefly present sufficient conditions for topological stability inside a class of diagram of fold mappings. In fact, such results are immediate consequence of the techniques and mechanisms employed to prove the main theorem (Theorem 1 below).

8.2 Preliminaries

We summarize a rough overall description of a few basic concepts and results in order to set the problem in question. Recall that, NSDS are generated by vector fields, locally given as systems of ordinary differential equation with nonsmooth right-hand side. For background information see [12].

8.2.1 Filippov System, Fold-Fold singularity, Diagram of Mappings

Let M be a small neighborhood of the origin in $\mathbb{R}^{n+1}$ with compact closure and let $h : (M, 0) \to (\mathbb{R}, 0)$ be a smooth function having 0 as a regular value, therefore $\Sigma = h^{-1}(0)$ is a compact embedded codimension one submanifold of M, known as **switching manifold**, which splits it into the sets $M^{\pm} = \{p \in M; \pm h(p) > 0\}$.

A **germ of vector field** of class $\mathscr{C}^r$ at a compact set $\Lambda \subset M$ is an equivalence class $\widetilde{X}$ of $\mathscr{C}^r$ vector fields defined in a neighborhood of Λ. More specifically, two $\mathscr{C}^r$ vector fields X_1 and X_2 are in the same equivalence class if:

1. X_1 and X_2 are defined in neighborhoods U_1 and U_2 of Λ in M, respectively;
2. There exists a neighborhood U_3 of Λ in M such that $U_3 \subset U_1 \cap U_2$;
3. $X_1|_{U_3} = X_2|_{U_3}$.

In this case, if X is an element of the equivalence class $\widetilde{X}$, then X is said to be a representative of $\widetilde{X}$. The set of germs of vector fields of class $\mathscr{C}^r$ at Λ will be denoted by $\chi^r(\Lambda)$, or simply χ^r. In what follows, we make no distinction between a germ of a mapping and any one of its representatives.

Remark 1. If Λ is a point, then we have the usual local approach.

Analogously, a **germ of piecewise smooth vector field** of class $\mathscr{C}^r$ at a compact set $\Lambda \subset M$ is an equivalence class $\widetilde{Z} = (\widetilde{X}, \widetilde{Y})$ of pairwise $\mathscr{C}^r$ vector fields defined as follows: $Z_1 = (X_1, Y_1)$ and $Z_2 = (X_2, Y_2)$ are in the same equivalence class if, and only if,

1. X_i and Y_i are defined in neighborhoods U_i and V_i of Λ in M, respectively, $i = 1, 2$.
2. There exist neighborhoods U_3 and V_3 of Λ in M such that $U_3 \subset U_1 \cap U_2$ and $V_3 \subset V_1 \cap V_2$.
3. $X_1|_{U_3 \cap \overline{M^+}} = X_2|_{U_3 \cap \overline{M^+}}$ and $Y_1|_{V_3 \cap \overline{M^-}} = Y_2|_{V_3 \cap \overline{M^-}}$.

In this case, if $Z = (X, Y)$ is an element of the equivalence class $\widetilde{Z}$, then Z is said to be a representative of $\widetilde{Z}$. The set of germs of piecewise smooth vector fields of class $\mathscr{C}^r$ at Λ will be denoted by $\Omega^r(\Lambda)$, or simply Ω^r.

We endow χ^r with the C^r topology and consider $\Omega^r = \chi^r \times \chi^r$ with the product topology and $r \geq 2$.

The classical singularities of X and Y in $\overline{M^+}$ and $\overline{M^-}$, respectively, are naturally singularities of the PSVF $Z = (X, Y)$, nevertheless, the contact of X and Y with Σ gives rise to new singularities in this context. In order to analyze such contact we consider the **Lie derivative** $Xh(p)$ of h in the direction of the vector field $X \in \chi^r$ at $p \in \Sigma$, which is defined as $Xh(p) = \langle X(p), \nabla h(p) \rangle$. In this way, the **tangency set** between X and Σ is given by $S_X = \{p \in \Sigma;\ Xh(p) = 0\}$.

For a PSVF $Z = (X, Y)$ the switching manifold Σ is generically the closure of the union of the following three distinct open regions:

- Crossing Region: $\Sigma^c(Z) = \{p \in \Sigma;\ Xf(p)Yf(p) > 0\}$.
- Stable Sliding Region: $\Sigma^{ss}(Z) = \{p \in \Sigma;\ Xf(p) < 0,\ Yf(p) > 0\}$.
- Unstable Sliding Region: $\Sigma^{us}(Z) = \{p \in \Sigma;\ Xf(p) > 0,\ Yf(p) < 0\}$.

The tangency set of Z will be referred as $S_Z = S_X \cup S_Y$. Notice that Σ is the disjoint union $\Sigma^c \cup \Sigma^{ss} \cup \Sigma^{us} \cup S_Z$. Herein, $\Sigma^s = \Sigma^{ss} \cup \Sigma^{us}$ is called **sliding region** of Z. See Figure 8.1.

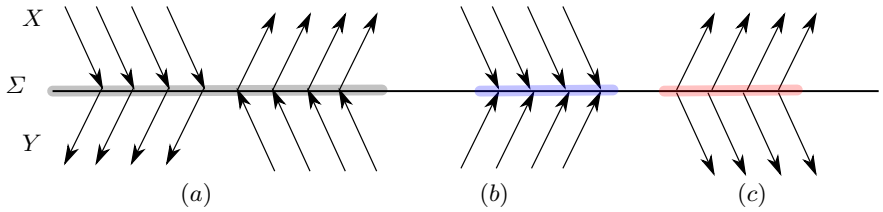


Fig. 8.1: Regions in Σ: Σ^c in (a), Σ^{ss} in (b) and Σ^{us} in (c).

The concept of solution of Z follows the Filippov's convention (see, for instance, [11, 14, 23]). The local solution of $Z = (X, Y) \in \Omega^r$ at $p \in \Sigma^s$ is given by the **sliding vector field**

$$F_Z(p) = \frac{1}{Yh(p) - Xh(p)} \left(Yh(p)X(p) - Xh(p)Y(p) \right). \tag{8.1}$$

Notice that F_Z is a $\mathscr{C}^r$ vector field tangent to Σ^s. The critical points of F_Z in Σ^s are called **pseudo-equilibria** of Z. Sometimes, it is useful to cosider the **normalized sliding vector field** F_Z^N of Z which is given by

$$F_Z^N(p) = Yh(p)X(p) - Xh(p)Y(p), \tag{8.2}$$

for every $p \in \Sigma^s$, since the normalized sliding vector field can be $\mathscr{C}^r$ extended beyond the boundary of Σ^s and, in addition, if R is a connected component of Σ^{ss}, then F_Z^N is a re-parameterization of F_Z in R, and so the phase portraits of both coincide. If R is a connected component of Σ^{us}, then F_Z^N is a (negative) re-parameterization of F_Z in R, then they have the same phase portrait, but the orbits are oriented in opposite direction.

On Σ, the orbit solutions of the system are governed by the Filippov rules. If $p \in \Sigma^c$, then the orbit of $Z = (X, Y) \in \Omega^r$ at p is defined as the concatenation of the orbits of X and Y at p. Nevertheless, if $p \in \Sigma \setminus \Sigma^c$, then it may occur a lack of uniqueness of solutions. In this case, the flow of Z is multivalued and any possible trajectory passing through p originated by the orbits of X, Y and F_Z is considered as a solution of Z. More details can be found in [11, 14].

The set of all singularities of X and Y in Σ, the tangency set S_Z of Z and the pseudo-equilibria of Z are referred as the **Σ-singularities** of Z.

Among all Σ-singularities, the tangential singularities, i.e. points of S_Z, are quite interesting, since they usually has no counterparts in the smooth world. In order, to characterize them, we introduce the higher order Lie derivatives of h: for $X_1, \cdots, X_k \in \chi^r$, it s defined recurrently as

$$X_k \cdots X_1 h(p) = X_k(X_{k-1} \cdots X_1 h)(p),$$

i.e. $X_k \cdots X_1 h(p)$ is the Lie derivative of the smooth function $X_{k-1} \cdots X_1 h$ in the direction of the vector field X_k at p. In particular, $X^k h(p)$ denotes $X_k \cdots X_1 h(p)$, where $X_i = X$, for $i = 1, \cdots, k$.

One of the most intriguing tangential singularities, due to its simplicity, is the **fold singularity** of X (resp. Y) which is characterized by a point p of Σ for which $Xh(p) = 0$ (resp. $Yh(p) = 0$) and $X^2h(p) \neq 0$ (resp. $Y^2h(p) \neq 0$). If the orbit though p is visible (resp. invisible) in the subset where the vector field is acting, then it is said to be a visible fold singularity (resp. invisible fold singularity). More specifically, when p is a fold singularity for both X and Y, it is said to be a **fold-fold singularity**. There are three types of fold-fold singularities: hyperbolic, parabolic and elliptic, which corresponds to the visible-visible, visible-invisible (and invisible-visible), and invisible-invisible case, respectively. In this paper we focus on the elliptic case.

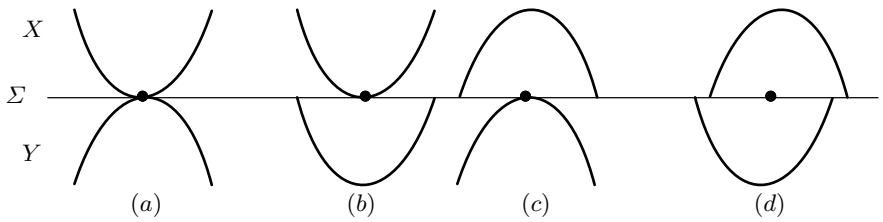


Fig. 8.2: Fold-Fold Singularity: (a) Hyperbolic, (b,c) Parabolic and (d) Elliptic.

Since we are dealing with local phenomena, there is no loss of generality in assuming that $h(x_1, x_2, ..., x_{n+1}) = x_{n+1}$ throughout the paper. In this way, the tangency set S_X (resp. S_Y), between $X = (X^1, X^2, ..., X^{n+1})$ (resp. $Y = (Y^1, Y^2, ..., Y^{n+1})$) and the hyperplane $\Sigma = \{x_{n+1} = 0\}$ is expressed by the equation $Xh = X^{n+1} = 0$ (resp. Y^{n+1}). Note that generically the dimension of S_X (resp. S_Y) is $(n-1)$. In this case, we can write $Z = (X, Y)$ as

$$Z(x) = \begin{cases} X(x), & x_{n+1} \geq 0, \\ Y(x), & x_{n+1} \leq 0, \end{cases}$$

and we consider that 0 is an invisible fold singularity of both vector fields, X and Y. That means that 0 is an elliptic fold-fold singularity of Z. So, in this coordinate system, our concern is to study the behavior of PSVF $Z = (X, Y) \in \Omega^r$ such that $Xh(0) = Yh(0) = 0$, $X^2h(0) < 0$ and $Y^2h(0) > 0$.

In this scenario, since 0 is a fold singularity of X (resp. Y) then there exists a map associated to X (resp. Y), $\phi_X : (\Sigma, 0) \to (\Sigma, 0)$ (resp. ϕ_Y), which is induced by the orbits of X (resp. Y) such that:

- $\phi_X^2 = id$ and $\phi_Y^2 = id$, i.e., ϕ_X and ϕ_Y are involutive maps;
- ϕ_X (resp. ϕ_Y) is a C^r-diffeomorfism at $(\Sigma, 0)$ with $\phi_X(0) = 0$ (resp. $\phi_Y(0) = 0$);
- The respective fixed-point sets F_X and F_Y of ϕ_X ϕ_Y, are given by the tangency sets S_X and S_Y, respectively, and have dimension $(n-1)$;
- 0 is a fixed point of the first return map $\phi = \phi_X \circ \phi_Y$.

As far as we know, [22] was the first work where the dynamics around a generic 3-dimensional fold-fold singularity was associated to the study of a 2-dimensional first return mapping ϕ that occurs around such singularity. In [24], the authors give necessary conditions for the structural stability of the 3-dimensional fold-fold singularity and proved that it is not a generic property. It is worth to point out the works [6, 7, 8].

8.2.2 Reversible Mapping and Diagram of Fold Mappings

It is acknowledged that time-reversal symmetry is one of the fundamental symmetries discussed in natural science and it arises in many branches in physics. Time-reversible systems share many properties of Hamiltonian systems. In [16], it is presented a interesting survey on reversibility in dynamical systems accompanied by an extensive bibliography. As we shall see, elliptic fold-fold singularities are naturally related with reversible maps which takes into account all the symmetric properties of the problem.

Let A^r be the space of all C^r-mappings $f : (\mathbb{R}^n, 0) \to (\mathbb{R}^n, 0)$ presenting a fold singularity at 0. In [17], the authors prove that there is a unique involutive symmetric diffeomorphism ϕ_f associated to f (i.e. $f \circ \phi_f = f$). The sets $Fix(\phi_f)$ and the singular set of $f \in A^r$ are coincident and has dimension $n-1$.

A normal form of a C^r fold mapping $f : (\mathbb{R}^n, 0) \to (\mathbb{R}^n, 0)$ is

$$f_0(u_1, u_2, ..., u_n) = ({u_1}^2, u_2, u_3, ..., u_n),$$

and its associated (symmetric) involution is

$$\phi_{f_0}(u_1, u_2, ..., u_n) = (-u_1, u_2, u_3, ..., u_n).$$

When we consider elliptic fold-fold singularities, we are dealing with two different vector fields X, Y, which give rise to two fold maps associated to them. So, it is not sufficient to study only fold mappings to understand such singularities, the study of diagrams of fold mappings are needed to its full comprehension.

Consider D^r as the set of all diagrams $D(f, g)$ of fold mappings at $(\mathbb{R}^n, 0)$.

$$D(f, g) : \{(\mathbb{R}^n, 0) \leftarrow_f (\mathbb{R}^n, 0) \ _g \to (\mathbb{R}^n, 0)\},$$

and let (ϕ_f, ϕ_g) be their associated pair of involutions.

One may get results on the topological classification of the diagram from a qualitative analysis of the behavior of the reversible mapping $\phi = \phi_f \circ \phi_g$. See [17, 18] for more details.

Remark 2. In dimension 2, we know that (see [22]) 0 is hyperbolic fixed point (saddle type) of ϕ only if the eigenvalues of $\phi'(0)$ are real and of the form λ and $\frac{1}{\lambda}$. Moreover ϕ is structurally stable at 0, under perturbations of f and g.

8.3 T-Singularity

Let $Z = (X, Y)$ be an n-dimensional Filippov system having an invisible two-fold singularity at $p \in \Sigma$ and assume $n \geq 4$. Let $\varphi_X : (\Sigma, S_X) \to (\Sigma, S_X)$ and $\varphi_Y : (\Sigma, S_Y) \to (\Sigma, S_Y)$ be the involutions associated to X and Y, respectively.

Now, S_X and S_Y are codimension 1 submanifolds of Σ and since S_X and S_Y are in general position, it follows that $M = S_X \cap S_Y$ is a codimension 2 submanifold of Σ. Notice that every point of M is a fixed point of $\varphi = \varphi_Y \circ \varphi_X$.

Therefore it follows that M is a submanifold of Σ having dimension $n-3$. In this case T_pM ia a subspace of $T_p\Sigma$ having dimension $n-3$. Let $v_1, v_2, \cdots, v_{n-3}$ be the elements of a basis of T_pM, then , for each $1 \leq i \leq n-3$, there exists a curve $\gamma_i : (-\varepsilon, \varepsilon) \to M$ such that $\gamma_i(0) = p$ and $\gamma'(0) = v_i$.

Notice that $\varphi \circ \gamma_i = \gamma_i$ and thus it follows that:

$$\begin{aligned}(\varphi \circ \gamma_i)'(0) &= d\varphi(p)(\gamma_i'(0)) \\ &= d\varphi_p(v_i)\end{aligned}$$

and thus $d\varphi_p(v_i) = v_i$, for each $1 \leq i \leq n-3$.

It means that v_i is an eigenvector associated to the eigenvalue 1 of $d\varphi_p$. Let $\mathscr{B} = \{w_1, w_2, v_1, v_2, \cdots, v_{n-3}\}$ be a basis of $T_p\Sigma$, then the matrix A of $d\varphi_p$ with respect to the basis $\mathscr{B}$ is given by

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \cdots & a_{1,n-1} \\ a_{2,1} & a_{2,2} & a_{2,3} & \cdots & a_{2,n-1} \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & 0 & \ddots & 0 \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

Thus, a simple change in the basis $\mathscr{B}$, allows us to express $d\varphi_p$ as

$$A = \begin{pmatrix} A_2 & 0 \\ 0 & I_{n-3} \end{pmatrix},$$

where I_{n-3} is the identity matrix of dimension $n-3$. Since $\varphi = \varphi_Y \circ \varphi_X$, it follows that the eigenvalues λ, μ of A_2 satisfy either

1. $\lambda, \mu \in \mathbb{C}$ and $\mu = \overline{\lambda}$ and $|\lambda| = 1$;
2. $\lambda \in \mathbb{R} \setminus \{0\}$ and $\mu = \lambda^{-1}$

So in the second case, when $\lambda \neq 1$, we have that φ has a semi-hyperbolic singularity of type saddle. In this case we have the existence of central, center-unstable and center-stable invariant manifolds. Moreover, there exists a change of coordinates which brings p to the origin and

$$\varphi(x_1, x_2, y) = (\lambda x_1, \lambda^{-1} x_2, y + f(y)),$$

which means that φ is decoupled. Notice that in this case W^{cs} and W^{cu} have dimension $n-2$ and the central manifold W^c has dimension $n-3$.

8.3.1 Invariant Manifolds

Notice that $W^c = S_X \cap S_Y$ is composed by fixed points of φ. When p is a semi-hyperbolic saddle, it follows that each point $q \in W^{cu}$ satisfies that

$$\lim_{n\to\infty} \varphi^{-n}(q) = r_q,$$

where r_q is a fixed point of φ in W^c.

Recall that $\varphi = \varphi_Y \circ \varphi_X$, and thus $\varphi^{-1} = \varphi_X^{-1} \circ \varphi_Y^{-1} = \varphi_X \circ \varphi_Y$. It follows that:

$$\begin{aligned}\varphi^n(\varphi_X(q)) &= (\varphi_Y \circ \varphi_X) \circ \cdots \circ (\varphi_Y \circ \varphi_X) \circ \varphi_X(q)\\ &= (\varphi_Y \circ \varphi_X) \circ \cdots \circ \varphi_Y(q)\\ &= \varphi_Y(\varphi^{-n+1}(q))\end{aligned}$$

Hence

$$\begin{aligned}\lim_{n\to\infty} \varphi^n(\varphi_X(q)) &= \varphi_Y\left(\lim_{n\to\infty} \varphi^{-n+1}(q)\right)\\ &= \varphi_Y(r_q)\\ &= r_q.\end{aligned}$$

It follows that, if $q \in W^{cu}$, then $\varphi_X(q) \in W^{cs}$. Moreover

$$\varphi_X(W^u(r_q)) \subset W^s(r_q).$$

Analogously, one can prove that

$$\varphi_Y(W^s(r_q)) \subset W^u(r_q).$$

It follows that, for each $r \in W^c$, we have an invariant cone with vertex at r foliated by crossing orbits of Z. So, in this case we have that:

$$W^{cs}(Z) = C^s \text{ and } W^{cu}(Z) = C^u,$$

where $C^{s,u}$ is the union of stable (unstable) invariant cones with vertexes at a point of $S_X \cap S_Y$. Also $W^c(Z) = W^c$.

So, it proves that Z has a center manifold of dimension $n-3$ and invariant center-stable and center-unstable manifolds of dimension 2.

In dimension $n = 3$ we have the classical nonsmooth diabolo illustrated in Figure 8.3. In higher dimensions, we may see $W^c(Z)$ as a $(n-3)$-parameter family of 2-dimensional diabolos.

8.4 A Bridge between Diagrams and T-Singularities

In this section, we describe how fold singularities of C^r vector fields with boundary relate to fold mappings. Such construction allows us to give a new interpretation for the results on T-singularities in the context of diagrams of fold mappings.

Let X_0 be a C^r vector field on $(\mathbb{R}^{n+1}, 0)$ and Σ, N be two n-dimensional hyperplanes in general position, with $X_0(p)$ transverse to N at 0. Using the Implicit Function Theorem, one can find neighborhood $\mathscr{U}$ of X_0 in χ^r, U of p in $\mathbb{R}^{n+1}$ and a C^r map $\tau : \mathscr{U} \times U \to \mathbb{R}$ such that, for each $X \in \mathscr{U}$, $\varphi_X(t, q) \in N$ with $q \in U$, if and only if, $t = \tau(X, q)$. So, for each $X \in \mathscr{U}$, we consider the C^r map

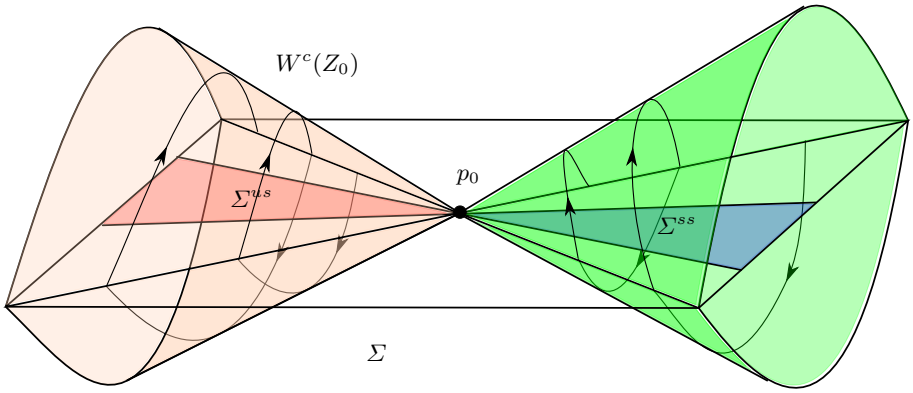


Fig. 8.3: Nonsmooth diabolo $W^c(Z_0)$ at a stable T-singularity p_0 of Z_0.

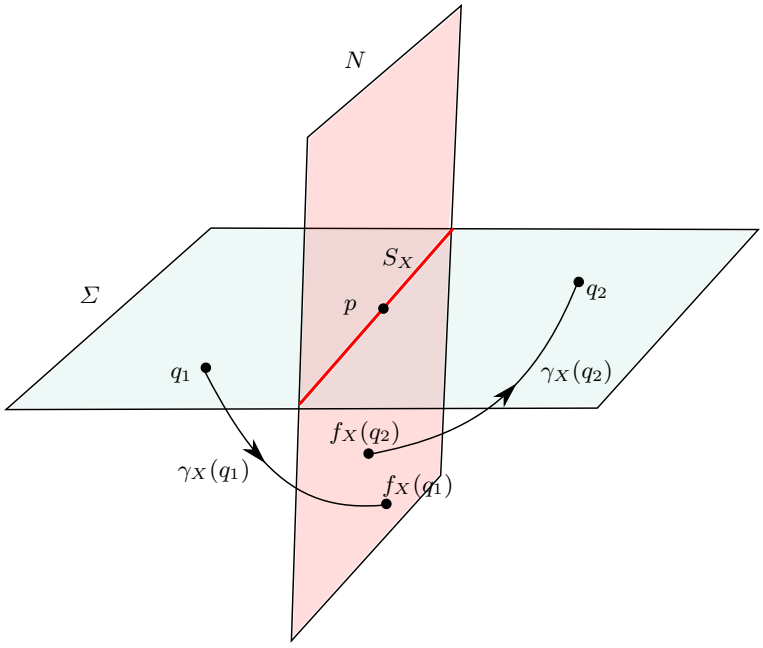


Fig. 8.4: Construction of the map f_X.

$$f_X : \Sigma \cap U \to N,$$

given by $f_X(q) = \varphi_X(\tau(X,q),q)$, which means that, for each $q \in \Sigma$, $f_X(q)$ is the point where the trajectory of X through q, $\gamma_X(q)$ reaches N (see Figure 8.4).

Remark 3. Notice that, in Figure 8.4, if $q_1 \neq q_2$ and q_1 and q_2 are in the same orbit of X, i.e. $\gamma_X(q_1) = \gamma_X(q_2)$, then $f_X(q_1) = f_X(q_2)$.

If the contact of $\gamma_{X_0}(p)$ with Σ is quadratic (i.e. p is a fold singularity), then f_{X_0} is a fold mapping at (Σ, p). Moreover, for each $X \in \mathscr{U}$, we have that all the following objects coincide:

1. S_X;
2. the singular set of f_X;
3. the fixed-points set, F_X, of the involution ϕ_X associated to the fold mapping f_X.

In short, an element $Z = (X, Y) \in \Omega^r$ at an elliptic fold-fold singularity, induces the emergence of a fold mapping diagram $D(f_X, f_Y)$ which is associated to a involutive mapping diagram $D(\phi_X, \phi_Y)$ which is related to the reversible map $\phi_Z = \phi_X \circ \phi_Y \in \Delta^r$. We would say that the dynamics ϕ_Z occupies an extremely important position in the study of a T-singularity.

8.5 Main Results

Now, we summarize the main results obtained from the discussions above. First, in the context of Filipov systems, we enunciate the results shown in Section 8.3.

Theorem 1. *Let $Z = (X, Y)$ be a Filippov system defined in $\mathbb{R}^n$ having an invisible two-fold singularity at $p \in \Sigma$ and assume that $n \geq 4$. The following statements hold:*

1. *S_X and S_Y are local submanifolds of Σ of dimension $n-2$.*
2. *If S_X and S_Y are in general position, then $S_X \cap S_Y$ is a submanifold of Σ of dimension $n-3$ which is composed by elliptic two-fold singularities. Moreover, $S_X \cap S_Y$ is the center manifold W^c of Z at p and it has dimension $n-3$.*
3. *In the conditions of item (2), Z has an invariant center-stable (resp. center stable) manifold W^{cs} (resp. W^{cu}) composed by a union of 2-dimensional cones. Each of this cones has its vertex at a point $p \in W^c$. Thus, W^{cs} (resp. W^{cu}) has dimension 2.*

Notice that the result above has been proven by analysing the local invariant manifolds of the first return map $\varphi : \Sigma \to \Sigma$ which is a composition $\varphi = \varphi_Y \circ \varphi_X$ of the two involutions φ_X and φ_Y associated to the elliptic T-singularity of $Z = (X, Y)$. Since, these are the unique ingredients to reach such result, it can be clearly extended to reversible mappings using exactly the same arguments.

Theorem 2. *Let $\phi = \phi_1 \circ \phi_2$ be a reversible mapping on $(\mathbb{R}^n, 0)$ where ϕ_1, ϕ_2 are involutions and 0 is a fixed point of ϕ. Assume that the respective fixed points set F_1 and F_2 of ϕ_1 and ϕ_2 have dimension $n-1$ and that they are in general position at 0. The following statements hold:*

1. *$F = F_1 \cap F_2$ is a $(n-2)$-dimension invariant submanifold of $\mathbb{R}^n$ which is composed by fixed points of ϕ. Moreover, $F_1 \cap F_2$ is the center manifold W^c of φ at 0.*
2. *If all the eigenvalues of $\phi'(0)$ are real, then there exists a basis of $\mathbb{R}^n$ for which $\phi'(0)$ is represented by a matrix A_Z having two blocks A_1 and A_2, where A_1 is a 2-dimensional diagonal matrix having non-zero entries λ and $1/\lambda$, for some $\lambda \neq \pm 1$, and A_2 is the identity matrix of dimension $n-2$.*
3. *In the conditions above, there exists a 1-dimensional center-stable manifold W^{cs} (resp center-unstable manifold W^{cu}) of ϕ at 0 which is composed by a union of lines transverse to F. It means that, at each point $p \in W^c$, there is a line transverse to F which is a stable invariant manifold of ϕ at p. Moreover, W^{cs} and W^{cu} are transverse.*

4. If all the eigenvalues of $\phi'(0)$ are real, then ϕ is locally structurally stable at 0 in the space of the reversible mappings $\psi = \psi_1 \circ \psi_2$ such that $\psi_{1,2}$ is an involution having an $(n-1)-$dimensional fixed point set F_1, F_2 such that F_1, F_2 are in general position.

8.5.1 Applications to Diagrams of Mappings

It is known that a diagram of involutions

$$D(\phi_1, \phi_2) : \{(\mathbb{R}^n, 0) \leftarrow_{\phi_1} (\mathbb{R}^n, 0) \ _{\phi_2} \rightarrow (\mathbb{R}^n, 0)\},$$

is simultaneously structurally stable at 0 if and only if $\phi_1 \circ \phi_2$ is locally structurally stable at 0, thus the next result follows directly from item 4 of Theorem 2.

Corollary 1. *Let $D(\phi_1, \phi_2)$ be a diagram of involutions at $(\mathbb{R}^n, 0)$ and assume that the respective fixed points set F_1 and F_2 of ϕ_1 and ϕ_2 have dimension $n-1$ and that they are in general position at 0 and $0 \in F_1 \cap F_2$. If the eigenvalues of $(\phi_1 \circ \phi_2)'(0)$ are real, then $D(\phi_1, \phi_2)$ is simultaneously structurally stable in the space of the diagrams of involutions satisfying the same initial assumptions.*

A diagram of fold mappings

$$D(f, g) : \{R^n, 0 \leftarrow_f R^n, 0 \ _g \rightarrow \ R^n, 0\},$$

is simultaneously structurally stable when its associated diagram of involutions $D(\phi_X, \phi_Y)$ is also simultaneously structurally stable, thus the corollary above gives a necessary condition for stability of diagram of fold mappings.

8.6 A discussion on Global Dynamics and Further Directions

In [13], an interesting phenomenon has been detected from the communication between the branches of a nonsmooth diabolo at a T-singularity. In fact, it has been shown that, under generic conditions, such communication lead us to a robust chaotic situation. More specifically, in this case, a Smale horseshoe is detected in the crossing region providing chaos. See [13] for more details.

In higher dimensions, we have that the dynamical behavior of the invariant manifolds of a T-singularity does not suffer a drastic change, since the center-stable and center-unstable manifolds remains with dimension 2. In this sense, the same global construction which communicates the branches of the diabolo in dimension 3 can be employed to give rise to a communication between $W^{cs}(Z)$ and $W^{us}(Z)$, and the same approach can be used to construct a Smale horseshoe for this systems.

It is worth to mention that, since $W^c(Z)$ is a $(n-3)$-parameter family of diabolos, and from construction, the variation between these diabolos is smooth, it folllows that, if we have a generic communication between two branches of a diabolo centered at a T-singularity p_0, then for T-singularities p near p_0, we have that the diabolo centered at p also presents the same communication. So in higher dimension one can construct a $(n-3)$-parameter family of Smale horseshoes using exactly the same process described in [13].

Remark 4. Notice that in higher dimensions $n \geq 4$, one can have also another type of chaos which is generated from the communication between the branches of diabolos centered at two distinct T-singularities p_1 and p_2, i.e. $W^{cs}(p_1) \cap W^{cu}(p_2) \neq \emptyset$. Nevertheless, the same constructions can be done to obtain a chaotic situation.

In light of this, a natural extension of this work is to investigate multiple connections arising from the communication of branches of T-singularities, which can give rise to deterministic and non-determinist chaos. Also, such characterization can be translated to the global characterization of chaos in diagrams of involutions arising from the intersections on center-stable and center-unstable manifolds.

Problems of stability in relay systems are discussed in numerous applications. In fact, Control Theory is a natural source of mathematical models of these systems. It is worth mentioning that Anosov, in his first paper, studied a class of relay systems in $\mathbb{R}^n$ of the form

$$u' = Au + \text{sgn}(u_1)k,$$

where $u = (u_1, u_2, \cdots, u_n)$, A is a $n \times n$ real matrix and $k = (k_1, k_2, \cdots, k_n)$ is a constant vector in $\mathbb{R}^n$. See [2] for more details. In this sense, fold-fold singularities naturally appear in this scenario, and the theoretical results developed here can be used to understand applied problems.

References

1. Andronov, A. , Leontovich, E. A. , I., Gordon, I., Maier A. G., Theory of bifurcations of dynamic systems on a plane. John Wiley and Sons, 1971.
2. Anosov, D. V. *Stability of the equilibrium positions in relay systems,Automation and remote control*, V. XX, 2, 1959.
3. Balakrishnan, A.V.: Applied Functional Analysis. Springer-Verlag, New York, (1976)
4. Barbashin, E. A. (1970) Introduction to the theory of stability. Wolters-Noordhoff Publishing, Groningen. Translated from the Russian by Transcripta Service, London. Edited by T. Lukes, pp 223.
5. Buzzi C. A., Medrado J. C. da R. , Teixeira M. A., . Generic bifurcation of refracted systems, Advances in Mathematics (NY, 1965),V. 234, 2013, p. 653-666.
6. Colombo, A., Jeffrey, M. R. The two-fold singularity of discontinuous vector fields. SIAM J. Applied Dynamical Systems, 8(2), p. 624-640, 2009.
7. Colombo, A., Jeffrey, M. R. Nondeterministic chaos, and the two-fold singularity in piecewise smooth flows. SIAM J. Applied Dynamical Systems, 10(2), p. 423-451, 2011.
8. Colombo, A., Jeffrey, M. R. The two-fold singularity of nonsmooth flows: Leading order dynamics in n-dimensions. Physica D Nonlinear Phenomena, 263, 1-10, 2013.
9. di Bernardo, M., Budd C.J., Champneys A. R., Kowalczyk P., Piecewise-Smooth Dynamical Systems: Theory and Applications, Appl. Math. Sci. Series 163, Springer-Verlag, London, 2008.
10. Ekeland, I., *Discontinuité des champs Hamiltoniens er existence de solutions optimales en calcul des variations*. Pub. IHES, 47, 5-32, 1977.
11. Filippov, A. F. Differential equations with discontinuous righthand sides. Kluwer, 1988.
12. Gomide, O. M. L., Teixeira, M A, On structural stability of 3D Filippov systems. Mathematische zeitschrift, v. 294, p. 419-449, 2019.
13. Gomide, O. M. L. , Teixeira, M. A., Chains in 3D Filippov systems: A chaotic phenomenon, Journal de Mathématiques Pures et Appliquées, Volume 159, 2022, Pages 168-195, ISSN 0021-7824, https://doi.org/10.1016/j.matpur.2021.12.002.
14. Guardia, M., Seara, T., and Teixeira, M. A. Generic bifurcations of low codimension of planar filippov systems. Journal of Differential Equations, 250(4):1967 – 2023, 2011.
15. Klok, F., Broken solutions of homogeneous variational problems, J. Differential Equations 55 (1984), 101134.
16. Lamb, J. S. W., Reversing symmetries in dynamical systems, Thesis, University of Amsterdam (1994).

17. Mancini, S. , Manoel, M. ; Teixeira, M. A.; Divergent diagrams of folds and simultaneous conjugacy of involutions. Discrete and Continuous Dynamical Systems, 2005, 12(4): 657-674. doi: 10.3934/dcds.2005.12.657
18. Mancini, S. , Manoel, M. ; Teixeira, M. A.; Simultaneous linearization of a class of pairs of involutions with normally hyperbolic composition. Bulletin des Sciences Mathematiques (Paris. 1885), v. 137, p. 418-433, 2013.
19. Minorsky, N., Theorie des oscillations, (French) Memorial des Sciences Mathematiques, Fasc. 163 Gauthier-Villars Editeur, Paris, (1967) i+114 pp.
20. Minorsky, N. Theory of nonlinear control systems, McGraw-Hill Book Co., New York-London-Sydney, (1969) xx+331 pp.
21. Quispe J. Estabilidade estrutural de campos de vetores suave por partes, PhD Thesis IMECC Unicamp, 2013.
22. Teixeira, M.A., On topological stability of divergent diagrams of folds, Math. Z., V. 180, 1982, 361-371.
23. Teixeira, M.A. , Perturbation Theory for Non-smooth Systems, in -Encyclopedia of Complexity and Systems Science,Perturbation Theory: Mathematics, Methods and Applications, 503-517, 2009.
24. Teixeira, M.A.; Gomide, O. M. L. . Generic Singularities of 3D Piecewise Smooth Dynamical Systems. In: Carlile Lavor and Francisco A. M. Gomes. (Org.). Advances in Mathematics and Applications. 1ed.: Springer International Publishing, 2018, v. , p. 373-404.-449, 2019.
25. Vishik, S. M., *Vector fields near the boundary of a manifold.* Vestnik Moskovskogo Universiteta Mathematika, **27(1)**, 21-28, 1972.
26. Zelikin, M.I., Borisov, V.F. Theory of chattering control with applications to Astronautics, Robotics, Economics, and Engineering, (1994), Birkhauser.

9

Exponential Stability of Heat Exchanger Systems and Heat-Plate Coupled Systems

Cheng-Zhong Xu[1] and Qiong Zhang[2]

[1] Department of Mechanics, LAGEPP, University Claude Bernard Lyon 1, 43 Boulevard du 11 novembre 1918, F-69622, Villeurbanne, France. cheng-zhong.xu@univ-lyon1.fr
[2] Department of Mathematics and Statistics, Beijing Institute of Technology, Beijing, China. zhangqiong@bit.edu.cn

Summary. This note proves exponential stability of a classical heat exchanger equation and a heat-plate transmission system. We give both a sharp estimate of the exponential decay rate and a precise spectrum characterization of the systems.

9.1 Introduction

The motivation of this note is to answer the open question left by [9], [36] and [40]. In [36] only strong stability was proved for the contraflow heat exchanger system from the point of view of linear distributed parameter systems. Exponential stability was a very desirable property for the H^∞ control design of this kind of systems (see [9]). On the other hand, the heat exchanger equation is also a typical and fundamental distributed parameter model for studying hyperbolic dynamic processes (see [8]). By using the Lyapunov direct method exponential stability has been established in [38] for a class of general hyperbolic heat exchanger systems. Today much more general results can be found in this aspect (see [4, 20]).

In this paper, using the necessary and sufficient characteristic condition due to F.L. Huang [15], we prove that the contraflow heat exchanger system is exponentially stable for every set of physical parameters. We give a precise spectrum characterization of the system and provide a sharp estimate of the exponential decay rate to stability. We show that the spectrum determined growth condition is satisfied for the heat exchanger system. These results would be helpful to understanding the dynamic behaviour of the exchanger system and instructive for the control purpose. We should recall that the result of [15] has been successfully applied for the first time in [7] to establish an exponential decay result for an Euler-Bernoulli beam with rate control of the beam moment only. It was difficult to prove exponential stability of the latter system without using the Huang-Prüss theorem.

One the other hand, we consider exponential stability of a two-dimensional heat-plate transmission system. By applying the frequency domain method, the regularity theory of elliptic partial differential equations etc., it is shown that dissipation solely from the heat equation is sufficient for exponential stability of the coupled system.

[3] This paper is dedicated to the memory of Professor I. Kupka

For convenience of the reader, we will restate Huang's result [15] in Proposition 1. Let H be a separable Hilbert space (with the inner product $< \cdot, \cdot >$). Consider on H the linear system : $\dot{x}(t) = Ax(t)$ where A is the infinitesimal generator of a C_0-semigroup $(e^{tA})_{t\geq 0}$ on H. We denote by $\sigma(A)$ the spectrum set of A and by $\rho(A)$ the resolvent set, i.e., $\rho(A) = \mathbb{C}\backslash\sigma(A)$. For each element $f \in H$ we write its norm as $\|f\| = \sqrt{< f, f >}$. Similarly, for each linear continuous mapping $T : H \to H$ we write its induced norm also as $\|T\|$ if there is no confusion from the context.

The order, or growth rate of the semigroup is defined by $\omega_0(A) = \lim_{t\to+\infty} t^{-1} \ln(\|e^{tA}\|)$ (which does exist). Let the spectral bound be $\sigma_{max}(A) = \sup\{\Re e(\lambda) \mid \lambda \in \sigma(A)\}$. Then it is always true that $\sigma_{max}(A) \leq \omega_0(A)$ (see p.20, [30]). It is evident that for each $\epsilon > 0$, there exists a $M_\epsilon > 0$ such that $\|e^{tA}\| \leq M_\epsilon e^{t(\omega_0(A)+\epsilon)}\ \forall\ t \geq 0$. If $\omega_0(A) < 0$, we say that the system is exponentially stable. In general, $\sigma_{max}(A) < 0$ does not imply that the system is exponentially stable. Two interesting examples can be found in [30] (p.117, [30]) and [15] in which $\sigma_{max}(A) < 0$ and however $\|e^{tA}\| \geq 1$. The following fundamental result was proved in [15].

Theorem 1. *$\sigma_{max}(A) = \omega_0(A)$ if and only if, for each $\sigma > \sigma_{max}(A)$, the following property holds:*

$$\sup_{\Re e(\lambda)\geq\sigma} \left\{\|(\lambda I - A)^{-1}\|\right\} < \infty.$$

We will use this proposition to prove exponential stability of the exchanger system. Another variant of Theorem 1 usually called Huang-Prüss theorem is specific but more useful in practice stated as follows.

Theorem 2. *Suppose that the semigroup $(e^{tA})_{t\geq 0}$ on H is uniformly bounded. It is exponentially stable if and only if $i\mathbb{R} \subset \rho(A)$ and*

$$\sup_{\omega\in\mathbb{R}}\{\|(i\omega I - A)^{-1}\|\} < \infty. \tag{9.1}$$

The classical contraflow heat exchanger with constant flowrates is modeled by the following linear partial differential equations:

$$\begin{array}{l} \dfrac{\partial R_1(x,t)}{\partial t} = -F_1\dfrac{\partial R_1(x,t)}{\partial x} + h_1(R_2(x,t) - R_1(x,t)) \\ \dfrac{\partial R_2(x,t)}{\partial t} = F_2\dfrac{\partial R_2(x,t)}{\partial x} + h_2(R_1(x,t) - R_2(x,t)) \end{array} \tag{9.2}$$

with the boundary conditions :

$$R_1(0,t) = 0 \ \text{ and } \ R_2(l,t) = 0. \tag{9.3}$$

In (9.2) and (9.3), $R_1(x,t)$ and $R_2(x,t)$ denote the temperature variation at the time t and at the point $x \in [0,l]$ with respect to an equilibrium point. The usual assumptions are made in (9.2) and (9.3) : the heat capacity, heat transfer coefficient and mass density are constant such that $F_1 > 0$, $F_2 > 0$, $h_1 > 0$, $h_2 > 0$ and $l > 0$. (The physical parameters are always positive.) For other physical aspects of the exchanger, the reader is referred to [8].

Let the state space for the system be the Hilbert space $H = L^2(0,l) \times L^2(0,l)$ with the inner product :

$$< f, g >= \int_0^l [f_1(x)g_1(x) + f_2(x)g_2(x)]dx.$$

We define the unbounded linear operator $A : \mathscr{D}(A) \subset H \to H$ by

$$\mathscr{D}(A) = \left\{ f = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \middle| \, f, \ f_x \in H, f_1(0) = f_2(l) = 0 \right\}$$

and

$$Af = \begin{pmatrix} -F_1 f_{1x} \\ F_2 f_{2x} \end{pmatrix} + \begin{pmatrix} -h_1 & h_1 \\ h_2 & -h_2 \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \ \forall \ f \in \mathscr{D}(A).$$

Proposition 1. *The unbounded linear operator A is the infinitesimal generator of a uniformly bounded C_0-semigroup $(e^{tA})_{t\geq 0}$ on H. Moreover it has compact resolvent operators.*

Proof. To prove that A is the generator of a C_0 semigroup on H we consider just the differential operator A_1 defined by $\mathscr{D}(A_1) = \mathscr{D}(A)$ and

$$A_1 f(x) = \begin{pmatrix} -F_1 f_1'(x) \\ F_2 f_2'(x) \end{pmatrix} \ \forall \ f \in \mathscr{D}(A).$$

We prove that A_1 is maximal dissipative: (i) it is dissipative because

$$< Af, f >= \int_0^l (-F_1 f_1'(x) f_1(x) + F_2 f_2'(x) f_2(x)) dx$$

$$= -\frac{F_1}{2} f_1^2(l) - \frac{F_2}{2} f_2^2(0) \leq 0 \ \forall \ f \in \mathscr{D}(A_1);$$

(ii) it is maximal because $Range(I - A_1) = H$. Indeed, for each $g \in H$ we solve the equation $f - A_1 f = g$ for $f \in \mathscr{D}(A_1)$. Obviously the solution is given by

$$f_1(x) = \frac{1}{F_1} \int_0^x e^{-(x-\xi)/F_1} g_1(\xi) d\xi, \tag{9.4}$$

$$f_2(x) = \frac{1}{F_2} \int_x^l e^{(x-\xi)/F_2} g_2(\xi) d\xi. \tag{9.5}$$

By the classical theorem [Brézis], A_1 generates a C_0 semigroup of contractions on H. The second part of A is a bounded operator from H to H,

$$f \to Bf = \begin{pmatrix} -h_1 & h_1 \\ h_2 & -h_2 \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}.$$

By the bounded pertubation theorem [Pazy], the operator $A = A_1 + B$ is also the generator of a C_0 semigroup. Now let un prove that the generated semigroup is a uniformly bounded one.

For the purpose we consider the following transformation $T : H \to H$ such that

$$Tf(x) = \begin{pmatrix} \sqrt{h_1} & 0 \\ 0 & \sqrt{h_2} \end{pmatrix} f(x).$$

It is easy to check that

$$\tilde{A} = T^{-1} A T = A_1 + T^{-1} B T = A_1 + \begin{pmatrix} -h_1 & \sqrt{h_1 h_2} \\ \sqrt{h_1 h_2} & -h_2 \end{pmatrix}$$

and the second part of $\tilde{A}$ is also dissipative. Hence $\tilde{A}$ generates a C_0 semigroup of contractions. By the inverse transformation we get that $\|e^{tA}\| \leq M$ for some positive constant M. Remark also that from (9.4)-(9.5) the resolvent operator $(I-A)^{-1}$ is compact, hence compact for all $\lambda \in \rho(A)$ [36]. Indeed every weakly converging (to zero) sequence (g_n) is by (9.4)-(9.5) transformed to a strongly converging sequence (f_n) that implies that $(I-A)^{-1}$ is compact. □

Definition 1. *We call the semigroup $(\mathbb{T}(t))_{t\geq 0}$ strongly stable if the following holds:*

$$\lim_{t\to\infty} \mathbb{T}(t)f = 0 \; \forall \; f \in H.$$

We get immediately strong asymptotic stability of the zero equilibrium state $(0,0)$.

Proposition 2. *The semigroup $(e^{tA})_{t\geq 0}$ is strongly stable on H.*

Proof. After the tansformation the haet exchanger system is written as

$$\dot{\varphi}(t) = (A_1 + B_1)\varphi(t), \; \varphi(0) = \varphi_0 \in H \tag{9.6}$$

where

$$B_1 = \begin{pmatrix} -h_1 & \sqrt{h_1 h_2} \\ \sqrt{h_1 h_2} & -h_2 \end{pmatrix}.$$

Take the initial condition $\varphi_0 \in \mathscr{D}(\mathscr{A})$. Consider the Lyapunov functional $V : H \to V$ such that $V(\varphi) = \|\varphi\|^2$. By differentiating $V(\varphi(t)$ along the trajectory of the system we get easily

$$\dot{V}(\varphi(t)) = -F_1\varphi_1^2(l,t) - F_2\varphi_2^2(0,t) - < B_1\varphi(t), \varphi(t) > .$$

By the LaSalle principle we have necessarily:

$$\varphi_1^2(l,t) = 0, \;\; \varphi_2^2(0,t) = 0, \;\; < B_1\varphi(t), \varphi(t) >= 0.$$

In other word the ω-limit set $\omega(\varphi_0)$ is contained in the subset E:

$$E = \{\varphi \in \mathscr{D}(A) \mid \varphi_i(0) = \varphi_i(l) = 0, i = 1,2, \; < B_1\varphi, \varphi >= 0\}.$$

Take a $\psi \in \omega(\varphi_0) \subset E$. We have the trajectory governed by

$$\dot{\varphi}(t) = A_1\varphi(t), \;\; \varphi(0) = \psi.$$

By the invariance principle of LaSalle, $y(t) = (\varphi_1(l,t), \varphi_2(0,t)) = (0,0) \; \forall \; t \geq 0$. By the method of characteristics it is easy to see that $\psi = 0$. Hence $\omega(\varphi_0) = \{0\}$. So the asymptotic stability is proved. □

Remark 1. Consequently the spectrum $\sigma(A)$ consists entirely of isolated eigenvalues with finite multiplicity (without accumulation point different from ∞) (p.187 [16]). Moreover the whole imaginary axis is contained in the resolvent set, i.e., $i\mathbb{R} \subset \rho(A)$.

In fact the semigroup is exponentially stable as stated in the following.

Theorem 3. *The semigroup $(e^{tA})_{t\geq 0}$ is exponentially stable on the state space H, i.e., there exist positive constant $M > 0$ et $\omega > 0$ such that*

$$\|e^{tA}\|_{\mathscr{L}(H)} \leq Me^{-\omega t} \;\; \forall \; t \geq 0.$$

Proof. The proof will be done by using Theorem 2. First $i\mathbb{R} \subset \rho(A)$ is true by strong stability of the semigroup since the spectrum of the generator is discrete. Now let us prove the condition (9.1) satisfied. We prove it by absurd argument.

Suppose that the condition (9.1) is not satisfied. Then there exist a sequence of real numbers (ω_n) ($\omega_n > 0$ without loss of generality) and a sequence (f_n) in $\mathscr{D}(A)$ such that

$$\lim_{n\to\infty} \omega_n = \infty, \ \ \|f_n\| = 1, \ \ \lim_{n\to\infty} \|(i\omega_n - A)f_n\| = 0.$$

It will lead to a contradiction. Indeed, set

$$(i\omega_n - A_1 - B_1)f_n = g_n \tag{9.7}$$

where $f_{n,1}(0) = 0$ and $f_{n,2}(l) = 0$. We claim that

$$\lim_{n\to\infty} f_{n,1}(l) = \lim_{n\to\infty} f_{n,2}(0) = 0. \tag{9.8}$$

Taking the inner product in the equation (9.7) we get

$$0 = \Re e < i\omega_n f_n, f_n >= \Re e < A_1 f_n, f_n > + < B_1 f_n, f_n > + \Re e < g_n, f_n > .$$

Direct computation gives us the following:

$$\frac{F_1}{2}|f_{n,1}(l)|^2 + \frac{F_2}{2}|f_{n,2}(0)|^2 +$$

$$\int_0^l \Big(h_1|f_{n,1}(x)|^2 + h_2|f_{n,2}(x)|^2 - 2\sqrt{h_1 h_2}\Re e(f_{n,2}(x)\bar{f}_{n,1}(x))\Big)\, dx$$

$$= \Re e < g_n, f_n > .$$

As the right-hand side tends to zero and that each left-hand term is positive, it follows that

$$\lim_{n\to\infty} f_{n,1}(l) = \lim_{n\to\infty} f_{n,2}(0) = 0,$$

$$\lim_{n\to\infty} \int_0^l \Big(h_1|f_{n,1}(x)|^2 + h_2|f_{n,2}(x)|^2 - 2\sqrt{h_1 h_2}\Re e(f_{n,2}(x)\bar{f}_{n,1}(x))\Big)\, dx = 0.$$

Next we prove that $\lim_{n\to\infty} \|f_n\| = 0$ being contradictory with the assumption $\|f_n\| = 1$. Indeed, rewrite the equation (9.7) as follows:

$$\begin{aligned} f'_{n,1}(x) &= -\frac{h_1 + i\omega_n}{F_1} f_{n,1}(x) + \frac{h_1}{F_1} f_{n,2}(x) + \frac{1}{F_1} g_{n,1}(x) \\ f'_{n,2}(x) &= \frac{h_2 + i\omega_n}{F_2} f_{n,2}(x) - \frac{h_2}{F_2} f_{n,1}(x) - \frac{1}{F_2} g_{n,2}(x) \end{aligned} \tag{9.9}$$

with the initial condition:

$$f_{n,1}(0) = 0, \ \lim_{n\to\infty} f_{n,2}(0) = 0.$$

Multiplying the first equation in (9.9) by the complex conjugate function $\bar{f}_{n,1}(x)$ and the second one by $\bar{f}_{n,2}(x)$, respectively, and then adding the real part together we get the following:

$$\frac{d}{dx}[|f_{n,1}(x)|^2+|f_{n,2}(x)|^2]=-\frac{2h_1}{F_1}|f_{n,1}(x)|^2+\frac{2h_2}{F_2}|f_{n,2}(x)|^2+$$
$$\frac{2(h_1F_2-h_2F_1)\Re e(f_{n,1}\bar{f}_{n,2})}{F_1F_2}+\frac{2\Re e(g_{n,1}\bar{f}_{n,1})}{F_1}-\frac{2\Re e(g_{n,2}\bar{f}_{n,2})}{F_2}.$$

As $|\Re e(g_{n,2}\bar{f}_{n,2})|\leq|g_{n,2}|\,|f_{n,2}|$, from the last equation we can find some positive constants $k_1>0$ and $k_2>0$ such that

$$\frac{d}{dx}\|f_n(x)\|^2_{\mathbb{R}^2}\leq k_1\|f_n(x)\|^2_{\mathbb{R}^2}+k_2\|g_n(x)\|^2_{\mathbb{R}^2}.$$

Solving the differential inequality leads to the following:

$$\|f_n(x)\|^2_{\mathbb{R}^2}\leq e^{k_1x}\|f_n(0)\|^2_{\mathbb{R}^2}+k_2\int_0^x e^{k_1(x-\xi)}\|g_n(\xi)\|^2_{\mathbb{R}^2}d\xi.$$

Taking the integral on $[0,l]$ from the last inequality gives us

$$\|f_n\|^2\leq\frac{e^{k_1l}-1}{k_1}\left[\|f_n(0)\|^2_{\mathbb{R}^2}+k_2\|g_n\|^2\right].$$

Since each right-hand term tends to zero as $n\to\infty$, we prove the contradiction that $\lim_{n\to\infty}\|f_n\|=0$. Hence the proof of exponential stability of the semigroup is complete. □

For the moment we have no idea about the exponential decay rate $\omega>0$ of the exponentially stable semigroup (e^{tA}). We will gave a shape estimate to the exponential decay rate.

Throughout the paper we set

$$\alpha_1=\frac{h_1}{F_1},\ \beta_1=\frac{h_2}{F_2},\ \alpha_2=\frac{1}{F_1},\ \text{ and }\beta_2=\frac{1}{F_2}\ \text{ and }\ a=\pm\frac{1}{l\sqrt{\alpha_1\beta_1}}. \tag{9.10}$$

9.2 Exponential Stability of Heat Exchangers

As we have already proved exponential stability of the semigroup we like to further know what is the maximal exponential decay rate that we can expect of the semigroup. In other words could we expect that the maximal decay rate is determined by the spectral bound of the semigroup generator, as the same as for linear systems of finite dimension ?

Our main result is the following :

Theorem 4. *(i) The heat exchanger system ((9.2) and (9.3)) is exponentially stable for every set of physical parameters. Moreover the spectrum determined condition is satisfied, i.e., the exponential decay rate is determined by the spectral bound: $\omega_0(A)=\sigma_{max}(A)$. (ii) For each $\epsilon>0$, there exists a constant $M_\epsilon>0$ such that if $a^2\geq1$,*

$$\|e^{tA}\|\leq M_\epsilon\exp\left(t\left(-\frac{(\sqrt{\alpha_1}+\sqrt{\beta_1})^2}{\alpha_2+\beta_2}+\epsilon\right)\right),$$

and if $a^2<1$,

$$\|e^{tA}\|\leq M_\epsilon\exp\left(t\left(-\frac{(\sqrt{\alpha_1}-\sqrt{\beta_1})^2+\delta}{\alpha_2+\beta_2}+\epsilon\right)\right)$$

where $\delta=2\sqrt{\alpha_1\beta_1}\left(1+\min\left\{a^{-1}|a|\cos(y)\ \middle|\ \begin{array}{l}\sin(y)=ay\\ a^2=(l^2\alpha_1\beta_1)^{-1}\end{array}\right\}\right)>0.$

We will prove Theorem 4 at the end of the paper through several lemmas. The first lemma gives a complete description of spectrum $\sigma(A)$ and a sharp estimate of spectral bound of A.

Lemma 1. *(i) Assume that $l\sqrt{\alpha_1\beta_1} \neq 1$. Then the spectrum $\sigma(A)$ constituted of eigenvalues only is described by*

$$\sigma(A) = S_A = \left\{ \lambda = -\frac{\alpha_1+\beta_1}{\alpha_2+\beta_2} - \frac{2\cos(y)\cosh(x)}{al(\alpha_2+\beta_2)} - i\frac{2\sin(y)\sinh(x)}{al(\alpha_2+\beta_2)} \;\middle|\; (x,y) \in \mathbb{R}^2, x \geq 0, \left\| \begin{bmatrix} x \\ y \end{bmatrix} \right\| > 0, \begin{matrix} \cos(y)\sinh(x) = ax \\ \sin(y)\cosh(x) = ay \end{matrix} \right\}.$$

(ii) If $l\sqrt{\alpha_1\beta_1} = 1$, then

$$\sigma(A) = S_A \cup \left\{ \lambda = \frac{-(\sqrt{\alpha_1}+\sqrt{\beta_1})^2}{\alpha_2+\beta_2} \right\}.$$

(iii) If $a^2 = \dfrac{1}{l^2\alpha_1\beta_1} \geq 1$, then

$$\sigma_{max}(A) = \sup_{\lambda\in\sigma(A)} \Re e(\lambda) \leq \frac{-(\sqrt{\alpha_1}+\sqrt{\beta_1})^2}{\alpha_2+\beta_2}.$$

(iv) If $a^2 < 1$, then

$$\sigma_{max}(A) \leq \frac{-(\sqrt{\alpha_1}-\sqrt{\beta_1})^2 - 2\sqrt{\alpha_1\beta_1}\left[1 + \min_{\sin(y)=ay}\{a^{-1}|a|\cos(y)\}\right]}{\alpha_2+\beta_2}.$$

Proof. A complex number $\lambda \in \sigma(A)$ if and only if the following differential equations have a non-zero solution :

$$\begin{aligned} &-F_1 f_{1x}(x) - h_1 f_1(x) + h_1 f_2(x) = \lambda f_1(x) \\ &F_2 f_{2x}(x) + h_2 f_1(x) - h_2 f_2(x) = \lambda f_2(x) \\ &f_1(0) = f_2(l) = 0. \end{aligned} \tag{9.11}$$

These equations are reduced to the following equivalent differential equation:

$$\begin{aligned} &f_{2xx}(x) + [(\alpha_2-\beta_2)\lambda + \alpha_1 - \beta_1] f_{2x}(x) - [\alpha_2\beta_2\lambda^2 + (\alpha_1\beta_2 + \beta_1\alpha_2)\lambda] f_2(x) = 0, \\ &(\beta_1 + \beta_2\lambda) f_2(0) - f_{2x}(0) = 0, \\ &f_2(l) = 0, \\ &f_1(x) = h_2^{-1}\left[(\lambda + h_2) f_2(x) - F_2 f_{2x}(x)\right]. \end{aligned} \tag{9.12}$$

Thus $\lambda \in \sigma(A)$ if and only if (9.12) has a non trivial solution $f_2 \neq 0$.

The two characteristic roots of (9.12) are

$$\mathbf{r_{1,2}} = \frac{1}{2}[\beta_1 - \alpha_1 + (\beta_2 - \alpha_2)\lambda \pm \sqrt{\Delta(\lambda)}], \tag{9.13}$$

where

$$\Delta(\lambda) = [\alpha_1 + \beta_1 + (\alpha_2 + \beta_2)\lambda]^2 - 4\alpha_1\beta_1, \tag{9.14}$$

and $\sqrt{\Delta(\lambda)}$ is the square root of a complex number $\Delta(\lambda)$ defined as being of positive or zero real part (see Appendix for definition). There are two cases to be considered : $r_1 = r_2$ and $r_1 \neq r_2$.

(a) $\mathbf{r_1 = r_2}$ if and only if $\lambda = -\dfrac{(\sqrt{\alpha_1} \pm \sqrt{\beta_1})^2}{\alpha_2 + \beta_2}$. Then the general solution of (9.12) is given by $f_2(x) = c_1 e^{r_1 x} + c_2 x e^{r_1 x}$. The constants c_1 and c_2 are to be determined to satisfy the boundary condition such that

$$\begin{pmatrix} \beta_1 + \beta_2\lambda - r_1 & -1 \\ 1 & l \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Therefore $f_2 \neq 0$ if and only if the determinant of the above matrix $= (\beta_1 + \beta_2\lambda - r_1)l + 1 = 0$. It is easy to see that, for $\lambda = -\dfrac{(\sqrt{\alpha_1} - \sqrt{\beta_1})^2}{\alpha_2 + \beta_2}$, the determinant is equal to $1 + \sqrt{\alpha_1\beta_1} > 0$. Therefore $\lambda = -\dfrac{(\sqrt{\alpha_1} - \sqrt{\beta_1})^2}{\alpha_2 + \beta_2}$ is never a spectrum point of spectrum $\sigma(A)$. For $\lambda = -\dfrac{(\sqrt{\alpha_1} + \sqrt{\beta_1})^2}{\alpha_2 + \beta_2}$ the determinant is equal to $1 - l\sqrt{\alpha_1\beta_1}$. Consequently $\lambda = -\dfrac{(\sqrt{\alpha_1} + \sqrt{\beta_1})^2}{\alpha_2 + \beta_2}$ is an eigenvalue of spectrum $\sigma(A)$ if and only if the physical parameters satisfy the condition that $1 - l\sqrt{\alpha_1\beta_1} = 0$.

(b) $\mathbf{r_1 \neq r_2}$. The solution of (9.12) is $f_2(x) = c_1 e^{r_1 x} + c_2 e^{r_2 x}$ satisfying the boundary conditions :

$$\begin{pmatrix} e^{r_1 l} & e^{r_2 l} \\ \beta_1 + \beta_2\lambda - r_1 & \beta_1 + \beta_2\lambda - r_2 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Hence $\lambda \in \sigma(A)$ if and only if the determinant of the left-hand matrix is zero:

$$(\beta_1 + \beta_2\lambda - r_2)e^{r_1 l} - (\beta_1 + \beta_2\lambda - r_1)e^{r_2 l} = 0,$$

or,

$$\eta \sinh(z) + \frac{2z}{l}\cosh(z) = 0, \tag{9.15}$$

where

$$\eta = \alpha_1 + \beta_1 + (\alpha_2 + \beta_2)\lambda \ \text{ and } z = \frac{l}{2}\sqrt{\Delta(\lambda)}. \tag{9.16}$$

Denote by Σ_0 the set of algebraic solutions $\lambda \in \mathbb{C}$ of (9.15) such that $z \neq 0$:

$$\Sigma_0 = \left\{ \lambda \in \mathbb{C} \;\middle|\; \eta \sinh(z) + \frac{2z}{l}\cosh(z) = 0,\ \Delta(\lambda) \neq 0 \right\}. \tag{9.17}$$

Define also the set S_A by

$$S_A = \left\{ \lambda = -\left(\frac{\alpha_1 + \beta_1}{\alpha_2 + \beta_2}\right) - \frac{2\cosh(\xi)}{al(\alpha_2 + \beta_2)} \;\middle|\; \xi \neq 0, \Re e(\xi) \geq 0,\ \sinh(\xi) = a\xi \right\}, \tag{9.18}$$

where a was defined in (9.10) in Introduction.

Lemma 2. *Assume that* $l\sqrt{\alpha_1\beta_1} \neq 1$. *Then* $\Sigma_0 = S_A = \sigma(A)$.

(The proof of Lemma 2 is postponed to Appendix.) Hence (i) and (ii) are proved. ∎

Set $\xi = x + iy$ ($x \geq 0$ and $y \in \mathbb{R}$) and consider the following equation:

$$\sinh(\xi) = a\xi. \tag{9.19}$$

Consider the real part of each $\lambda \in S_A$:

$$\Re e(\lambda) = -\frac{\alpha_1 + \beta_1}{\alpha_2 + \beta_2} - \frac{2\cos(y)\cosh(x)}{al(\alpha_2 + \beta_2)}, \tag{9.20}$$

where x and y satisfy the algebraic equations equivalent to (9.19):

$$\begin{array}{l} \cos(y)\sinh(x) = ax, \\ \sin(y)\cosh(x) = ay. \end{array} \tag{9.21}$$

Assume that $a^2 \geq 1$. If $x = 0$, (9.21) has no trivial solution for $a^2 \geq 1$. Assuming $x > 0$ and replacing (9.21) into (9.20) leads to the following:

$$\begin{aligned} \Re e(\lambda) &= -\frac{\alpha_1 + \beta_1}{\alpha_2 + \beta_2} - \frac{2x\cosh(x)}{l(\alpha_2 + \beta_2)\sinh(x)} < -\frac{\alpha_1 + \beta_1}{\alpha_2 + \beta_2} - \frac{2}{l(\alpha_2 + \beta_2)} \\ &\leq -\frac{\alpha_1 + \beta_1}{\alpha_2 + \beta_2} - \frac{2\sqrt{\alpha_1\beta_1}}{\alpha_2 + \beta_2} = \frac{-(\sqrt{\alpha_1} + \sqrt{\beta_1})^2}{\alpha_2 + \beta_2} \end{aligned} \tag{9.22}$$

because $|a| \geq 1$ and $\inf_{x>0} \frac{x\cosh(x)}{\sinh(x)} = 1$. So (ii) is proved. ∎

Assume that $a^2 < 1$. We have to consider a finite number of the elements of S_A with $x = 0$:

$$\begin{aligned} \lambda &= -\frac{\alpha_1 + \beta_1}{\alpha_2 + \beta_2} - \frac{2\cos(y)}{al(\alpha_2 + \beta_2)} = -\frac{(\sqrt{\alpha_1} - \sqrt{\beta_1})^2 + 2\sqrt{\alpha_1\beta_1}(1 + a^{-1}|a|\cos(y))}{\alpha_2 + \beta_2} \\ &\leq -\frac{(\sqrt{\alpha_1} - \sqrt{\beta_1})^2 + 2\sqrt{\alpha_1\beta_1}\left(1 + \min\left\{a^{-1}|a|\cos(y) \;\middle|\; \begin{array}{l}\sin(y) = ay \\ a^2 = \frac{1}{\alpha_1\beta_1 l^2}\end{array}\right\}\right)}{\alpha_2 + \beta_2}. \end{aligned}$$

This proves (iv). Therefore, the proof of Lemma 1 is complete. □

From equation (9.21), we observe that if x is bounded, then the equation has only a finite number of solutions $y \in \mathbb{R}$. Therefore we would have only a finite number of eigenvalues. However the operator A being unbounded has a sequence of countable eigenvalues whose module tends to infinity. Hence the real part of the sequence eigenvalues $\lambda \in \sigma(A)$ tend to infinity (see (i) of Lemma 1). On the other hand, if $y = 0$ the equation (9.21) has only one possible solution. Moreover, if y is a solution then $(-y)$ is also a solution. From now on it is sufficient to look at the solution (x, y) with $x, y > 0$.

Indeed, depending on the case where $a > 0$ or $a < 0$, we can go further in the study of solutions of equation (9.21), or, eigenvalues of A.

a) for $a > 0$, the equation (9.21) is equivalent to the following:

$$\cos\left(\frac{\cosh(x)}{a}\sqrt{1 - \left(\frac{ax}{\sinh(x)}\right)^2}\right) = \frac{ax}{\sinh(x)}, \tag{9.23}$$

$$y = \frac{\cosh(x)}{a}\sqrt{1-\Big(\frac{ax}{\sinh(x)}\Big)^2} \in [2k\pi, 2k\pi + \frac{\pi}{2}], \quad k \in \mathbb{N}. \tag{9.24}$$

b) for $a < 0$, the equation (9.21) is equivalent to the following:

$$\cos\left(\frac{\cosh(x)}{a}\sqrt{1-\Big(\frac{ax}{\sinh(x)}\Big)^2}\right) = \frac{ax}{\sinh(x)}, \tag{9.25}$$

$$y = \frac{-\cosh(x)}{a}\sqrt{1-\Big(\frac{ax}{\sinh(x)}\Big)^2} \in [(2k+1)\pi, (2k+1)\pi + \frac{\pi}{2}], \quad k \in \mathbb{N}. \tag{9.26}$$

It is not difficult to see that both (9.23) and (9.25) have countably many solutions with $x > 0$. Indeed, taking $a > 0$, we see that the right-hand term of the equation of (9.23) is positive and strictly decreasing with $x > 0$ and tends to zero for $x \to +\infty$. The left-hand term passes from -1 to $+1$ infinitely many times because the argument of the cosine function is a strictly increasing function of $x > 0$ tending to infinity as $x \to \infty$. Consequently both (9.23) and (9.25) have a countable number of solutions x's which tend to $+\infty$ (analyticity!) as well as y.

We define the subset $\Sigma_\infty \subset \sigma(A)$ as follows:

$$\Sigma_\infty = \Big\{ \lambda = -\frac{\alpha_1+\beta_1}{\alpha_2+\beta_2} - \frac{2\cos(y)\cosh(x)}{al(\alpha_2+\beta_2)} \pm i\frac{2\sin(y)\sinh(x)}{al(\alpha_2+\beta_2)} \;\Big|\; x > 0,\ y > 0 \text{ solving (9.23)} - \text{(9.24) and (9.25)} - \text{(9.26)} \Big\}.$$

Here is an another characterization of the spectrum $\sigma(A)$:

Lemma 3. *Excepting a finite number of them, all the spectrum points $\sigma(A)$ (or, eigenvalues of A) are described by the set Σ_∞. Moreover the real parts of these eigenvalues $\lambda \in \sigma(A)$ tend to $-\infty$ as well as the imaginary parts.*

Proof. Since (x, y) is a solution of (9.23)-(9.24) if and only if $(x, -y)$ is, we consider only the $x > 0$ and $y > 0$. For each $\lambda \in \Sigma_\infty$, from (9.23), $\cos(y)\sinh(x) = ax$. If $a > 0$ and $2k\pi \le y \le 2k\pi + \frac{\pi}{2}$, then

$$\sin(y)\cosh(x) = \sqrt{1-\Big(\frac{ax}{\sinh(x)}\Big)^2}\cosh(x) = ay.$$

The argument is similar for $a < 0$. Hence (x, y) satisfies the equation (9.21) and $\lambda \in \sigma(A)$. As discussed above there are only a finite number of eigenvalues that are not covered by the solutions of equation (9.21).

From the proof of Lemma 1, we get:

$$\Re e(\lambda) = -\frac{\alpha_1+\beta_1}{\alpha_2+\beta_2} - \frac{2x\cosh(x)}{l(\alpha 2+\beta_2)\sinh(x)},$$

$$\Im m(\lambda) = \pm\frac{2y\sinh(x)}{l(\alpha 2+\beta_2)\cosh(x)},$$

which tends to $-\infty$ for $x \to +\infty$. This proves Lemma 3. □

Remark 2. In a very similar way we can understand the location of the solutions $\lambda \in \mathbb{C}$ of the following equation :

$$\eta \sinh(z) - \frac{2z}{l}\cosh(z) = 0,$$

where z and η are defined in (9.16). If $z = x + iy$ and x is bounded, this equation gives only a finite number of solutions λ's. Similarly as in the proof of Lemma 1, the real part of the solution λ goes to $+\infty$ for $x \to +\infty$ because

$$\Re e(\lambda) = -\frac{\alpha_1 + \beta_1}{\alpha_2 + \beta_2} + \frac{2x\cosh(x)}{l(\alpha_2 + \beta_2)\sinh(x)}.$$

This fact will be used in the proof of Lemma 4.

Lemma 4. *For each $\epsilon > 0$ there is a positive real number M such that the following estimates hold:*

$$\begin{aligned} &0 \le \Re e(\sqrt{\Delta(\lambda)}) \le 4\left[(\alpha_1 + \beta_1 + (\alpha_2 + \beta_2)\epsilon\right], \\ &|\Im m(\sqrt{\Delta(\lambda)})| \ge \left(\tfrac{\alpha_2+\beta_2}{2}\right)|\Im m(\lambda)|, \end{aligned}$$

$\forall\ \lambda \in \mathbb{C}$ *satisfying* $\sigma_{max}(A) + \epsilon \le \Re e(\lambda) \le \epsilon$ *and* $|\Im m(\lambda)| \ge M$.

Proof. See Appendix.

Lemma 5. *For each $\epsilon > 0$, the following property holds:*

$$\sup\left\{\|(\lambda - A)^{-1}\|\ \ |\Re e(\lambda) \ge \sigma_{max}(A) + \epsilon\right\} < \infty.$$

Proof of Lemma 5: Since the semigroup e^{tA} is uniformly bounded, it follows from Hille-Yosida's theorem (see p.20, [30]) that, for each $\epsilon > 0$,

$$\|(\lambda I - A)^{-1}\| \le \frac{M}{\epsilon}\ \ \forall\ \Re e(\lambda) \ge \epsilon.$$

It rests only to prove that

$$\sup\left\{\|(\lambda I - A)^{-1}\|\ \ |\sigma_{max}(A) + \epsilon \le \Re e(\lambda) \le \epsilon\right\} < \infty.$$

For each $h > 0$ the set E defined below is a compact set contained in $\rho(A)$ (see Lemma 1):

$$E = \{\lambda\ |\ \sigma_{max}(A) + \epsilon \le \Re e(\lambda) \le \epsilon, |\Im m(\lambda)| \le h\}.$$

Hence, $\sup_{\lambda\in E}\|(\lambda I - A)^{-1}\| < \infty$. Therefore it is sufficient to prove that for some $h > 0$,

$$\sup_{\lambda\in E_h}\{\|(\lambda I - A)^{-1}\|\} < \infty \tag{9.27}$$

where $E_h = \{\lambda\ |\ \sigma_{max}(A) + \epsilon \le \Re e(\lambda) \le \epsilon, |\Im m(\lambda)| \ge h\}$. It amounts to solving the differential equation : $(\lambda I - A)f = g$ for all $g \in H$ and estimating $\|(\lambda - A)^{-1}g\|$:

$$\begin{aligned} &\frac{d}{dx}\begin{pmatrix} f_1(x) \\ f_2(x)\end{pmatrix} = \Lambda\begin{pmatrix} f_1(x) \\ f_2(x)\end{pmatrix} + \begin{pmatrix} \alpha_2 & 0 \\ 0 & -\beta_2\end{pmatrix}\begin{pmatrix} g_1(x) \\ g_2(x)\end{pmatrix} \\ &f_1(0) = f_2(0) = 0, \end{aligned} \tag{9.28}$$

where $\Lambda = \begin{pmatrix} -(\alpha_1+\alpha_2\lambda) & \alpha_1 \\ -\beta_1 & \beta_1+\beta_2\lambda \end{pmatrix}$. The two eigenvalues of Λ are r_1 and r_2 given in (9.13). (We will write simply Δ instead of $\Delta(\lambda)$.) We take a h sufficiently large such that $r_1 \neq r_2$. By direct computations, it is found that

$$e^{\Lambda x} = \begin{pmatrix} \dfrac{(\eta+\sqrt{\Delta})e^{r_2x} + (\sqrt{\Delta}-\eta)e^{r_1x}}{2\sqrt{\Delta}} & \dfrac{\alpha_1(e^{r_1x}-e^{r_2x})}{\sqrt{\Delta}} \\ \dfrac{\beta_1(e^{r_2x}-e^{r_1x})}{\sqrt{\Delta}} & \dfrac{(\eta+\sqrt{\Delta})e^{r_1x} + (\sqrt{\Delta}-\eta)e^{r_2x})}{2\sqrt{\Delta}} \end{pmatrix}.$$

Notice the fact that $4\alpha_1\beta_1 = (\eta+\sqrt{\Delta})(\eta-\sqrt{\Delta})$ that has been used in computing $e^{\Lambda x}$. In particular we get the following:

$$(0\ \ 1)e^{\Lambda l}\begin{pmatrix}0\\1\end{pmatrix} = \frac{e^{\frac{l}{2}[\beta_1-\alpha_1+(\beta_2-\alpha_2)\lambda]}}{\sqrt{\Delta}}\left\{\eta\sinh(z)+\sqrt{\Delta}\cosh(z)\right\}, \tag{9.29}$$

where η and z were defined previously in (9.16). Since $E_h \subset \rho(A)$, $\eta\sinh(z)+\sqrt{\Delta}\cosh(z) \neq 0$ (see (9.15)).

The solution of (9.28) is given by

$$\begin{pmatrix} f_1(x) \\ f_2(x) \end{pmatrix} = (\lambda I - A)^{-1}\begin{pmatrix} g_1 \\ g_2 \end{pmatrix}(x) \tag{9.30}$$

$$= -\left[(0\ \ 1)e^{\Lambda l}\begin{pmatrix}0\\1\end{pmatrix}\right]^{-1}\int_0^l e^{\Lambda\xi}\begin{pmatrix}0&0\\0&1\end{pmatrix}e^{\Lambda(l-\xi)}\begin{pmatrix}\alpha_2&0\\0&-\beta_2\end{pmatrix}\begin{pmatrix}g_1(\xi)\\g_2(\xi)\end{pmatrix}d\xi$$
$$+\int_0^x e^{\Lambda(x-\xi)}\begin{pmatrix}\alpha_2&0\\0&-\beta_2\end{pmatrix}\begin{pmatrix}g_1(\xi)\\g_2(\xi)\end{pmatrix}d\xi.$$

According to Lemma 4, we take a h large enough such that on E_h, $|\Re e(z)|$ is bounded and $|\Im m(\sqrt{\Delta})|$ is proportional to $|\Im m(\lambda)|$. So there exist some constants $h>0$ and $M_1>0$ such that

$$\left|\frac{\eta}{\sqrt{\Delta}}\right| + |\sinh(z)| + |\cosh(z)| \leq M_1, \quad \left|\frac{\sinh(z)}{\sqrt{\Delta}} \pm a\right| \geq \frac{|a|}{2}, \quad \forall\ \lambda \in E_h.$$

Direct elementary computations from (9.29) lead to

$$\left[(0\ \ 1)e^{\Lambda l}\begin{pmatrix}0\\1\end{pmatrix}\right]^{-1} = e^{\frac{l}{2}[\alpha_1-\beta_1+(\alpha_2-\beta_2)\lambda]}\frac{\frac{\eta}{\sqrt{\Delta}}\sinh(z)-\cosh(z)}{\frac{4\alpha_1\beta_1\sinh^2(z)}{\Delta}-1}.$$

For $h>0$ big enough there is some positive constant M_2 such that

$$\sup_{x\in[0,l]}\|e^{\Lambda x}\|_{\mathscr{L}(\mathbb{R}^2)} + \left|\left[(0\ \ 1)e^{\Lambda l}\begin{pmatrix}0\\1\end{pmatrix}\right]^{-1}\right| \leq M_2, \quad \forall\ \lambda \in E_h.$$

Substituting these estimates into the solution (9.30) we get some constante $M_3>0$ such that

$$\left\|\begin{pmatrix}f_1(x)\\f_2(x)\end{pmatrix}\right\|_{\mathbb{R}^2} \leq M_3\int_0^l\left\|\begin{pmatrix}g_1(\xi)\\g_2(\xi)\end{pmatrix}\right\|_{\mathbb{R}^2}d\xi \leq M_3\sqrt{l}\left(\int_0^l\left\|\begin{pmatrix}g_1(\xi)\\g_2(\xi)\end{pmatrix}\right\|^2_{\mathbb{R}^2}d\xi\right)^{\frac{1}{2}}.$$

Therefore,

$$\left\|\begin{pmatrix} f_1 \\ f_2 \end{pmatrix}\right\|_H \leq M_3 l \left\|\begin{pmatrix} g_1 \\ g_2 \end{pmatrix}\right\|_H .$$

This implies (9.27) and so Lemma 4 is proved. □

Proof of Theorem 4: The first part (i) is proved by applying Lemma 4 and Lemma 5. The second part (ii) is proved by applying Lemma 1. □

Remark 3. Theorem 4 gives us a sharp estimate of the exponential decay rate for the semigroup. If only exponential stability is sought, the Lyapunov direct method is the most efficient way to get it. For example, consider the Lyapunov functional $V : H \to \mathbb{R}^+$ such that

$$V(\varphi) = \int_0^l \varphi(x)^\top \begin{bmatrix} e^{-\theta x} & 0 \\ 0 & e^{\theta x} \end{bmatrix} \varphi(x) dx,\ \theta > 0.$$

By differentiating $V(\varphi(\cdot, t))$ along the trajectory of system (9.6) the following inequality holds:

$$\dot{V}(\varphi(\cdot, t)) \leq -\theta \left(\min\{F_1, F_2\} - \theta l \|B_1\|_{\mathscr{L}(\mathbb{R})} e^{\theta l}\right) V(\varphi(\cdot, t)) \ \ \forall\, t \geq 0.$$

By taking θ sufficiently small it follows that

$$\dot{V}(\varphi(\cdot, t)) \leq \frac{-\theta \min\{F_1, F_2\}}{2} V(\varphi(\cdot, t)) \ \ \forall\, t \geq 0.$$

This proves exponential stability of the semigroup with decay rate greater than $\frac{\theta \min\{F_1, F_2\}}{2}$. The interested reader is referred to [38] for more general results concerned with the Lyapunov direct method.

9.3 Stability of Plate-Heat Transmission System

In what follows, we shall investigate the stability properties of a coupled system composed of plate and heat equations in adjacent regions, where the coupling arises from the transmission boundary conditions on the interface between the two regions. In this transmission system, the heat equation acts as a controller, dissipating energy through the interface and influencing the plate equation. Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with smooth boundary $\Gamma \triangleq \partial\emptyset$. Suppose $\overline{\emptyset} = \overline{\emptyset}_1 \cup \overline{\emptyset}_2$, $\emptyset_i$ is non-empty subdomain with smooth boundary $\partial\emptyset_i$, $i = 1, 2$. Denote by $\gamma \neq \emptyset$ the interface between $\emptyset_1$ and $\emptyset_2$ (see Fig. 9.1). We assume that $\partial\Omega_1 = \gamma$, $\partial\Omega_2 = \Gamma \cup \gamma$.

The PDE model is as follows:

$$\begin{cases} w_{tt}(x,t) + \Delta^2 w(x,t) = 0 & \text{in } \Omega_1 \times \mathbb{R}^+, \\ \theta_t(x,t) - \Delta\theta(x,t) = 0 & \text{in } \Omega_2 \times \mathbb{R}^+, \\ \mathscr{B}_1 w(x,t) = 0,\ w_t(x,t) = \partial_{\nu_2}\theta(x,t) & \text{on } \gamma \times \mathbb{R}^+, \\ \mathscr{B}_2 w(x,t) - w(x,t) = \Delta\theta(x,t) & \text{on } \gamma \times \mathbb{R}^+, \\ \theta(x,t) = 0 & \text{on } \Gamma \times \mathbb{R}^+, \\ w(x,0) = w_0(x),\ w_t(x,0) = w_1(x) & \text{in } \emptyset_1, \\ \theta(x,0) = \theta_0(x) & \text{in } \emptyset_2, \end{cases} \tag{9.31}$$

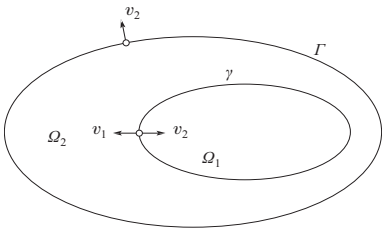


Fig. 9.1: Configuration of domains Ω, Ω_1 and Ω_2

where $\mathscr{B}_1,\ \mathscr{B}_2$ are boundary operators:

$$\begin{cases} \mathscr{B}_1 w = \Delta w + (1-\mu)\left(2\nu_{11}\nu_{12}\dfrac{\partial^2 w}{\partial x_1 \partial x_2} - \nu_{11}^2\dfrac{\partial^2 w}{\partial x_2^2} - \nu_{12}^2\dfrac{\partial^2 w}{\partial x_1^2}\right), \\ \mathscr{B}_2 w = \partial_{\nu_1}\Delta w + (1-\mu)\partial_{\tau_1}\left(\left(\nu_{11}^2 - \nu_{12}^2\right)\dfrac{\partial^2 w}{\partial x_1 \partial x_2} + \nu_{11}\nu_{12}\left(\dfrac{\partial^2 w}{\partial x_2^2} - \dfrac{\partial^2 w}{\partial x_1^2}\right)\right), \end{cases}$$

$\nu_i = (\nu_{i1}, \nu_{i2})$ is the unit outward normal vector of $\partial\Omega_i,\ \tau_i = (-\nu_{i2}, \nu_{i1})$ is the unit tangent vector of $\partial\Omega_i,\ i = 1, 2.\ 0 < \mu < \frac{1}{2}$ is the Poisson ratio. $\alpha,\ \beta$ are nonnegative constants. $\theta_0,\ w_0,\ w_1$ are initial data.

We introduce the following Hilbert space:

$$\begin{aligned} &\mathscr{H} \triangleq H^2(\emptyset_1) \times L^2(\emptyset_1) \times H^1_\Gamma(\emptyset_2), \\ &\|Z\|^2_{\mathscr{H}} = a(w) + \|w\|^2_{L^2(\gamma)} + \|y\|^2_{L^2(\emptyset_1)} + \|\nabla\theta\|^2_{L^2(\emptyset_2)},\ \ \forall\ Z = (w, y, \theta) \in \mathscr{H}. \end{aligned}$$

where $a(w) = a(w, w)$ and

$$\begin{aligned} a(w_1, w_2) = \int_{\Omega_1} \Big[&\frac{\partial^2 w_1}{\partial x_1^2}\overline{\frac{\partial^2 w_2}{\partial x_1^2}} + \frac{\partial^2 w_1}{\partial x_2^2}\overline{\frac{\partial^2 w_2}{\partial x_2^2}} + \mu\Big(\frac{\partial^2 w_1}{\partial x_1^2}\overline{\frac{\partial^2 w_2}{\partial x_2^2}} + \frac{\partial^2 w_1}{\partial x_2^2}\overline{\frac{\partial^2 w_2}{\partial x_1^2}}\Big) + \\ &2(1-\mu)\frac{\partial^2 w_1}{\partial x_1 \partial x_2}\overline{\frac{\partial^2 w_2}{\partial x_1 \partial x_2}}\Big] \mathrm{d}x, \quad \forall w_1, w_2 \in H^2(\Omega_1). \end{aligned}$$

Define an unbounded linear operator $\mathscr{A}$ on $\mathscr{H}$:

$$\begin{aligned} &\mathscr{A} Z = (y, -\mathscr{D}^2 w, \mathscr{D}\theta), \quad \forall\ Z = (w, y, \theta) \in D(\mathscr{A}), \\ &D(\mathscr{A}) = \Big\{ Z = (w, y, \theta) \in \mathscr{H} \,|\, \Delta^2 w \in L^2(\Omega_1),\ y \in H^2(\Omega_1),\ \Delta\theta \in H^1(\Omega_2), \\ &\qquad \theta \in H^2(\Omega_2),\ \mathscr{B}_1 w|_\gamma = 0,\ (\mathscr{B}_2 w - w)|_\gamma = \Delta\theta|_\gamma,\ y|_\gamma = \varphi_{\nu_2}\theta|_\gamma \Big\}. \end{aligned} \tag{9.32}$$

where $H^1_\Gamma(\Omega_2) = \{\theta \in H^1(\Omega_2) \,|\, \theta = 0 \ \text{ on } \Gamma\}$. Then, by letting $Z(t) = (w(\cdot, t), w_t(\cdot, t), \theta(\cdot, t))$, system (9.31) can be rewritten as an evolution equation on $\mathscr{H}$:

$$\dot{Z}(t) = \mathscr{A} Z(t),\ \ Z(0) = Z_0 = (w_0, w_1, \theta_0) \in \mathscr{H}. \tag{9.33}$$

The energy of system (9.31) is

$$E(t) = \frac{1}{2}\|(w(\cdot,t), w_t(\cdot,t), \theta(\cdot,t))\|^2_{\mathscr{H}}. \tag{9.34}$$

It is clear that the energy is non-increasing since

$$\frac{\mathrm{d}}{\mathrm{d}t}E(t) = -\int_{\varnothing_2} |\Delta\theta|^2 dx = -\int_{\varnothing_2} |\theta_t|^2 dx \le 0. \tag{9.35}$$

By the classical method, one can obtain the well-posedness of system (9.31) ([40]) as follows:

Lemma 6. *The operator $\mathscr{A}$ generates a C_0 semigroup of contractions on Hilbert space $\mathscr{H}$, and $0 \in \rho(\mathscr{A})$.*

The coupled system (9.31) is also called transmission system since it consists of partial differential equations defined in different regions and connected through boundary conditions. The stability analysis for transmission systems are closely related to the properties of each partial differential equation, the transmission boundary conditions, the position and type of controllers etc. We refer to [1, 5, 12, 14, 19, 23, 26, 27, 28, 39, 40, 42, 43] for stability and control problems for wave-plate transmission systems, heat-wave transmission systems, wave-viscoelastic wave transmission systems etc.

In [40], 1-D and 2-D plate-heat transmission systems (9.31) were discussed. The paper first analyzed the Riesz basis properties corresponding to the 1-D system (9.31), obtaining exponential stability and Gevrey-type regularity for the 1-D plate-heat system. The paper also studied exponential stability of the 2-D transmission system (9.31), but two conditions were required: 1) one boundary control is needed; 2) regions $\Omega, \Omega_1, \Omega_2$ satisfy appropriate geometric conditions. Therefore, natural questions arise: Is the controller necessary? Is the dissipation effect of the heat equation on Ω_2 sufficient to make the 2-D system exponentially stable? We shall explore these questions, specifically discussing the exponential decay properties of coupled system (9.31).

To analyze the stability of system (9.31), we shall use the following lemmas.

Lemma 7. *([39]) Let $x_0 \in \mathbb{R}^2$, $m(x) \triangleq x - x_0$. Assume $w \in H^2(\varnothing_1)$ satisfying $\mathscr{D}^2 w \in L^2(\varnothing_1)$. Then, we have*

$$\begin{aligned}\int_{\Omega_1} \Delta^2 w(m\cdot\nabla w)dx = a(w) + \int_{\varphi\varnothing_1} \left[\mathscr{B}_2 w\,\overline{(m\cdot\nabla w)} - \mathscr{B}_1 w\,\overline{\partial_{\nu_1}(m\cdot\nabla w)}\right] d\Gamma \\ + \frac{1}{2}\int_{\varphi\varnothing_1} (m\cdot\nu_1) b(w) d\Gamma,\end{aligned} \tag{9.36}$$

where

$$b(w(x)) = \left|\frac{\partial^2 w}{\partial x_1^2}\right|^2 + \left|\frac{\partial^2 w}{\partial x_2^2}\right|^2 + 2\mu\frac{\partial^2 w}{\partial x_1^2}\frac{\partial^2 w}{\partial x_2^2} + 2(1-\mu)\left|\frac{\partial^2 w}{\partial x_1 \partial x_2}\right|^2.$$

The following theorem asserts exponential stability of coupled system (9.31):

Theorem 5. *Let $\varnothing \subset \mathbb{R}^2$ be a bounded domain with partition $\overline{\varnothing} = \overline{\varnothing}_1 \cup \overline{\varnothing}_2$ satisfying $\varnothing_1 \cap \varnothing_2 = \emptyset$, $\partial\Omega_1 = \gamma$ and $\partial\Omega_2 = \Gamma \cup \gamma$. Assume*

$$(m\cdot\nu_1)|_\gamma \ge 0. \tag{H}$$

Then the energy of (9.31) is exponentially stable, i.e., there exist positive constants M, $\o$ such that

$$E(t) \le Me^{-\o t}E(0), \quad \forall\, t \ge 0.$$

Proof. From Theorem 2, it suffices to prove

$$\inf_{\lambda\in\mathbb{R}} \{\|\mathrm{i}lZ - \mathscr{A}Z\|_{\mathscr{H}} \mid Z \in D(\mathscr{A}),\ \|Z\|_{\mathscr{H}} = 1\} \geq r, \quad r > 0. \tag{9.37}$$

Suppose (9.37) is not true. Then there exists $\{\lambda_n,\, Z_n\} \subset R \times D(\mathscr{A})$ with $Z_n = (w_n, y_n, \theta_n)$ and $\|Z_n\|_{\mathscr{H}} = 1$ such that

$$\|(\mathrm{i}l_n I - \mathscr{A})Z_n\|_{\mathscr{H}} \to 0. \tag{9.38}$$

We assume $\lambda_n > 0$ without affecting the results. It follows from (9.38) that as $n \to \infty$,

$$f_{1,n} \triangleq \mathrm{i}l_n w_n - y_n \to 0 \qquad \text{in } \ H^2(\mathcal{O}_1), \tag{9.39}$$

$$f_{2,n} \triangleq \mathrm{i}l_n y_n + \mathscr{D}^2 w_n \to 0 \qquad \text{in } \ L^2(\mathcal{O}_1), \tag{9.40}$$

$$f_{3,n} \triangleq \mathrm{i}l_n \theta_n - \mathscr{D}\theta_n \to 0 \qquad \text{in } \ H^1(\mathcal{O}_2). \tag{9.41}$$

Furthermore, the dissipativeness of $\mathscr{A}$ yields

$$\Re e\langle \mathscr{A}Z_n, Z_n\rangle_{\mathscr{H}} = -\|\Delta\theta_n\|^2_{L^2(\mathcal{O}_2)}.$$

Consequently, one can deduce from (9.38) that

$$\|\Delta\theta_n\|_{L^2(\mathcal{O}_2)}, \ \ \|\lambda_n\theta_n\|_{L^2(\mathcal{O}_2)} \to 0. \tag{9.42}$$

Firstly taking the inner product of (9.41) with θ_n in $L^2(\Omega_2)$,

$$\int_{\Omega_2} (\mathrm{i}\lambda_n|\theta_n|^2 + |\nabla\theta_n|^2)dx - \int_\gamma \partial_{\nu_2}\theta_n\, \overline{\theta_n} d\Gamma \to 0,$$

and then substituting the boundary conditions on the interface, (9.39) and (9.42) into the above equation, we get the following:

$$\begin{aligned} \lim_{n\to\infty} \|\nabla\theta_n\|^2_{L^2(\mathcal{O}_2)} &\leq \lim_{n\to\infty} \|y_n\|_{H^{\frac{1}{2}}(\gamma)} \|\theta_n\|_{H^{-\frac{1}{2}}(\gamma)} \\ &\leq \lim_{n\to\infty} \lambda_n \|w_n\|_{H^1(\Omega_2)} \|\theta_n\|_{L^2(\Omega_2)} \\ &= 0. \end{aligned} \tag{9.43}$$

Combining $\theta_n|_\Gamma = 0$, (9.42) and (9.43), we obtain

$$\|\theta_n\|_{H^2(\mathcal{O}_2)} \to 0. \tag{9.44}$$

From (9.39)-(9.40), the variables w_n satisfy

$$\begin{cases} -l_n^2 w_n + \mathscr{D}^2 w_n = \mathrm{i}\lambda_n f_{1,n} + f_{2,n} & \text{in } \ \Omega_1, \\ \mathscr{B}_1 w_n = 0, \ \ \mathscr{B}_2 w_n - w_n = \Delta\theta_n, \ \ \mathrm{i}\lambda_n w_n = \partial_{\nu_2}\theta_n + f_{1,n} & \text{on } \ \gamma. \end{cases} \tag{9.45}$$

We introduce the linear operator:

$$\mathbb{G}\phi = \varphi \ \Leftrightarrow \ \begin{cases} \mathscr{D}^2\varphi = 0 & \text{in } \ \Omega_1, \\ \mathscr{B}_1\varphi = 0, \ \ \mathscr{B}_2\varphi - \varphi = \phi & \text{on } \ \gamma. \end{cases} \tag{9.46}$$

Due to the theory of regularity ([2, 22]),

$$\mathbb{G} \in \mathscr{L}(H^s(\gamma),\, H^{s+\frac{7}{2}}(\Omega_1)), \quad s \in \mathbb{R}. \tag{9.47}$$

we can define

$$\phi_n \triangleq (\mathscr{B}_2 w_n - w_n)|_\gamma, \quad u_n \triangleq w_n - \mathbb{G}\phi_n. \tag{9.48}$$

Therefore, variables u_n satisfy

$$\begin{cases} -l_n^2 u_n + \mathscr{D}^2 u_n = V_n & \text{in } \ \Omega_1, \\ \mathscr{B}_1 u_n = \mathscr{B}_2 u_n - u_n = 0 & \text{on } \ \gamma, \\ \mathrm{i}\lambda_n u_n = -\mathrm{i}\lambda_n \mathbb{G}\phi_n + \partial_{\nu_2}\theta_n + f_{1,n} & \text{on } \ \gamma. \end{cases} \tag{9.49}$$

where

$$V_n \triangleq l_n^2 \mathbb{G}\phi_n + \mathrm{i}\lambda_n f_{1,n} + f_{2,n}.$$

We shall prove that the sequence of variables Z_n satisfies the following

$$\|Z_n\|_{\mathscr{H}} \to 0 \ \text{ as } n \to \infty, \tag{9.50}$$

that contradicts the fact that $\|Z_n\|_{\mathscr{H}} = 1$, thus the proof of the theorem will be complete.

For the purpose we shall apply the following lemma whose proof is postponed to Appendix.

Lemma 8. *As $n \to \infty$, there exist positive constants C_1 and C_2 such that*

$$\begin{aligned} &a(u_n) + 2\lambda_n^2 \int_{\Omega_1} |u_n|^2 dx \\ &\le 2\Re e \int_{\Omega_1} V_n \overline{(m \cdot \nabla u_n)} dx + 2(C_1\lambda_n^2 + C_2)\|u_n\|^2_{L^2(\gamma)}, \end{aligned} \tag{9.51}$$

and

$$\lambda_n \|\mathbb{G}\phi_n\|_{H^1(\Omega_1)} \to 0, \tag{9.52}$$

$$\lambda_n \|u_n\|_{L^2(\gamma)} \to 0, \tag{9.53}$$

$$\Re e \int_{\Omega_1} V_n \overline{(m \cdot \nabla u_n)} dx \to 0. \tag{9.54}$$

Substituting (9.53) and (9.54) into (9.51) yields

$$a(u_n), \ \ \lambda_n \|u_n\|_{L^2(\Omega_1)} \to 0. \tag{9.55}$$

From (9.39), (9.48), (9.52) and (9.55), we have

$$\|y_n\|_{L^2(\Omega_1)} \to 0. \tag{9.56}$$

Taking the $L^2(\Omega)$ inner product of (9.39) with y_n and (9.40) with w_n, respectively, adding the results up, one gets.

$$a(w_n) + \|w_n\|^2_{L^2(\gamma)} - \|y_n\|^2_{L^2(\Omega_1)} = -\Re e \int_\gamma \Delta\theta_n \overline{(\mathrm{i}\lambda_n)^{-1}(\partial_{\nu_2}\theta_n + f_{1,n})} d\Gamma + o(1). \tag{9.57}$$

Obviously, by (9.41), (9.44) and the trace theorem, we can obtain

$$\begin{aligned} &\lim_{n\to\infty} \|\lambda_n^{-1}\Delta\theta_n\|_{L^2(\gamma)} = \lim_{n\to\infty} \|\theta_n\|_{L^2(\gamma)} = 0, \\ &\lim_{n\to\infty} \|\partial_{\nu_2}\theta_n\|_{L^2(\gamma)} = 0. \end{aligned} \tag{9.58}$$

Substituting (9.58) into (9.57), we obtain: as $n \to \infty$,

$$a(w_n) + \|w_n\|^2_{L^2(\gamma)} - \|y_n\|^2_{L^2(\Omega_1)} \to 0. \tag{9.59}$$

Finally (9.50) is proved by using (9.44), (9.56) and (9.59). Hence the proof of Theorem 5 is complete. □

Remark 4. In the plate-heat transmission system (9.31), the control is applied to the subregion Ω_2. It can be easily verified that the regions $\Omega, \Omega_1, \Omega_2$ satisfying the assumptions of Theorem 5 also satisfy the geometric control conditions [3]. From the above discussion it can be inferred that both the type of controller and the geometric properties of the control region may affect the stability properties of the transmission system.

In Theorem 5, exponential stability of the transmission system (9.31) has been proved. When the coupling conditions on the interface γ change, one can still investigate the exponential stability property by similar arguments. For example, the following model is exponentially stable (see [41]):

$$\begin{cases} w_{tt} + \Delta^2 w = 0, & (x,t) \in \Omega_1 \times \mathbb{R}^+, \\ \theta_t - \Delta\theta = 0, & (x,t) \in \Omega_2 \times \mathbb{R}^+, \\ \mathscr{B}_1 w = 0,\ w_t = \theta,\ \mathscr{B}_2 w - w = \partial_{\nu_2}\theta, & (x,t) \in \gamma \times \mathbb{R}^+, \\ \theta = 0, & (x,t) \in \Gamma \times \mathbb{R}^+, \\ w(0) = w_0,\ w_t(0) = w_1, & x \in \Omega_1, \\ \theta(0) = \theta_0, & x \in \Omega_2, \end{cases} \tag{9.60}$$

where the 2-D region Ω and its subregions Ω_1, Ω_2 all satisfy the relevant assumptions in Theorem 5.

9.4 Appendix

Proof of Lemma 2: As $l\sqrt{\alpha_1\beta_1} \neq 0$, we consider only $\lambda \in \mathbb{C}$ such that $\Delta(\lambda) \neq 0$. Recall that

$$\Sigma_0 = \left\{ \lambda \in \mathbb{C} \;\middle|\; \eta \sinh(z) + \frac{2z}{l}\cosh(z) = 0,\ \Delta(\lambda) \neq 0 \right\},$$

$$S_A = \left\{ \lambda = -\left(\frac{\alpha_1+\beta_1}{\alpha_2+\beta_2}\right) - \frac{2\cosh(\xi)}{al(\alpha_2+\beta_2)} \;\middle|\; \xi \neq 0, \Re e(\xi) \geq 0,\ \sinh(\xi) = a\xi \right\}.$$

For each $\lambda \in \Sigma_0$, the real part of z by (9.16) is positive or zero, and λ satisfies the equation:

$$\eta^2 \sinh^2(z) = \frac{4z^2}{l^2} \cosh^2(z). \tag{9.61}$$

Using the fact that $\eta^2 - \Delta(\lambda) = 4\alpha_1\beta_1$ and $\cosh^2(z) - \sinh^2(z) = 1$, the equation (9.61) is equivalent to the following:

$$\sinh(z) = az. \tag{9.62}$$

From (9.15) and (9.62), we have:

$$\eta = -\frac{2z\cosh(z)}{l\sinh(z)} = -\frac{2\cosh(z)}{al}, \quad \lambda = -\frac{\alpha_1+\beta_1}{\alpha_2+\beta_2} - \frac{2\cosh(z)}{al(\alpha_2+\beta_2)}.$$

Hence $\lambda \in S_A$. It is meant that $\Sigma_0 \subset S_A$.

Now we prove the converse: $S_A \subset \Sigma_0$. It is sufficient to prove that each $\lambda \in S_A$ does satisfy (9.15). It is clear that

$$\eta = \alpha_1 + \beta_1 + (\alpha_2+\beta_2)\lambda = \frac{-2\cosh(\xi)}{al},$$
$$z = \frac{l}{2}\sqrt{\Delta(\lambda)} = \frac{l}{2}\sqrt{\eta^2 - 4\alpha_1\beta_1} = l\sqrt{\alpha_1\beta_1}|a|\xi = \xi.$$

It follows that

$$\eta\sinh(z) + \frac{2z}{l}\cosh(z) = \frac{-2}{al}\cosh(\xi)\sinh(\xi) + \frac{2\xi}{l}\cosh(\xi) = 0.$$

Hence $\lambda \in \Sigma_0$ and $S_A \subset \Sigma_0$. The proof of Lemma 2 is complete. □

Proof of Lemma 4: Recall the formula of complex square root: for each $(X, Y) \in \mathbb{R}^2$,

$$\sqrt{X+iY} = \left(\frac{X+\sqrt{X^2+Y^2}}{2}\right)^{\frac{1}{2}} + i\,\mathrm{sign}(Y)\left(\frac{-X+\sqrt{X^2+Y^2}}{2}\right)^{\frac{1}{2}}, \tag{9.63}$$

where $\mathrm{sign}(Y) = 1$ if $Y \geq 0$ and $\mathrm{sign}(Y) = -1$ if $Y < 0$.

By setting $\lambda = x + iy$ with $x, y \in \mathbb{R}$ and from the definition of $\Delta(\lambda)$ (see (9.14)) it is easy to see that

$$X = \Re e(\Delta(\lambda)) = (\alpha_1+\beta_1+(\alpha_2+\beta_2)x)^2 - (\alpha_2+\beta_2)^2y^2 - 4\alpha_1\beta_1$$
$$Y = \Im m(\Delta(\lambda)) = 2\,(\alpha_1+\beta_1+(\alpha_2+\beta_2)x)\,(\alpha_2+\beta_2)y.$$

For any $\sigma_{max}(A) + \epsilon \leq x \leq \epsilon$, we have

$$\sqrt{X^2+Y^2} - X \geq -X = y^2\left\{(\alpha_2+\beta_2)^2 + \frac{4\alpha_1\beta_1 - [\alpha_1+\beta_1+(\alpha_2+\beta_2)x]^2}{y^2}\right\}$$
$$\geq \frac{y^2(\alpha_2+\beta_2)^2}{4} \tag{9.64}$$

provided that

$$|y| \geq M = \left(\sup_{x\in[\sigma_{max}(A)+\epsilon,\epsilon]}\left|\frac{4\alpha_1\beta_1 - [\alpha_1+\beta_1+(\alpha_2+\beta_2)x]^2}{3(\alpha_2+\beta_2)^2}\right|\right)^{1/2}.$$

The first inequality is proved by using (9.63) and (9.64).

The second inequality results from the first one as follows:

$$\left(\sqrt{X^2+Y^2}+X\right)^{1/2} = \frac{|Y|}{\sqrt{\sqrt{X^2+Y^2}-X}}$$

$$\leq \frac{4(\alpha_1+\beta_1+(\alpha_2+\beta_2)x)(\alpha_2+\beta_2)|y|}{(\alpha_2+\beta_2)|y|} \leq 4[\alpha_1+\beta_1+(\alpha_2+\beta_2)\epsilon].$$

The proof of Lemma 4 is complete. □

Proof of Lemma 8: By Green formula and (9.36),

$$\begin{aligned}\int_{\Omega_1} \Delta^2 u_n \overline{(m\cdot\nabla u_n)}dx = a(u_n) + \int_\gamma \mathscr{B}_2 u_n \overline{(m\cdot\nabla u_n)}d\Gamma \\ +\frac{1}{2}\int_\gamma (m\cdot\nu_1)b(u_n)d\Gamma,\end{aligned} \tag{9.65}$$

and

$$-\lambda_n^2 \Re e\int_{\Omega_1} u_n\overline{(m\cdot\nabla u_n)}dx = \frac{-\lambda_n^2}{2}\int_\gamma (m\cdot\nu_1)|u_n|^2 d\Gamma + \lambda_n^2\int_{\Omega_1}|u_n|^2dx. \tag{9.66}$$

Then it follows from (9.49), (9.65) and (9.66) that

$$\begin{aligned}a(u_n)+\lambda_n^2\int_{\Omega_1}|u_n|^2dx = \Re e\int_{\Omega_1} V_n\overline{(m\cdot\nabla u_n)}dx + \frac{1}{2}\lambda_n^2\int_\gamma (m\cdot\nu_1)|u_n|^2d\Gamma \\ -\Re e\int_\gamma u_n\overline{(m\cdot\nabla u_n)}d\Gamma - \frac{1}{2}\int_\gamma (m\cdot\nu_1)b(u_n)d\Gamma.\end{aligned}$$

By Cauchy-Schwarz inequlity, there exist positive constants C_1, C_2 such that

$$\begin{aligned}a(u_n)+2\lambda_n^2\int_{\Omega_1}|u_n|^2dx \leq 2\Re e\int_{\Omega_1} V_n\overline{(m\cdot\nabla u_n)}dx + 2(C_1\lambda_n^2+C_2)\int_\gamma |u_n|^2d\Gamma \\ -\int_\gamma (m\cdot\nu_1)b(u_n)d\Gamma,\end{aligned} \tag{9.67}$$

where $C_1 \triangleq \max\{|m| : x\in\gamma\}$, $C_2 = C_1^2C_3^2$ with C_3 being the constant in $\|u\|_{H^1(\gamma)} \leq C_3\|u\|_{H^2(\Omega)}$. Therefore, (9.51) is proved by using (9.67), (H) and

$$b(u_n) \geq (1-\mu)\left(\left|\frac{\partial^2 u_n}{\partial x_1^2}\right|^2 + \left|\frac{\partial^2 u_n}{\partial x_2^2}\right|^2 + 2\left|\frac{\partial^2 u_n}{\partial x_1\partial x_2}\right|^2\right).$$

(ii) From the boundary conditions on the interface, we have

$$\|\mathbb{G}\phi_n\|_{H^1(\Omega_1)} \leq C\|\Delta\theta_n\|_{H^{-\frac{5}{2}}(\gamma)} \leq C\|\theta_n\|_{L^2(\Omega_1)}. \tag{9.68}$$

Furthermore, from (9.42), it can be seen that as $n\to\infty$,

$$\lambda_n\|\theta_n\|_{L^2(\Omega_1)} \le \|\Delta\theta_n\|_{L^2(\Omega_1)} \to 0. \tag{9.69}$$

Therefore, combining the above two equations yields equation (9.52).

(iii) It follows from (9.49) that

$$\begin{aligned}\lambda_n\|u_n\|_{L^2(\gamma)} &\le \|\lambda_n\mathbb{G}\phi_n\|_{L^2(\gamma)} + \|\partial_{\nu_2}\theta_n\|_{L^2(\gamma)} + \|f_{1,n}\|_{L^2(\gamma)} \\ &\le \lambda_n\|\mathbb{G}\phi_n\|_{H^1(\Omega_1)} + \|\theta_n\|_{H^2(\Omega_1)} + \|f_{1,n}\|_{H^1(\Omega_1)}.\end{aligned} \tag{9.70}$$

Substituting (9.39), (9.44) and (9.52) into the above inequality yields (9.53).

(iv) From definition of V_n and u_n, one can get

$$\int_{\Omega_1} V_n\overline{(m\cdot\nabla u_n)}dx = \Re e\int_{\Omega_1}(l_n^2\mathbb{G}\phi_n + \mathrm{i}\lambda_n f_{1,n} + f_{2,n})\overline{(m\cdot\nabla(w_n - \mathbb{G}\phi_n))}dx. \tag{9.71}$$

Due to (9.39), (9.52) and Cauchy-Schwarz inequality,

$$\lim_{n\to\infty}\lambda_n^2|(\mathbb{G}\phi_n,\ m\cdot\nabla w_n)_{L^2(\Omega_1)}| \le \lim_{n\to\infty}\|\lambda_n\mathbb{G}\phi_n\|_{H^1(\Omega_1)}\|\lambda_n w_n\|_{L^2(\Omega_1)} = 0. \tag{9.72}$$

Similarly, as $n\to\infty$,

$$\lambda_n^2|(\mathbb{G}\phi_n,\ m\cdot\nabla\mathbb{G}\phi_n)_{L^2(\Omega_1)}| \le C\lambda_n^2\|\mathbb{G}\phi_n\|^2_{H^1(\Omega_1)} = 0, \tag{9.73}$$

$$\lim_{n\to\infty}\lambda_n|(f_{1,n},\ m\cdot\nabla w_n)_{L^2(\Omega_1)}| \le C\lim_{n\to\infty}\|f_{1,n}\|_{H^1(\Omega_1)}\|y_n\|_{L^2(\Omega_1)} = 0, \tag{9.74}$$

$$\lim_{n\to\infty}\lambda_n|(f_{1,n},\ m\cdot\nabla\mathbb{G}\phi_n)_{L^2(\Omega_1)}| \le C\lim_{n\to\infty}\|f_{1,n}\|_{H^1(\Omega_1)}\|\lambda_n\mathbb{G}\phi_n\|_{H^1(\Omega_1)} = 0, \tag{9.75}$$

and

$$\begin{aligned}&\lim_{n\to\infty}|(f_{2,n},\ m\cdot\nabla(w_n - \mathbb{G}\phi_n))_{L^2(\Omega_1)}| \\ &\le C\lim_{n\to\infty}\|f_{2,n}\|_{L^2(\Omega_1)}(\|w_n\|_{H^2(\Omega_1)} + \|\mathbb{G}\phi_n\|_{H^1(\Omega_1)}) = 0.\end{aligned} \tag{9.76}$$

In summary, one can obtain (9.54) by substituting (9.72)–(9.76) into (9.71). □

Acknowledgment

This work was supported by the National Natural Science Foundation of China (grants No.12271035, 12131008) and Beijing Municipal Natural Science Foundation (grant No.1232018).

References

1. K. Ammari and S. Nicaise, Stabilization of a transmission wave/plate equation, *J. Differential Equations*, **249** (2010), 707-727.
2. G. Avalos and I. Lasiecka, Exponential stability of a thermoelastic system with free boundary conditions without mechanical dissipation, *SIAM J. Math. Anal.*, **29** (1998), No.1, 155-182.
3. C. Bardos, G. Lebeau, and J. Rauch, Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary, *SIAM J. Control Optim.*, **30** (1992), 1024-1065.

4. G. Bastin and J.M. Coron, *Stability and boundary stabilization of 1D hyperbolic systems*, Springer International Publishing Switzerland, 2016.
5. C.J.K. Batty, L. Paunonen, and D. Seifert, Optimal energy decay for the wave-heat system on a rectangular domain, *SIAM J. Math. Anal.*, **51** (2018), No.2, 808-819.
6. C. Castro and E. Zuazua, Exact boundary controllability of two Euler-Bernoulli beams connected by a point mass, *Mathematical and Computer Modelling*, **32** (2000), 955-969.
7. G. Chen, S.G. Krantz, D.W. Ma, C.E. Wayne and H.H. West, The Euler-Bernoulli beam equation with boundary energy dissipation, *Operator Method For Control Problem*, S.J. Lee (Ed.), Lecture Notes in Pure and Applied Mathematics, Marcell-dekker Inc, New York, 1987, pp.67-96.
8. J.C. Friedly, *Dynamic behavior of processes*, Prentice-Hall Inc.
9. J.P. Gauthier and C.Z. Xu, H^∞-control of a distributed parameter system with non-minimum phase, *Int. J. Control*, 1991, vol.53, pp.45-79.
10. D. Gilbarg and N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin Heidelberg, 2001.
11. P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, London, 1985.
12. Y. Guo, J. Wang and D. Zhao, Energy decay estimates for a two-dimensional coupled wave-plate system with localized frictional damping, *Z. Angew. Math. Mech.*, **100**(2019), No.2, 201900030.
13. B. Guo and C.-Z. Xu, On the spectrum-determined growth condition of a vibration cable with a tip mass, *IEEE Transactions on Automatic Control* 45 (2000), pp.89-93.
14. F. Hassine, Logarithmic stabilization of the Euler-Bernoulli transmission plate equation with locally distributed Kelvin-Voigt damping, *J. Math. Anal. Appl.*, **455**(2017), 1765-1782.
15. F.L. Huang, Characteristic conditions for exponential stability of linear dynamical systems in Hilbert spaces, *Ann. Differential Equations*, 1985, pp.43-56.
16. T. Kato, *Perturbation theory for linear operators*, Springer-Verlag 1976.
17. V. Komornik, *Exact Controllability and Stabilization: The Multiplier Method*, Research in Applied Mathematics, V.36, Wiley-Masson, 1994.
18. J.E. Lagnese, *Boundary Stabilization of Thin Plates*, Philadelphia: SIAM, 1989.
19. I. Lasiecka, *Mathematical Control Theory of Coupled PDEs*, Philadelphia: SIAM, 2002.
20. T.T. Li, *Global classical solutions for quasilinear hyperbolic systems*, Research in Applied Mathematics, J Wiley/Masson, 1994.
21. T.T. Li and B. Rao, Criteria of Kalman's type to the approximate controllability and the approximate synchronization for a coupled system of wave equations with Dirichlet boundary controls, *SIAM J. Control Optim.*, **54** (2016), 49-72.
22. J.L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, V.I, Springer-Verlag, New York, 1972.
23. K. Liu and B. Rao, Exponential stability for the wave equation with local Kelvin-Voigt damping, *Z. angew. Math. Phys.*, **57** (2006), No.3, 419-432.
24. Z. Liu, B. Rao and Q. Zhang, Polynomial stability of the Rao-Nakra beam with a single internal viscous damping, *J. Differential Equations*, **269** (2020), No.7, 6125-6162.
25. Z. Liu, S.A. Trogdon and J. Yong, Modeling and analysis of a laminated beam, *Math. Comput. Modeling*, **30** (1999), No.1-2, 149-167.
26. Z. Liu and Q. Zhang, Stability of a string with local Kelvin-Voigt damping and non-smooth coeficient at interface, *SIAM J. Control Optim.*, **54** (2016), No. 4, 1859-1871.
27. Z. Liu and B. Rao, Characterization of polynomial decay rate for the solution of linear evolution equation, *Z. Angew. Math. Phys.*, **56** (2005), 630-644.
28. K. Liu, Z. Liu and Q. Zhang, Eventual differentiability of a string with local Kelvin-Voigt damping, *ESAIM Control Optim. Calculus Var.*, **23** (2017), No. 2, 443-454.
29. A. Osses and J.-P. Puel, Approximate controllability for a linear model of fluid structure interaction, *ESAIM Control Optim. Calc. Var.* , **4** (1999), 497-513.
30. A. Pazy, *Semigroup of linear operators and applications to partial differential equations*, Springer-Verlag, New York, 1983.

31. J. Prüss, On the spectrum of C_0-semigroups, *Trans. Amer. Math. Soc.*, **284** (1984), 847-857.
32. B. Rao, Stabilization of elastic plates with dynamical boundary control, *SIAM J. Control Optim.*, **36** (1998), 148-163.
33. J.-M. Wang and B.-Z. Guo, Analyticity and dynamic behavior of a three-layer damped sandwich beam, *J. Optim. Theory Appl.*, **137** (2008), 675-689.
34. J.-M. Wang, B. Ren and M. Krstic, Stabilization and Gevrey regularity of a Schrödinger equation in boundary feedback with a heat equation, *IEEE Trans. Automat. Control*, **57** (2012), 179-185.
35. G.G. Xu, S.P. Yung and L.K. Li, Stabilization of wave systems with input delay in the boundary control, *ESAIM Control Optim. Calc. Var.*, **12** (2006), 770-785.
36. C.-Z. Xu and J.-P. Gauthier, Analyse et commande d'un échangeur thermique à contre-courant, *RAIRO APPII* 25, 1991, 377-396.
37. C.-Z. Xu, J.-P. Gauthier and I. Kupka, Exponential stability of the heat exchanger equation, *The Proceedings of European Control Conference 1993*, pp.303-307, 1993, Groningen, The Netherlands.
38. C.-Z. Xu and G. Sallet, Exponential stability and transfer functions of processes governed by symmetric hyperbolic systems, *ESAIM: Control, Optimisation and Calculus of Variations* 7, pp.421-442, 2002.
39. Q. Zhang, Polynomial decay of an elastic/viscoelastic waves interaction system, *Z. Angew. Math. Phys.*, **69** (2018), 88.
40. Q. Zhang, J.-M. Wang and B.-Z. Guo, Stabilization of the Euler-Bernoulli equation via boundary connection with heat equation, *Math. Control. Sign. Syst.*, **26** (2014), 77-118.
41. Q. Zhang, Stability analysis of an interaction system coupling plate equation and heat equation, *Control Theory and Application*, **39** (2022), N.9, 1587-1593.
42. X. Zhang and E. Zuazua, Polynomial decay and control of a 1d hyperbolic-parabolic coupled system, *J. Differential Equations*, **204** (2004), 380-438.
43. X. Zhang and E. Zuazua, Asymptotic behavior of a hyperbolic-parabolic coupled system arising in fluid-structure interaction, *Internat. Ser. Numer. Math.*, **154** (2006), 445-455.

10

Numerical Tools for Geometric Optimal Control and the Julia control-toolbox Package

Olivier Cots[1] and Joseph Gergaud[2]

[1] Institut de Recherche en Informatique de Toulouse, UMR CNRS 5505, Université de Toulouse, INP-ENSEEIHT, France. olivier.cots@irit.fr
[2] Institut de Recherche en Informatique de Toulouse, UMR CNRS 5505, Université de Toulouse, INP-ENSEEIHT, France. joseph.gergaud@irit.fr

Summary. We present in this article numerical techniques — the Julia `control-toolbox` package — for computing geometric optimal control concepts: Hamiltonian flows associated to the optimal control problem, Jacobi flows, Poisson brackets of Hamiltonians to define the singular control associated to a singular arc, etc. With these tools, it becomes easy to solve an optimal control problem by indirect methods, and to compute conjugate times together with the cut locus. We present the numerical tools on two test bed examples: the surface of revolution of minimum area and the Goddard problem. The first problem comes from calculus of variations and thus is regular. We also compute the conjugate locus for this example. On the other hand, the optimal solution of the Goddard problem contains bang and singular arcs.

10.1 Introduction

When we aim to solve an optimal control problem via the indirect methods, first we have to apply the Pontryagin Maximum Principle [33]. Then, by numerical integration of the underlying Hamiltonian, we obtain the flow of this Hamiltonian system. For obtaining this flow with our Julia `control-toolbox` package, we only have to define the optimal control problem and to give the function which computes the control with respect to the state and the costate (obtained by solving analytically the maximization of the pseudo-Hamiltonian); then, the flow is automatically computed thanks to automatic differentiation. Next, it is easy to define the shooting function and to compute extremals. Still with the use of automatic differentiation, we can compute Jacobi fields and conjugate points in relation with second-order conditions of local optimality. We present the use of our Julia `control-toolbox` package on two test bed examples: the surface of revolution of minimum area from calculus of variations and the well-known Goddard problem.

The article is organized as follows. Section 10.2 is devoted to the theory of geometric optimal control: singular control, weak principle and conjugate point, maximum principle and the Hamiltonian frame, problems affine in the (scalar) control. In Section 10.3, we present the simple and multiple indirect shooting methods but also the differential homotopy methods in the frame of geometric control. Finally, in Section 10.4, we introduce the Julia `control-toolbox` package and demonstrate how to use it on the two examples.

Remark 1. In the spirit of the reproducible research, the reader will find at the github repository `https://github.com/control-toolbox/Kupka`, Julia notebooks for the two numerical examples.

10.2 Geometric Optimal Control

This section is inspired by references about optimal control theory and more specifically by references about geometric control. We refer to [1, 9, 13, 14, 27, 28, 29, 31, 33, 34, 36, 37, 39, 40] for more details.

10.2.1 Singular control

We consider a $\mathscr{C}^1$ mapping

$$\begin{aligned} f\colon \mathbb{R}^n \times \mathbb{R}^m &\longrightarrow \mathbb{R}^n \\ (x,u) &\longmapsto f(x,u) \end{aligned}$$

and define the controlled dynamical system

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

A control law $t \mapsto u(t)$ is an essentially bounded mapping defined on an interval of the form $[0\,,\tau_u)$, with $\tau_u \in \mathbb{R}_+^* \cup \{+\infty\}$, and taking values in $\mathbb{R}^m$. We note the set of control laws

$$\mathscr{U} \coloneqq \left\{ u \in L^\infty([0\,,\tau_u),\mathbb{R}^m) \;\middle|\; \tau_u \in \mathbb{R}_+^* \cup \{+\infty\} \right\}.$$

For any pair $(x_0,u) \in \mathbb{R}^n \times \mathscr{U}$, there exists a unique maximal solution (in Carathéodory sense) of the Cauchy problem $\dot{x}(t) = f(x(t),u(t))$, $x(0) = x_0$. We denote by

$$t \mapsto x(t,x_0,u)$$

this solution and $I(x_0,u)$ its interval of definition. Let $t \geq 0$ and $x_0 \in \mathbb{R}^n$ be fixed. We introduce the set $\mathscr{U}_{t,x_0} \subset \mathscr{U}$ of admissible control laws over $[0\,,t]$ with initial condition x_0, as the subset of control laws $u \in \mathscr{U}_{t,x_0}$ such that $x(\cdot,x_0,u)$ is well defined over $[0\,,t]$. With this notations, the mapping $(t,x_0,u) \mapsto x(t,x_0,u)$ is defined on the set:

$$\mathscr{D} \coloneqq \{(t,x_0,u) \mid t \geq 0,\ x_0 \in \mathbb{R}^n,\ u \in \mathscr{U}_{t,x_0}\}.$$

We introduce the two following partial mappings which are of crucial interest. First, we define the flow mapping at time $t \geq 0$ by:

$$\begin{aligned} \Phi_t\colon \quad \mathscr{D}_t &\longrightarrow \mathbb{R}^n \\ (x_0,u) &\longmapsto \Phi_t(x_0,u) \coloneqq x(t,x_0,u), \end{aligned}$$

with $\mathscr{D}_t \coloneqq \{(x_0,u) \mid (t,x_0,u) \in \mathscr{D}\}$. Then, we introduce the *endpoint* mapping at time $t \geq 0$ from $x_0 \in \mathbb{R}^n$ as:

$$\begin{aligned} E_{t,x_0}\colon \mathscr{U}_{t,x_0} &\longrightarrow \mathbb{R}^n \\ u &\longmapsto E_{t,x_0} \coloneqq x(t,x_0,u). \end{aligned}$$

Let $t \geq 0$ and $x_0 \in \mathbb{R}^n$. Then, $\mathscr{U}_{t,x_0}$ is an open subset of $L^\infty([0\,,t],\mathbb{R}^n)$ and the endpoint mapping is of class $\mathscr{C}^1$. The *reachable set* at time $t \geq 0$ from $x_0 \in \mathbb{R}^n$ is defined as $\mathscr{A}(t,x_0) \coloneqq E_{t,x_0}(\mathscr{U}_{t,x_0})$. Let $t > 0$ be a fixed positive time. We say that the system (Σ_u) is *controllable* from $x_0 \in \mathbb{R}^n$ in time t if $\mathscr{A}(t,x_0) = \mathbb{R}^n$, and is *locally controllable* from x_0 in time t around $x_1 \in \mathbb{R}^n$ if x_1 belongs to the interior of $\mathscr{A}(t,x_0)$, that is if $x_1 \in \mathrm{Int}(\mathscr{A}(t,x_0))$. We can notice that if E_{t,x_0} is surjective then (Σ_u) is controllable and if its Fréchet differential $E'_{t,x_0}(u)$ is surjective then, by the following nonlinear open mapping theorem, the system is locally controllable around $E_{t,x_0}(u)$.

Theorem 1. *Let $F\colon U \subset E \to \mathbb{R}^n$ be a function of class $\mathscr{C}^1$ on U, defined on the open set U of a Banach space E. Let $x \in U$ be a regular point of F (i.e. $F'(x)$ surjective), then F is locally open at x.*

A control $u \in \mathscr{U}_{t,x_0}$ is said to be *regular* if $E'_{t,x_0}(u)$ is surjective. Otherwise, it is called *singular*. Hence, if u is regular, then, the system (Σ_u) is locally controllable. By contraposition, if $E_{t,x_0}(u)$ belongs to the boundary of the reachable set, then, the control u is singular. We are now in position to introduce the (pseudo-)Hamiltonian characterization of the singular controls. For this purpose, we introduce the pseudo-Hamiltonian associated to (Σ_u):

$$\begin{aligned} H\colon \mathbb{R}^n \times (\mathbb{R}^n)^* \times \mathbb{R}^m &\longrightarrow \mathbb{R} \\ (x,p,u) &\longmapsto H(x,p,u) \coloneqq p \cdot f(x,u). \end{aligned}$$

Let consider now a singular control u over $[0\,,t]$ and denote by $x(\cdot) \coloneqq x(\cdot, x_0, u)$ the associated *state trajectory*. Since $E'_{t,x_0}(u)$ is not surjective, then there exists $\lambda \in (\mathbb{R}^n)^* \setminus \{0\}$ orthogonal to the linear subspace $\operatorname{Im} E'_{t,x_0}(u)$, *i.e.* such that $\forall\, \delta u \in L^\infty([0\,,t], \mathbb{R}^m)$:

$$\lambda \cdot (E'_{t,x_0}(u) \cdot \delta u) = \lambda \cdot \int_0^t R(t,s)\, B(s)\, \delta u(s)\, \mathrm{d}s = 0,$$

with $B(s) \coloneqq \partial_u f(x(s), u(s))$ and where $R(t,s)$ is the state transition matrix of the linear differential equation $\dot{X}(\tau) = \partial_x f(x(\tau), u(\tau)) \cdot X(\tau)$, $X(s) = I_n$. Setting[3]

$$p(s) \coloneqq \lambda\, R(t,s) \in (\mathbb{R}^n)^*,$$

we obtain that the *covector* mapping $p\colon [0\,,t] \to (\mathbb{R}^n)^* \setminus \{0\}$ is such that for almost every $s \in [0\,,t]$:

$$\dot{x}(s) = \frac{\partial H}{\partial p}[s], \quad \dot{p}(s) = -\frac{\partial H}{\partial x}[s], \quad 0 = \frac{\partial H}{\partial u}[s],$$

with $[s] \coloneqq (x(s), p(s), u(s))$. We give on Figure 10.1, a two-dimensional illustration of the geometric interpretation of the covector p for a singular control u satisfying $E_{s,x_0}(u) \in \operatorname{Fr}(\mathscr{A}(s,x_0))$ for every $s \in [0\,,t]$.

We are interested now in the computation of singular controls, that is in the resolution of the equation $\partial_u H(x,p,u) = 0$. For any given $(\bar{x}, \bar{p}, \bar{u})$, if $\partial_u H(\bar{x}, \bar{p}, \bar{u}) = 0$ and if $\partial^2_{uu} H(\bar{x}, \bar{p}, \bar{u})$ is invertible, then, by the implicit function theorem, one can find an implicit mapping, denoted $u_s(x,p)$, such that locally

$$\frac{\partial H}{\partial u}(x, p, u_s(x,p)) = 0,$$

and such that $u_s(\bar{x}, \bar{p}) = \bar{u}$.

Example 1. Let us consider two examples for which $\partial^2_{uu} H(x,p,u)$ is invertible et compute the singular control.

- Consider a pseudo-Hamiltonian of the form

  $$H(x,p,u) \coloneqq H_0(x,p) + u\, p_1 + 0.5\, u^2\, p_2$$

  with x, p in $\mathbb{R}^2$, u in $\mathbb{R}$ and where H_0 is a smooth mapping. We have $\partial_u H(x,p,u) = 0$ if and only if $p_1 + up_2 = 0$ so the singular control is of the form

  $$u_s(x,p) = -p_1/p_2 \quad \text{if} \quad p_2 \neq 0.$$

[3] The notation $\lambda R(t,s)$ stands for $\lambda \circ R(t,s)$.

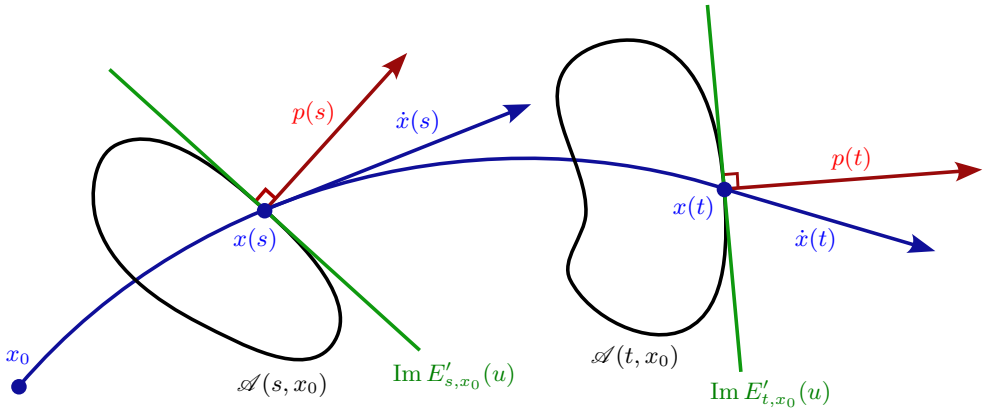


Fig. 10.1: Illustration of the geometric intepretation of the covector p, in the particular case where $E_{s,x_0}(u) \in \mathrm{Fr}(\mathscr{A}(s,x_0))$ for every $s \in [0\,,t]$.

- Consider
$$H(x,p,u) = H_0(x,p) + \sum_{i=1}^{m} u_i\, H_i(x,p) + 0.5\,\|u\|^2\, p_n$$
with $u \in \mathbb{R}^m$, $x \in \mathbb{R}^n$, $p \in (\mathbb{R}^n)^*$ and where H_0, H_1, …, H_m are smooth. Let us introduce
$$\Phi \coloneqq (H_1, \ldots, H_m).$$
Then, we have that $\partial_u H(x,p,u) = 0$ if and only if $\Phi + p_n u = 0$ so the singular control is of the form
$$u_s(x,p) = -\Phi(x,p)/p_n \quad \text{if} \quad p_n \neq 0.$$

Example 2. Consider
$$H(x,p,u) = p_1 x_2^2/2 + p_2 u$$
with $u \in \mathbb{R}$, $x \in \mathbb{R}^2$ and $p \in \mathbb{R}^2$. In this example, the matrix $\partial^2_{uu} H(x,p,u)$ is not invertible, the pseudo-Hamiltonian being affine in the control u. Let us assume that along a given extremal of reference: $\partial_u H(x(t),p(t),u(t)) = p_2(t) = 0$ almost everywhere on an interval of non-empty interior. So, for almost every time t we have:
$$\frac{\mathrm{d}}{\mathrm{d}t} p_2(t) = \dot{p}_2(t) = -\frac{\partial H}{\partial x_2}(x(t),p(t),u(t)) = p_1(t)\, x_2(t) = 0.$$
Derivating $t \mapsto p_1(t)\, x_2(t)$, we get
$$\frac{\mathrm{d}}{\mathrm{d}t}(p_1(t)\, x_2(t)) = \dot{p}_1(t) x_2(t) + p_1(t)\dot{x}_2(t) = \alpha u(t) = 0, \quad \alpha = p_1(t) \neq 0.$$
The singular control is thus $u \equiv 0$.

In the previous example, we can do the computations in a more systematic way. We need for that to introduce the following definition.

Definition 1. *Let f and g be two smooth mappings on $\mathbb{R}^n \times (\mathbb{R}^n)^*$. We define for $z \coloneqq (x,p) \in \mathbb{R}^n \times (\mathbb{R}^n)^*$:*

$$\vec{f}(z) \coloneqq \left(\frac{\partial f}{\partial p}(z), -\frac{\partial f}{\partial x}(z)\right),$$

$$\{f,g\}(z) \coloneqq g'(z)\cdot\vec{f}(z) = \sum_{i=1}^{n} \frac{\partial f}{\partial p_i}(z)\frac{\partial g}{\partial x_i}(z) - \frac{\partial f}{\partial x_i}(z)\frac{\partial g}{\partial p_i}(z).$$

The bracket $\{f,g\}$ is the Poisson bracket *of f and g, and $\vec{f}$ is the* Hamiltonian vector field *(or Hamiltonian system, or symplectic gradient) associated to f.*

We recall that the Poisson bracket is bilinear, skew-symmetric, and it satisfies the Leibniz rule and the Jacobi identity:

- $\{f,g\} = -\{g,f\}$,
- $\{fh,g\} = f\{h,g\} + h\{f,g\}$,
- $\{\{f,g\},h\} + \{\{h,f\},g\} + \{\{g,h\},f\} = 0$.

We can define the following procedure to compute the singular control (of *minimal order*) when $\partial^2_{uu}H(x,p,u)$ is not invertible.

Example 3. Consider a pseudo-Hamiltonian of the form

$$H(x,p,u) = H_0(x,p) + u\,H_1(x,p),$$

with $u \in \mathbb{R}$, $x \in \mathbb{R}^n$, $p \in (\mathbb{R}^n)^*$ and where H_0 and H_1 are smooth. Then,

$$\frac{\partial H}{\partial u}(x,p,u) = H_1(x,p).$$

We note $z \coloneqq (x,p)$. If $H_1(z(t)) = 0$ over a time interval I not reduced to a point, then, for any time $t \in I$, all the existing derivatives of $t \mapsto H_1(z(t))$ are equal to 0. Derivating as many times as needed, we can make appear the control. Let $t \in I$, then, setting $H_u(z) \coloneqq H(z,u)$, we have:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}H_1(z(t)) &= H_1'(z(t))\cdot\dot{z}(t) \\
&= \{H_u, H_1\}(z(t)) && \text{(by definition)} \\
&= \{H_0, H_1\}(z(t)) + u(t)\,\{H_1, H_1\}(z(t)) && \text{(by linearity)} \\
&= \{H_0, H_1\}(z(t)) && \text{(by skew-symmetry)} \\
&\eqqcolon H_{01}(z(t)).
\end{aligned}$$

The control does not appear, we differentiate twice:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}H_{01}(z(t)) &= H_{01}'(z(t))\cdot\dot{z}(t) \\
&= \{H_u, H_{01}\}(z(t)) \\
&= \{H_0, H_{01}\}(z(t)) + u(t)\,\{H_1, H_{01}\}(z(t)) \quad \text{(by linearity)} \\
&\eqqcolon H_{001}(z(t)) + u(t)\,H_{101}(z(t)).
\end{aligned}$$

Hence, the singular control is of the form

$$u_s(z(t)) = -\frac{H_{001}(z(t))}{H_{101}(z(t))}$$

if $H_{101}(z(t)) \neq 0$. Otherwise, if $H_{101}(z(t)) = 0$, then, we must differentiate again.

10.2.2 Weak principle and conjugate point

To the control system (Σ_u), we add fixed limit conditions and an objective function in integral form involving a smooth mapping $(x,u) \mapsto f^0(x,u)$. We thus consider an optimal control problem in Lagrange form with simple limit conditions, that is the initial and final conditions are respectively of the form $x(0) = x_0$ and $x(t_f) = x_f$, $t_f > 0$ given, and besides, there is no constraint on the control:

$$(P_L) \qquad \begin{cases} \min J(x,u) \coloneqq \displaystyle\int_0^{t_f} f^0(x(t),u(t))\,\mathrm{d}t \\ \dot{x}(t) = f(x(t),u(t)), \quad u(t) \in \mathbb{R}^m, \quad t \in [0\,,t_f] \text{ a.e.}, \\ x(0) = x_0, \quad x(t_f) = x_f. \end{cases}$$

We introduce the *augmented state* $\widetilde{x} \coloneqq (x, x^0)$ and the *augmented system* $\widetilde{f} \coloneqq (f, f^0)$ defined by:

$$\widetilde{f}(t,\widetilde{x},u) \coloneqq (f(t,x,u), f^0(t,x,u)).$$

Then, setting $\widetilde{x}_0 \coloneqq (x_0, 0)$, one can write Problem (P_L) in the *reduced* form:

$$\min \left\{ (\pi_{x^0} \circ \widetilde{E}_{t_f,\widetilde{x}_0})(u) \;\middle|\; u \in \widetilde{\mathscr{U}}_{t_f,\widetilde{x}_0},\ E_{t_f,x_0}(u) = x_f \right\},$$

where $\widetilde{\mathscr{U}}_{t_f,\widetilde{x}_0}$ is the set of admissible control laws for the augmented system and where $\widetilde{E}_{t_f,\widetilde{x}_0}$ is the endpoint mapping associated to the augmented system. We have also introduced the canonical projection $\pi_{x^0}(\widetilde{x}) = x^0$, with $\widetilde{x} = (x, x^0) \in \mathbb{R}^n \times \mathbb{R}$. Considering the augmented reachable set $\widetilde{\mathscr{A}}(t,\widetilde{x}_0) \coloneqq \widetilde{E}_{t,\widetilde{x}_0}(\widetilde{\mathscr{U}}_{t_f,\widetilde{x}_0})$, then, necessarily

$$\widetilde{E}_{t_f,\widetilde{x}_0}(u) \in \mathrm{Fr}(\widetilde{\mathscr{A}}(t_f,\widetilde{x}_0)),$$

otherwise, we could decrease the cost, see the illustration Figure 10.2. Indeed, if not, then there would exist a neighbourhood of the point $\widetilde{x}(t_f) = \widetilde{E}_{t_f,\widetilde{x}_0}(u) = (x(t_f), x^0(t_f))$ in $\widetilde{\mathscr{A}}(t_f,\widetilde{x}_0)$ containing a point (y, y^0) such that $y^0 < x^0(t_f)$, which would contradict the optimality of the control u. Hence, u is singular for the augmented endpoint mapping and we get the following necessary conditions of optimality.

Proposition 1. *If (x,u) is a solution to Problem (P_L), then, there exists a covector mapping $p\colon [0\,,t_f] \to (\mathbb{R}^n)^*$ absolutely continuous, a scalar $p^0 \in \{-1,0\}$, such that $(p,p^0) \neq (0,0)$, and such that the following conditions are satisfied for almost every $t \in [0\,,t_f]$:*

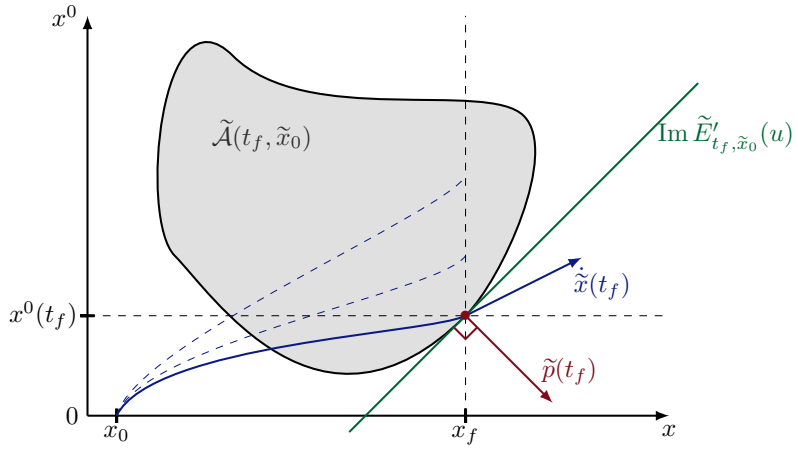


Fig. 10.2: Illustration of the optimality of the control u.

$$\begin{aligned} \dot{x}(t) &= \frac{\partial H}{\partial p}(x(t), p(t), p^0, u(t)), \\ \dot{p}(t) &= -\frac{\partial H}{\partial x}(x(t), p(t), p^0, u(t)), \\ 0 &= \frac{\partial H}{\partial u}(x(t), p(t), p^0, u(t)), \end{aligned} \tag{10.1}$$

where $H(x, p, p^0, u) := p \cdot f(x, u) + p^0 f^0(x, u)$.

Remark 2. If $E'_{t_f,x_0}(u)$ is surjective, *i.e.* if u is not a singular control for the (non-augmented) endpoint mapping, then $p^0 \neq 0$. If $E_{t_f,x_0}(u) \in \operatorname{Fr}(\mathscr{A}(t_f, x_0))$, then u is singular for E_{t_f,x_0} and $p^0 = 0$, see Figure 10.3.

Definition 2. *An* extremal *is a quadruplet* (x, p, p^0, u) *solution to the constrained pseudo-Hamiltonian equations* (10.1)*. It is a* BC-extremal *if it satisfies the limit conditions* $x(0) = x_0$ *and* $x(t_f) = x_f$*. An extremal* (x, p, p^0, u) *is said to be* abnormal *if* $p^0 = 0$ *and* normal *if* $p^0 = -1$*.*

Along an extremal, we have:

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

Let us call this condition the *Euler-Hamilton* condition. The Euler-Hamilton condition is obtained through the (first-order) Fréchet differential of the augmented endpoint mapping. The key result was the (first-order) nonlinear open mapping theorem, cf. Theorem 1. Similarly, from the following second-order nonlinear open mapping theorem, see [1, Theorem 20.3], and using the second-order differential of the augmented endpoint mapping, one can obtain a necessary condition of order 2, called the *Legendre-Clebsch condition*. Let us first recall the second-order nonlinear open mapping theorem and then give the condition.

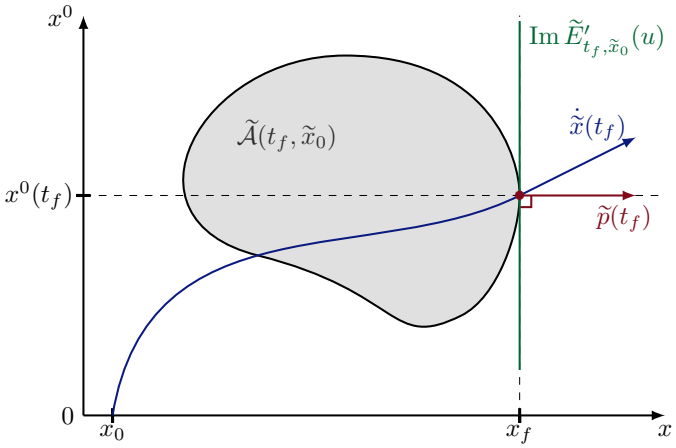


Fig. 10.3: On this illustration, we have $p^0 = 0$.

Theorem 2. *Let $F\colon E \to \mathbb{R}^n$ be a smooth function defined on a Banach space E. Let u be a singular point of corank one (*$\operatorname{codim} \operatorname{Im} F'(u) = 1$*). Let $\lambda \in (\mathbb{R}^n)^* \setminus \{0\}$ in $(\operatorname{Im} F'(u))^{\perp}$. If*[4] *$\lambda F''(u)$ is indefinite on $\operatorname{Ker} F'(u)$, then F is locally open at u.*

Remark 3. The bilinear form

$$\lambda F''(u) \in \mathscr{L}(E, \mathscr{L}(E, \mathbb{R})) \simeq \mathscr{L}_2(E \times E, \mathbb{R})$$

is called the *second-order intrinsic derivative* of F at u, and is defined up to a scalar in the case of corank one.

Along an extremal, the following Legendre-Clebsch condition, see [1, Proposition 20.11], is satisfied:

$$\frac{\partial^2 H}{\partial u^2}(x(t), p(t), p^0, u(t)) \cdot (v, v) \leq 0, \quad \forall\, v \in \mathbb{R}^m, \quad \forall\, t \in [0\,, t_f] \text{ a.e..}$$

Let $t \in [0\,, t_f]$ and write $F_t(u) \coloneqq H(t, x(t), p(t), p^0, u)$. If the Euler-Hamilton and Legendre-Clebsch conditions are satisfied at time t along the extremal, then it means that $u(t)$ satisfies the necessary local optimality conditions of order 1 and 2 of the following unconstrained optimization problem:

$$\max_{u \in \mathbb{R}^m} F_t(u). \tag{10.2}$$

If besides, the Legendre-Clebsch condition is strict, then $u(t)$ satisfies the second-order sufficient condition of strict local optimality for Problem (10.2). But, even if the sufficient condition is satisfied all along the interval $[0\,, t_f]$, *i.e.* if

$$\frac{\partial^2 H}{\partial u^2}(x(t), p(t), p^0, u(t)) \cdot (v, v) < 0, \quad \forall\, v \in \mathbb{R}^m, \quad \forall\, t \in [0\,, t_f] \text{ a.e.,}$$

[4] The notation $\lambda F''(u)$ stands for $\lambda \circ F''(u)$.

this does not give us a sufficient condition of local optimality for Problem (P_L). We need an additional condition called the *Jacobi condition* [1], even in the frame of calculus of variations [34]. The rest of this section is dedicated to the Jacobi condition, see Definition 4 and Theorem 3.

As mentioned before, if $\bar{u}$ is solution to Problem (P_L), then, necessarily

$$\widetilde{E}_{t_f,\widetilde{x}_0}(\bar{u}) \in \mathrm{Fr}(\widetilde{\mathscr{A}}(t_f,\widetilde{x}_0))$$

and so not only $\bar{u}$ is singular for the augmented endpoint mapping, but also this mapping is not locally open at $\bar{u}$. The fact that $\bar{u}$ is singular ensures the existence of a non-zero linear form $\bar{\tilde{\lambda}} \in (\mathbb{R}^{n+1})^*$ orthogonal to $\operatorname{Im} \widetilde{E}'_{t_f,\widetilde{x}_0}(\bar{u})$, that is

$$\bar{\tilde{\lambda}} \in \left(\operatorname{Im} \widetilde{E}'_{t_f,\widetilde{x}_0}(\bar{u})\right)^{\perp}, \quad \bar{\tilde{\lambda}} \neq 0.$$

The fact that the mapping is not locally open at $\bar{u}$ can be used to obtain new necessary local optimality conditions. For now, we assume that the control $\bar{u}$ is of corank one. Hence, the associated trajectory $\bar{x}$ admits a unique lift (up to a scalar) $(\bar{x},\bar{p},p^0,\bar{u})$ on $[0\,,t_f]$, that we suppose to be normal $(p^0 \neq 0)$. In this context, we apply Theorem 2 to the augmented endpoint mapping. We thus obtain that

$$\bar{\tilde{\lambda}}\widetilde{E}''_{t_f,\widetilde{x}_0}(\bar{u}) \text{ is semi-definite on } \operatorname{Ker} \widetilde{E}'_{t_f,\widetilde{x}_0}(\bar{u}). \tag{10.3}$$

Remark 4. Let us relate this to an optimization point of view. Consider Problem (P_L) in its reduced form and define the Lagrangian (we omit indices):

$$L(u,\tilde{\lambda}) = \lambda^0\,\pi_{x^0}(\widetilde{E}(u)) + \lambda\cdot(E(u)-x_f), \quad \tilde{\lambda} =: (\lambda,\lambda^0) \in (\mathbb{R}^n)^* \times \mathbb{R}_-.$$

We note $x^0(u) \coloneqq \pi_{x^0}(\widetilde{E}(u))$ and $x(u) \coloneqq E(u)$ so that $\widetilde{E}(u) = (x(u),x^0(u))$. From the optimization point of view, we have the second-order necessary condition of local optimality: $\partial^2_{uu}L(\bar{u},\bar{\tilde{\lambda}}) = \bar{\tilde{\lambda}}\widetilde{E}''(\bar{u})$ negative semi-definite on the tangent space to the constraints. In the case of qualified constraints, that is $\lambda^0 < 0$, the tangent space is given by $\operatorname{Ker} x'(\bar{u})$ and not by $\operatorname{Ker} \widetilde{E}'(\bar{u})$, as written in Equation (10.3). But these two kernels are the same since for every $v \in \operatorname{Ker} x'(\bar{u})$, we have from the first-order necessary local optimality condition:

$$0 = \frac{\partial L}{\partial u}(\bar{u},\bar{\tilde{\lambda}})\cdot v = \lambda^0\,x^{0\prime}(\bar{u})\cdot v + \lambda\cdot(x'(\bar{u})\cdot v) = \lambda^0\,x^{0\prime}(\bar{u})\cdot v$$

and so $x^{0\prime}(\bar{u})\cdot v = 0$ since $\lambda^0 \neq 0$. Hence, $\operatorname{Ker} x'(\bar{u}) \subset \operatorname{Ker} \widetilde{E}'(\bar{u})$. Besides, $\operatorname{Ker} \widetilde{E}'(\bar{u}) \subset \operatorname{Ker} x'(\bar{u})$ since $\widetilde{E}(u) = (x(u),x^0(u))$. In conclusion, $\operatorname{Ker} \widetilde{E}'(\bar{u}) = \operatorname{Ker} x'(\bar{u})$ and so the second-order necessary local optimality condition is indeed that $\bar{\tilde{\lambda}}\widetilde{E}''(\bar{u})$ has to be negative semi-definite on $\operatorname{Ker} \widetilde{E}'(\bar{u})$. This ends the remark.

For $t \in [0\,,t_f]$, we define the symmetric bilinear form

$$B_t \coloneqq \bar{\tilde{\lambda}}\widetilde{E}''_{t,\widetilde{x}_0}(\bar{u})$$

and we introduce $K_t \coloneqq \operatorname{Ker} \widetilde{E}'_{t,\widetilde{x}_0}(\bar{u})$. We have the following result, see [16] for more details.

Proposition 2. *If the extremal $(\bar{x},\bar{p},-1,\bar{u})$ satisfies the strong Legendre-Clebsch condition, then there exists $\varepsilon > 0$ such that $B_t|_{K_t}$ is negative definite for every $t \in [0\,,\varepsilon]$.*

Clearly, if $s \leq t$, then $B_t|_{K_t} \prec 0$ (*i.e.* negative definite) implies $B_s|_{K_s} \prec 0$, whence the following definition. We define the *first conjugate time* t_{1c}, along a normal extremal satisfying the strong Legendre-Clebsch condition, as the supremum of times t such that B_t is negative definite:

$$t_{1c} = \sup \{t > 0 \mid B_t|_{K_t} \prec 0\}.$$

From [1], then $B_{t_{1c}}|_{K_{t_{1c}}}$ has a non-trivial kernel. Hence, the conjugate times are defined as the times t_c such that $B_{t_c}|_{K_{t_c}}$ is degenerate.

Let us take a normal extremal of reference $(\bar{x}, \bar{p}, -1, \bar{u})$ and assume that the strong Legendre-Clebsch condition is satisfied:

$$\frac{\partial^2 H}{\partial u^2}(\bar{x}(t), \bar{p}(t), -1, \bar{u}(t)) \cdot (v, v) < 0, \quad \forall\, v \in \mathbb{R}^m, \quad \forall\, t \in [0\,, t_f] \text{ a.e..}$$

Under this assumption, the equation $\partial_u H = 0$ may be solved in a neighbourhood of the reference extremal and we can define the control as a function of the state and the costate, that is in feedback form, that we note $u_s(x, p)$. Setting on this neighbourhood the Hamiltonian $\mathbf{H}(z) \coloneqq H(z, -1, u_s(z))$, $z \coloneqq (x, p)$, we get

$$\mathbf{H}'(z) = \frac{\partial H}{\partial z}(z, -1, u_s(z)) + \frac{\partial H}{\partial u}(z, -1, u_s(z))\, u_s'(z) = \frac{\partial H}{\partial z}(z, -1, u_s(z))$$

since $\partial_u H(z, -1, u_s(z)) = 0$. Hence, on this neighbourhood, the system

$$\dot{x}(t) = \frac{\partial H}{\partial p}[t], \quad \dot{p}(t) = -\frac{\partial H}{\partial x}[t], \quad 0 = \frac{\partial H}{\partial u}[t],$$

with $[t] \coloneqq (x(t), p(t), -1, u(t))$, from Proposition 1, is equivalent to the *Hamiltonian system*

$$\dot{z}(t) = \vec{\mathbf{H}}(z(t)), \quad \vec{\mathbf{H}}(z) = \left(\frac{\partial \mathbf{H}}{\partial p}(z), -\frac{\partial \mathbf{H}}{\partial x}(z)\right).$$

Definition 3. *A solution to the linearized differential equation along z, called* Jacobi equation*,*

$$\dot{\widehat{\delta z}}(t) = \vec{H}'(z(t)) \cdot \delta z(t), \tag{10.4}$$

is called a Jacobi field*. A Jacobi field $\delta z = (\delta x, \delta p) \in \mathbb{R}^n \times (\mathbb{R}^n)^*$, is said to be* vertical at time t *if* $\delta x(t) = 0$.

We can give now a geometric characterization of the first conjugate time.

Proposition 3. *A time $t_c \in (0\,, t_f]$ is a conjugate time along a normal extremal satisfying the strong Legendre-Clebsch condition if and only if there exists a Jacobi field $\delta z = (\delta x, \delta p)$ vertical at 0 and t_c, and such that $\delta x \not\equiv 0$ on $[0\,, t_f]$.*

Definition 4 (Jacobi condition). *For an extremal defined on an interval $[a\,, b]$, we say that the* weak Jacobi condition *is satisfied if the open interval $(a\,, b)$ does not contain any conjugate time. We say that the* strong Jacobi condition *is satisfied if the semi-open interval $(a\,, b]$ does not contain any conjugate time.*

Remark 5. In the previous proposition, it is δx and not $\delta\tilde{x}$ which is under concern. This is possible since along a normal extremal satisfying the strong Legendre-Clebsch condition, we have $\delta\tilde{x} = (\delta x, \delta x^0)$ is vertical at t if and only if δx is vertical at t. To note that, it suffices to notice that $\delta x(t) = E'(u)\cdot\delta u = x'(u)\cdot\delta u$ for a given δu (with the notations of the previous remark). Hence, from the firs-order necessary optimality condition:

$$0 = \lambda^0\, x^{0\prime}(u)\cdot\delta u + \lambda\cdot(x'(u)\cdot\delta u) =: \lambda^0\,\delta x^0(t) + \lambda\cdot\delta x(t)$$

and so, since $\lambda^0 \neq 0$, we get $\delta x(t) = 0$ if and only if $\delta\tilde{x}(t) = (\delta x(t), \delta x^0(t)) = 0$.

The question we can ask now is: does the trajectory become necessarily non optimal after the first conjugate time? To answer yes to this question, the quadratic form Q_t associated to $B_t|_{K_t}$ must be indefinite for $t > t_{1c}$. However, it can happen in degenerate cases, that Q_t stays semi-definite on an interval $[t_{1c}\,, t_{1c}+\eta]$, $\eta > 0$. In the analytic frame, this cannot happen and necessarily for $t > t_{1c}$, Q_t is indefinite. We thus obtain the following (weak) second-order local optimality condition in the analytic case of corank one, see [1] for more details.

Theorem 3. *For a normal extremal satisfying the strong Legendre-Clebsch condition, whose associated analytic control is of corank one, the weak Jacobi condition is a necessary local optimality condition for the L^∞ topology (on u).*

Remark 6. This result is related to the notion of weak local solution. See [1, 8] for more details about this notion. For sufficient weak local conditions, we need to take into account the two-norm discrepancy, see [18]. We prefer to present next a more geometric point of view in relation with strong local optimality, see Theorem 4.

10.2.3 Pontryagin Maximum Principle and Hamiltonian frame

In the weak principle the control is unconstrained. We consider now an optimal control problem in which the control takes its values in any arbitrary set. We consider the following optimal control problem in Bolza form:

$$\text{(OCP)}\qquad \begin{cases} \min\, J(x,u) \coloneqq g(x(t_f)) + \displaystyle\int_0^{t_f} f^0(x(t),u(t))\,\mathrm{d}t \\ \dot{x}(t) = f(x(t),u(t)), \quad u(t)\in U, \quad t\in[0\,,t_f] \text{ a.e.,} \\ x(0) = x_0, \quad c(x(t_f)) = 0_{\mathbb{R}^p}, \end{cases}$$

where f, f^0, g and c are smooth functions. The initial and final times are fixed respectively to 0 and t_f, and the initial condition is simply $x(0) = x_0$, with $x_0 \in \mathbb{R}^n$ given. We have some final conditions of the form $c(x) = 0 \in \mathbb{R}^p$, with $p \leq n$. The set $U \subset \mathbb{R}^m$ is arbitrary. We assume also that c is a submersion on the set $c^{-1}(0)$, that is $c'(x)$ is surjective for any x such that $c(x) = 0$.

From the classical Pontryagin Maximum Principle [33], we have the following. If (x,u) is solution to Problem (OCP), then, there exists a covector mapping $p\colon [0\,,t_f] \to (\mathbb{R}^n)^*$ absolutely continuous,[5] a scalar $p^0 \in \{-1, 0\}$ and a linear form $\lambda \in (\mathbb{R}^p)^*$, such that $(p(\cdot), p^0) \neq (0,0)$ and such that the following equations are satisfied for almost every $t \in [0\,,t_f]$:

[5] Actually, the covector mapping is even Lipschitz in our setting.

$$\begin{aligned} \dot{x}(t) = &\quad \frac{\partial H}{\partial p}(x(t), p(t), p^0, u(t)), \\ \dot{p}(t) = &-\frac{\partial H}{\partial x}(x(t), p(t), p^0, u(t)), \\ H(x(t), p(t), p^0, u(t)) = &\max_{w \in U} H(x(t), p(t), p^0, w), \end{aligned} \tag{10.5}$$

where $H(x, p, p^0, u) \coloneqq p \cdot f(x, u) + p^0 f^0(x, u)$. The limit conditions $x(0) = x_0$ and $c(x(t_f))$ are satisfied. Besides, setting

$$\xi(x) \coloneqq p^0\, g(x) + \sum_{i=1}^{p} \lambda_i\, c_i(x),$$

we have the transversality condition:

$$p(t_f) = \xi'(x(t_f)). \tag{10.6}$$

The proof is based upon the construction of needle variations, see Figure 10.4, introduced by Boltyanski [7]. As we can see on the figure, the variations are applied on the control. They are not small in L^∞ norm but in L^1 norm, in comparison with the weak principle. The variation is coming from the constant $\bar{u}$ which replace the reference control on a small time interval of length dt.

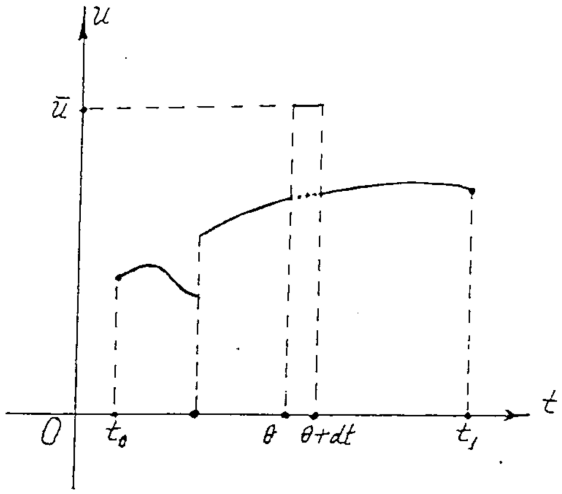


Fig. 10.4: Illustration of the needle variation excerpted from [7].

The following definition replace Definition 2 in this context.

Definition 5. *A* Pontryagin-Boltyanski extremal *is a quadruplet* (x, p, p^0, u) *solution to the constrained pseudo-Hamiltonian equations* (10.5)*. It is a* BC-extremal*, for Problem* (OCP)*, if it satisfies the limit conditions* $x(0) = x_0$*,* $c(x(t_f)) = 0$ *and the transversality condition* (10.6)*. An extremal* (x, p, p^0, u) *is still said to be* abnormal *if* $p^0 = 0$ *and* normal *if* $p^0 = -1$*.*

Let us give a more geometric point of view and reveal the Hamiltonian frame: in the following proposition, **H** is a Hamiltonian.

Proposition 4. *Let $(\bar{x}, \bar{p}, p^0, \bar{u})$ be a Pontryagin-Boltyanski extremal. We note $\bar{z} \coloneqq (\bar{x}, \bar{p})$. If for almost every $t \in [0\,, t_f]$, in a neighbourhood of $\bar{z}(t)$, the maximized pseudo-Hamiltonian defines a smooth mapping*

$$z \mapsto \mathbf{H}(z) \coloneqq \max_{u \in U} H(z, p^0, u),$$

then, for almost every $t \in [0\,, t_f]$, we have $\dot{\bar{z}}(t) = \overrightarrow{\mathbf{H}}(\bar{z}(t))$.

Proof. We introduce the following notation for a pseudo-Hamiltonian:

$$\overrightarrow{H}(x, p, p^0, u) \coloneqq \left(\frac{\partial H}{\partial p}(x, p, p^0, u), -\frac{\partial H}{\partial x}(x, p, p^0, u)\right).$$

Since for almost every $t \in [0\,, t_f]$, we have from the Pontryagin Maximum Principle:

$$\dot{\bar{z}}(t) = \overrightarrow{H}(\bar{z}(t), p^0, \bar{u}(t)),$$

it is sufficient to prove that for almost every $t \in [0\,, t_f]$ we have: $\mathbf{H}'(\bar{z}(t)) = \partial_z H(\bar{z}(t), p^0, \bar{u}(t))$. Let $t \in [0\,, t_f]$ for which (10.5) is satisfied and for which $\mathbf{H}$ is well defined and smooth, on an open neighbourhood of $\bar{z}(t)$. On this neighbourhood, we set $F(z) \coloneqq \mathbf{H}(z) - H(z, p^0, \bar{u}(t))$. Then, we have $F(z) \geq 0$ and $F(\bar{z}(t)) = 0$. So F is minimized on this open neighbourhood at the point $z = \bar{z}(t)$, which implies $F'(\bar{z}(t)) = 0$. The result is proved.

Definition 6 (Hamilton extremal). *A Pontryagin-Boltyanski extremal satisfying the assumptions of Proposition 4 is called a* Hamilton extremal.

Let us consider a smooth pseudo-Hamiltonian and fix p^0 so we do note write p^0 in the pseudo-Hamiltonian. Let us assume that U is an open subset of $\mathbb{R}^m$ and that $u \mapsto H(z, u)$, $z = (x, p)$, admits a unique maximum over U at $u = u_m(z)$, for any z. We assume also that u_m is smooth. Then, the maximized pseudo-Hamiltonian furnishes a smooth Hamiltonian given by:

$$\mathbf{H}(z) = H(z, u_m(z)).$$

From the Pontryagin Maximum Principle, along a Pontryagin-Boltyanski extremal is satisfied

$$\dot{z}(t) = \overrightarrow{H}(z(t), u_m(z(t))).$$

From the previous corollary, it satisfies also

$$\dot{z}(t) = \overrightarrow{\mathbf{H}}(z(t)),$$

which can be easily checked since $\mathbf{H}'(z(t)) = \partial_z H[t] + \partial_u H[t] \cdot u_m'(z(t)) = \partial_z H[t]$ since $\partial_u H[t] = 0$, with $[t] \coloneqq (z(t), u_m(z(t)))$. Besides, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{H}(z(t)) &= \mathbf{H}'(z(t)) \cdot \dot{z}(t) \\ &= \mathbf{H}'(z(t)) \cdot \overrightarrow{\mathbf{H}}(z(t)) = \{\mathbf{H}, \mathbf{H}\}(z(t)) = 0,\end{aligned}$$

hence, the Hamiltonian is constant along the extremals.

In this Hamiltonian frame, we can complete Theorem 3 and give a sufficient condition of strong local optimality. We have the following result.

Theorem 4. *For a normal Hamilton BC-extremal of Problem (P_L), the strong Jacobi condition is a sufficient condition of strict local optimality for the $\mathscr{C}^0$ topology (on x).*

Proof. A brief proof is given in [15, Appendix A]. See also [1, Chapter 21].

To summarize, combining Theorems 3 and 4, before the first conjugate time, the local optimality is satisfied in a big neighbourhood of the trajectory for the $\mathscr{C}^0$ topology, while after the first conjugate time, the local optimality is lost, even in small neighbourhoods of the control for the L^∞ topology.

To complete our geometric point of view, let us give a brief insight of the symplectic origins of this Hamiltonian frame. The Hamilton extremals bring us to the theories of Hamiltonian dynamical system and symplectic geometry [1, 3, 6, 30]. Roughly speaking, there are two central objects which play a crucial role in our frame: the Hamiltonian $\mathbf{H}$ and the symplectic form denoted ω. The Hamiltonian is a function defined on the phase space, while the symplectic form is a (differential) two-form on the phase space. It is this symplectic form which permits to define the notion of Hamiltonian vector field (or symplectic gradient) leading to the Hamilton equations. This Hamiltonian together with the symplectic form define how the Hamilton extremals evolve in the phase space, this phase space being the intrinsic geometric (or symplectic) structure in which we can search the solutions of our optimal control problem. We have already seen from where comes the Hamiltonian. Let us see how to obtain the symplectic form. First, consider that the state space is a differential manifold denoted Ω. We define the phase space, or cotangent space, as the space

$$T^*\Omega \coloneqq \{(x,p) \mid (x,p) \in \Omega \times T_x^*\Omega\},$$

where $T_x^*\Omega$ is the dual space of $T_x\Omega$. In the case where Ω is an open subset of $\mathbb{R}^n$, we have

$$T_x\Omega \simeq \mathbb{R}^n, \quad T_x^*\Omega \simeq \mathbb{R}^n \quad \text{and} \quad T^*\Omega \simeq \Omega \times \mathbb{R}^n.$$

From there, we can define in a canonical way a (differential) one-form on $T^*\Omega$ given by

$$\alpha_{(x,p)} \coloneqq \sum_{i=1}^{n} p_i \mathrm{d}x_i.$$

This one-form is called the *Liouville form*. The symplectic structure of $T^*\Omega$ is given by the exterior derivative of the Liouville form, that is the two-form ω defined by

$$\omega_{(x,p)} \coloneqq -\mathrm{d}\alpha_{(x,p)} = \sum_{i=1}^{n} \mathrm{d}x_i \wedge \mathrm{d}p_i.$$

The symplectic form is thus a non-degenerate differential form. The symplectic structure of $T^*\Omega$ leads to the definition of a Hamiltonian vector field. Let H be a function on $T^*\Omega$, that is a Hamiltonian. The associated Hamiltonian vector field of H is the vector field $\vec{H}$ defined by

$$\omega(\vec{H}, \cdot) = \mathrm{d}H.$$

Besides, the symplectic structure permits us to define the notion of Hamiltonian flow. Let $\vec{H}$ be a Hamiltonian vector field. The Hamiltonian flow is the one-parameter family of diffeomorphisms ϕ_t defined by

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi_t(z_0) = \overrightarrow{H}(\phi_t(z_0)), \quad \phi_0(z_0) = z_0.$$

The Hamiltonian flow is thus the solution of the Cauchy problem:

$$\dot{z}(t) = \overrightarrow{H}(z(t)), \quad z(0) = z_0.$$

Remark 7. One of the main results in symplectic geometry is Darboux's theorem which states that, locally, any symplectic form is equivalent to the canonical symplectic form ω_0 given by

$$\omega_0 = \sum_{i=1}^{n} \mathrm{d}x_i \wedge \mathrm{d}p_i.$$

This means that, locally, we can find a change of coordinates to pass from ω to ω_0. A change of coordinates on a symplectic structure is called a *symplectomorphism*, it is a diffeomorphism that preserves the symplectic two-form. In our frame, from a change of coordinates on Ω, we can define canonically a symplectomorphism on $T^*\Omega$. Indeed, if φ is a change of coordinates on the state, then, the canonical change of coordinates on $T^*\Omega$ is given by

$$\begin{aligned} \varphi^* \colon T^*\Omega &\longrightarrow T^*\Omega \\ (x,p) &\longmapsto \varphi^*(x,p) \coloneqq (\varphi(x), p\,\varphi'(x)^{-1}). \end{aligned}$$

One can check that φ^* is a symplectomorphism. Besides, another fundamental theorem of Hamiltonian systems is the Arnold-Liouville theorem and the notion of action-angle variables defined on a torus called the Liouville torus. This theorem permits us to classify the Hamilton extremals with respect to their topological properties and so to study how the extremals describe the torus. In the particular case of classical mechanics, the Morse theory [32] applied to the elevation function given by the mechanical energy leads to the description of the Liouville torus. This is also the case for instance in Riemannian geometry of revolution in dimension 2. The key point being the existence of a non-degenerate potential. This can be extended to Zermelo geometry, see [10]. This ends the remark.

10.2.4 Problems affine in the control

We call *affine control system* any system of the form

$$\dot{x}(t) = F_0(x(t)) + \sum_{i=1}^{m} u_i(t)\, F_i(x(t))$$

where the vector field F_0 is called the *drift* and $D = \{F_1, \ldots, F_m\}$ is called the *control distribution*. The Legendre-Clebsch condition is meaningless in the affine case where $\partial^2_{uu} H \equiv 0$. We thus introduce the following definition. We say that an extremal (x, p, p^0, u) defined on $[0\,, t_f]$ satisfies the *degenerate Legendre-Clebsch condition* if for almost every $t \in [0\,, t_f]$:

$$\frac{\partial^2 H}{\partial u^2}(x(t), p(t), p^0, u(t)) = 0.$$

In this case, for an affine system for instance, we have the following necessary condition. If the control is interior to the constraint, then the *Goh condition* [24] is necessary:

$$\left\{\frac{\partial H_t}{\partial u_i}, \frac{\partial H_t}{\partial u_j}\right\}(z(t)) = 0, \quad i,j = 1,\dots,m, \quad t \in [0\,, t_f] \text{ a.e..} \tag{10.7}$$

We recall that $z = (x,p)$ and we mention that we have introduced the notations $H_t(z) \coloneqq H(z, p^0, u(t))$ and

$$\frac{\partial H_t}{\partial u_i}(z) \coloneqq \frac{\partial H}{\partial u_i}(z, p^0, u(t)), \quad i = 1,\dots,m.$$

In addition to the Goh condition, in the case that the extremal satisfies the degenerate Legendre-Clebsch condition, under some additional assumptions, the extremal has to satisfy the *generalized Legendre-Clebsch condition*, see [1]. In the particular case of mono-input and autonomous affine pseudo-Hamiltonian:

$$H(x,p,u) = H_0(x,p) + u\,H_1(x,p),\ u \in \mathbb{R},$$

with $H_i(x,p) \coloneqq p \cdot F_i(x)$ the Hamiltonian lift of F_i, the generalized Legendre-Clebsch condition is given by:

$$H_{101}(z(t)) = \{H_1, \{H_0, H_1\}\}(z(t)) \geq 0, \quad t \in [0\,, t_f] \text{ a.e..}$$

Let us consider a mono-input autonomous affine control system of the form $\dot{x} = F_0(x) + u\,F_1(x)$, where $u \in \mathbb{R}$ and $x \in \mathbb{R}^n$. Let H_0, H_1 be the Hamiltonian lifts of F_0, F_1. Let (x,p,u) be an extremal[6] on $I \subset \mathbb{R}$, I an open interval of non-empty interior. Along this extremal, we have $\partial_u H(z(t), u(t)) = H_1(z(t)) = 0$, $z = (x,p)$. We are in the frame of Section 10.2.1 and we call *singular* this extremal. The value of the control can be obtained differentiating at least twice with respect to the time the mapping H_1 along the extremal, see Example 2:

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}H_1(z(t)) = H_{01}(z(t)) = 0 \\ &\frac{\mathrm{d}}{\mathrm{d}t}H_{01}(z(t)) = H_{001}(z(t)) + u(t)\,H_{101}(z(t)) = 0 \end{aligned} \tag{10.8}$$

and (10.8) provides the singular control in feedback form:

$$u_s(z) \coloneqq -\frac{H_{001}(z)}{H_{101}(z)}$$

outside

$$\Sigma_{101} \coloneqq \{z \in T^*\mathbb{R}^n \mid H_{101}(z) = 0\}\,.$$

A singular extremal defined outside Σ_{101} is called a *singular extremal of minimal order*. Denote

$$\Sigma_1 \coloneqq \{z \in T^*\mathbb{R}^n \mid H_1(z) = 0\}\,, \quad \Sigma_{01} \coloneqq \{z \in T^*\mathbb{R}^n \mid H_{01}(z) = 0\}\,,$$

and define the *singular manifold* $\Sigma_s \coloneqq \Sigma_1 \cap \Sigma_{01}$. Plugging u_s inside the pseudo-Hamiltonian, we get the Hamiltonian

$$\mathbf{H_s}(z) \coloneqq H(z, u_s(z)).$$

We have the following result.

[6] We do not mention p^0 in the extremal since it has no role here.

Proposition 5. *Let $\bar{z} \in \Sigma_s \setminus \Sigma_{101}$. Then, there is exactly one singular extremal of minimal order passing through $\bar{z}$. It is contained in Σ_s and it is solution to the Hamiltonian system*

$$\dot{z}(t) = \overrightarrow{\mathbf{H_s}}(z(t)),$$

the control being given by $u_s(z(t))$.

Proof. Denote by $z(\cdot)$ the integral curve of $\overrightarrow{\mathbf{H_s}}$ passing through $\bar{z}$ at time $t = 0$. Let $\varphi(t) \coloneqq \xi(z(t))$, with $\xi(z) \coloneqq (H_1(z), H_{01}(z))$. Then, $\varphi(0) = 0_{\mathbb{R}^2}$ and $\varphi'(t) = \xi'(z(t))\,\overrightarrow{\mathbf{H_s}}(z(t))$.

• To prove that Σ_s is invariant by the Hamiltonian flow of $\mathbf{H_s}$ we prove that $\varphi(t) = 0$ for every t. We have:

$$\begin{aligned}
\varphi_1'(t) &= \{\mathbf{H_s}, H_1\}(z(t)) = \{H_0 + u_s\, H_1, H_1\}(z(t)) \\
&= \Big(\{H_0, H_1\} + u_s\,\{H_1, H_1\} + H_1\,\{u_s, H_1\}\Big)(z(t)) \\
&= \{H_0, H_1\}(z(t)) + H_1(z(t))\,\{u_s, H_1\}(z(t))
\end{aligned}$$

and

$$\begin{aligned}
\varphi_2'(t) &= \{\mathbf{H_s}, H_{01}\}(z(t)) = \{H_0 + u_s\, H_1, H_{01}\}(z(t)) \\
&= \Big(\underbrace{\{H_0, H_{01}\} + u_s\,\{H_1, H_{01}\}}_{=0 \text{ by definition of } u_s} + H_1\,\{u_s, H_{01}\}\Big)(z(t)) \\
&= H_1(z(t))\,\{u_s, H_{01}\}(z(t)).
\end{aligned}$$

Hence, $\varphi'(t) = A(t)\,\varphi(t)$, with

$$A(t) \coloneqq \begin{pmatrix} \{u_s, H_1\}(z(t)) & 1 \\ \{u_s, H_{01}\}(z(t)) & 0 \end{pmatrix}$$

and $\varphi(0) = 0$. Thus, $\varphi(t) = 0$ for every t.

• Now, since

$$\begin{aligned}
\mathbf{H_s}'(z) &= \frac{\partial H}{\partial z}(z, u_s(z)) + \frac{\partial H}{\partial u}(z, u_s(z))\,u_s'(z) \\
&= \frac{\partial H}{\partial z}(z, u_s(z)) + H_1(z)\,u_s'(z),
\end{aligned}$$

we have $\mathbf{H_s}'(z(t)) = \partial_z H(z(t), u_s(z(t)))$ along the integral curve $z(\cdot)$ of $\overrightarrow{\mathbf{H_s}}$. Hence, $(z, u_s(z))$ is a singular extremal (of minimal order). Conversely, let us consider a singular extremal of minimal order passing through $\bar{z}$. Then, by definition, it is contained in Σ_s, the control is given by u_s and it is also solution to the Hamiltonian system $\dot{z}(t) = \overrightarrow{\mathbf{H_s}}(z(t))$ since $\mathbf{H_s}'(z(t)) = \partial_z H(z(t), u_s(z(t)))$ as soon as $H_1(z(t)) = 0$. This ends the proof.

Remark 8. By the previous proposition, Σ_s is invariant by the Hamiltonian flow of $\mathbf{H_s}$. Assume that ξ is a submersion on $\Sigma_s \neq \emptyset$. Then, $\Sigma_s = \xi^{-1}(0)$ is a submanifold of $T^*\mathbb{R}^n$ of codimension 2, and the sets Σ_1 and Σ_{01} are transverse at each point of Σ_s.

Assume now that the control is bounded between -1 and 1. The classification of the bang extremals (of maximal norm) and the singular extremals depends on the order of contact of the bang extremals with the *switching manifold* Σ_1 and the signs of the Poisson brackets H_{001} and H_{101} at the contact points. The local optimality of the extremals depends on this classification and the existence of conjugate points. See [8, 9, 11] for more details. See also [19, Section 1.2] for algorithms to compute conjugate points in the regular and singular cases.

10.3 Indirect Numerical Methods for Geometric Control

We present in this part, some indirect numerical methods. An indirect method aims to solve the equations given by the Pontryagin Maximum Principle, that is to compute BC-extremals.

10.3.1 Indirect simple shooting

Let consider Problem (P_L) and assume that for any extremal (z, p^0, u), we can write $u(t) = u(z(t))$, with $z \mapsto u(z)$ at least $\mathscr{C}^1$. Plugging $u(z)$ in $\vec{H}$, then, finding a BC-extremal amounts to solve the (Two-Point) Boundary Value Problem:

$$\text{(BVP)} \qquad \begin{cases} \dot{z}(t) &= \vec{H}(z(t), p^0, u(z(t))), \\ 0_{\mathbb{R}^{2n}} &= b(z(0), z(t_f)) \coloneqq (x(0) - x_0, x(t_f) - x_f). \end{cases}$$

We can rewrite (BVP) as a set on nonlinear equations introducing the *simple shooting function*:

$$\begin{aligned} S \colon \mathbb{R}^n &\longrightarrow \mathbb{R}^n \\ p_0 &\longmapsto S(p_0) \coloneqq \pi(z(t_f, x_0, p_0)) - x_f, \end{aligned}$$

where $\pi(x, p) = x$ and where $z(\cdot, x_0, p_0)$ is the solution of the Cauchy problem $\dot{z}(t) = \vec{H}(z(t), p^0, u(z(t)))$, $z(0) = (x_0, p_0)$. Solving (BVP) amounts to solve $S(p_0) = 0$. This is the *indirect simple shooting method*, see Figure 10.5.

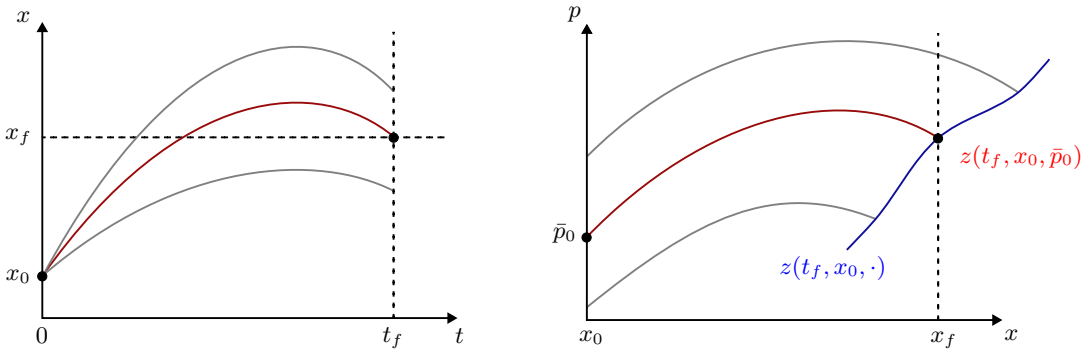


Fig. 10.5: Illustration of the indirect simple shooting method.

Remark 9. If $\bar{p}_0$ satisfies $S(\bar{p}_0) = 0$, then, the integral curve $\bar{z}(\cdot) \coloneqq z(\cdot, x_0, \bar{p}_0)$, with the control $\bar{u}(\cdot) \coloneqq u(\bar{z}(\cdot))$ and p^0, is a BC-extremal of Problem (P_L).

Remark 10. From Proposition 4, the integral curve $z(\cdot, x_0, p_0)$ is also solution of $\dot{z}(t) = \overrightarrow{\mathbf{H}}(z(t))$, $z(0) = (x_0, p_0)$, where $\mathbf{H}(z) \coloneqq H(z, p^0, u(z))$.

To solve the shooting equations, we need to compute $z(t_f, x_0, p_0)$. This is usually computed by Runge-Kutta solvers. Then, to find a zero of the shooting function, we can use Newton-like solvers. The Newton methods are known to be sensitive with respect to the initial iterate. One difficulty is to provide a good enough initial guess to make the Newton solver converge. We recall the Newton iteration:

$$p_0^{(k+1)} = p_0^{(k)} + d^{(k)},$$

with $d^{(k)}$ the solution of the linear system

$$S'(p_0^{(k)}) \cdot d = -S(p_0^{(k)}).$$

The Jacobian of the shooting function is given by:

$$S'(p_0) \cdot d = \pi \left(\frac{\partial z}{\partial p_0}(t_f, x_0, p_0) \cdot d \right) = \pi \left(\frac{\partial z}{\partial z_0}(t_f, x_0, p_0) \cdot (0_{\mathbb{R}^n}, d) \right),$$

where z_0 stands for (x_0, p_0). We need to compute $\partial_{z_0} z(\cdot, x_0, p_0) \cdot \delta z_0$, $\delta z_0 = (0_{\mathbb{R}^n}, d)$, solution of the *variational equations*:

$$\dot{\delta z}(t) = \overrightarrow{\mathbf{H}}'(z(t, x_0, p_0)) \cdot \delta z(t), \quad \delta z(0) = \delta z_0$$

that we recognize to be Jacobi equations. Hence, the invertibility of the Jacobian of the shooting function is directly related to the absence of conjugate points.

To compute the directional derivative $\partial_{z_0} z(t_f, z_0) \cdot \delta z_0$, a first possibility is to use finite differences. However, it is crucial to use in practise adaptive step-length Runge-Kutta integrators and in this case, the finite differences are not well suited, since the two grids dynamically evaluated, involved for instance in the computation of $z(t_f, z_0)$ and $z(t_f, z_0 + \delta z_0)$, may be different and could lead to artificial non-differentiability. The key point is thus to force the grid to be the same. This is known as *Internal Numerical Derivative* (IND) [4]. Another possibility is to use *Automatic Differentiation* (AD) on the integration solver code. Because of the adaptive step-length, the code only defines a function which is piecewise differentiable but DA may lead to the same accuracy that IND for adaptive Runge-Kutta scheme [26]. The last option is to assemble explicitly the variational equations and compute their solutions, that is Jacobi fields. Since $\overrightarrow{H'}$ is evaluated along the current solution, in practise we need to integrate simultaneously the systems in z and δz (the dimension of the full system is $2n + 4n^2$ to get the whole derivative):

$$(\dot{z}(t), \dot{\delta z}(t)) = \left(\overrightarrow{\mathbf{H}}(z(t)), \overrightarrow{\mathbf{H}}'(z(t)) \cdot \delta z(t) \right), \quad (z(0), \delta z(0)) = (z_0, \delta z_0).$$

Considering a one-step explicit adaptive Runge-Kutta scheme, DA on the integration code and VAR (integration of the augmented variational system) are equivalent if the step-length control is done only on the components of z (and not on $(z, \delta z)$), and so in this case the following diagram commutes:

$$\begin{array}{ccc} \text{(IVP)} & \xrightarrow{\text{Numerical integration}} & z(t,z_0) \\ \text{Derivation}\downarrow & & \downarrow\text{Derivation} \\ \text{(VAR)} & \xrightarrow{\text{Numerical integration}} & \dfrac{\partial z}{\partial z_0}(t,z_0) \end{array}$$

where (IVP) stands for the Initial Value Problem: $\dot{z}(t) = \vec{\mathbf{H}}(z(t))$.

10.3.2 Numerical difficulties of the indirect simple shooting

We present in this section, the illustration of some numerical difficulties from the indirect simple shooting that can we found in [22]. This presentation is a motivation to the introduction to structural multiple shooting and homotopy techniques. We refer to [22] for more details and for a presentation of the homotopy methods in the frame of topological degree.

To illustrate the numerical issues, we consider the double integrator problem with L^1-minimization cost:

$$\begin{cases} \min \displaystyle\int_0^{t_f} |u(t)|\,\mathrm{d}t, \\ \dot{x}_1(t) = x_2(t),\ \dot{x}_2(t) = u(t),\ |u(t)| \leq \gamma, \\ x(0) = x_0, \quad x(t_f) = x_f, \end{cases}$$

with x_0, x_f belonging to $\mathbb{R}^2$, $t_f \geq 0$ and $\gamma > 0$ fixed. The pseudo-Hamiltonian (in normal form) is:

$$H(x,p,u) = p_1 x_2 + p_2 u - |u|.$$

The maximizing control is given by $u = 0$ if $|p_2| < 1$, $u = \gamma \mathrm{sign} p_2$ if $|p_2| > 1$ and $u \in [-\gamma\,,\gamma]$ if $|p_2| = 1$. We denote by $u(p_2)$ this maximizing control. At the end, we are leading to solve the following non-smooth boundary value problem (even badly defined when $p_2 = \pm 1$):

$$\begin{cases} \dot{x}_1(t) = x_2(t), \quad \dot{x}_2(t) = u(p_2(t)), \quad \dot{p}_1(t) = 0, \quad \dot{p}_2(t) = -p_1(t), \\ x(0) = x_0, \quad x(t_f) = x_f. \end{cases}$$

The shooting function is given in Figure 10.6 with the initial covector denoted $p_0 =: (\alpha, \beta)$.

We can notice that the shooting function is not continuous at the points $(0, \pm 1)$, it is not differentiable at the interfaces of the different regions, and it is moreover constant on the blue regions. The nine regions are characterized by different control structures, see Figure 10.7. One of the consequences about the use of Newton solver is that if the initial guess p_0 is not chosen in the right region, that is if it does not give a control with the optimal structure, then the algorithm may not converge. On this simple example, the algorithm can generate a point inside the regions where the shooting function is constant. But, in practise, we do not know in advance the optimal structure. One usage of the homotopy methods presented in the next section is to provide the optimal structure together with a good initial guess to make the Newton solver converge. Once the optimal structure is revealed, we can define an appropriate multiple shooting function to get with high accuracy a BC-extremal of the problem, see the next section for examples of this.

Remark 11. An alternative of the use of homotopic methods is the use of direct algorithms to determine the optimal structure together with a good enough initial guess, in order to define and solve the multiple shooting equations.

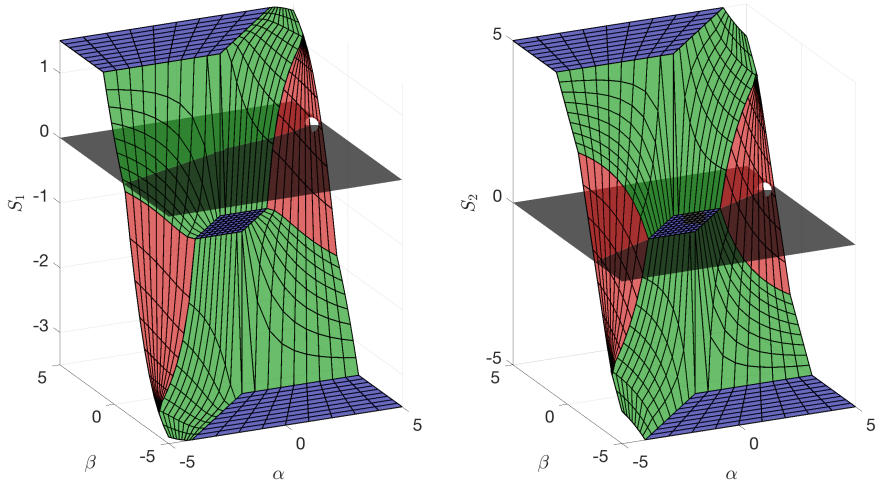


Fig. 10.6: Shooting function for the double integrator problem with L^1-minimization. On this example, $\gamma = 5$, $t_f = 1$, $x_0 = (-1, 0)$ and $x_f = (0, 0)$. The white ball corresponds to the solution, at the intersection between the graph of the shooting function and the planes $S_1 = 0$ and $S_2 = 0$.

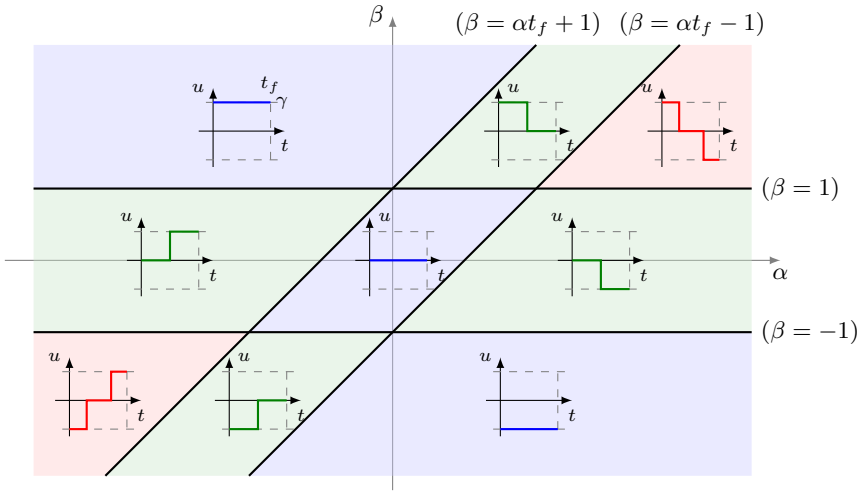


Fig. 10.7: Diagram of the different structures (on $[0\,, t_f]$) in the plane (α, β). Note that the extremals satisfy $p_2(t) = -\alpha t + \beta$ and the control is given by $u(p_2) = 0$ if $p_2 \in (-1\,, 1)$, $u(p_2) = \gamma \text{sign}(p_2)$ if $|p_2| > 1$ and $u(p_2) \in [-\gamma\,, \gamma]$ otherwise.

10.3.3 Structural indirect multiple shooting

Let us consider a scalar and affine control system of the form $H = H_0 + u\, H_1$ with the constraint $|u| \leq 1$. If the optimal extremal is the concatenation of bang arcs (where the control satisfies $|u| = 1$)

and singular arcs (where $H_1 = 0$), then, it is needed to use a multiple shooting method that we qualify as *structural indirect multiple method*. Let us give an example. Assume that the solution is composed of three arcs: a bang arc followed by a singular arc followed by another bang arc. We write the structure: bang-singular-bang. In this case, the shooting function must have in addition to the initial covector, two more unknowns: the two switching times. Since there are two more unknowns, we expect to find two more conditions to fulfill. These conditions are given by Proposition 5: since the singular surface Σ_s is invariant by the Hamiltonian flow of $\overrightarrow{\mathbf{H_s}}$, it is sufficient to impose that the first bang arc join Σ_s at the first switching time. Imposing the singular control on the second arc ensures that the extremal stays on Σ_s. Finally, for the third arc, it is sufficient to impose the bang control to leave the singular surface at the second switching time.

Remark 12. The maximization condition of the Pontryagin Maximum Principle gives us a stratification of the cotagent space where some regions are associated each to a unique Hamiltonian, while other regions like the singular surface are associated to several Hamiltonians. In our case, the region $H_1 < 0$ is associated to the Hamiltonian $H_0 - H_1$, the region $H_1 > 0$ is associated to the Hamiltonian $H_0 + H_1$, while in the region $H_1 = 0$ we have to deal with the two previous Hamiltonians and the Hamiltonian $\mathbf{H_s}$ defined by the singular control. Hence, finding a BC-extremal is a combination of dealing with the competition between Hamiltonians and the research of the optimal path in the cotangent space.

Example 4. Consider the following double integrator problem:

$$\begin{cases} \min \dfrac{1}{2}\displaystyle\int_0^{t_f} (x_1^2(t) + x_2^2(t))\,\mathrm{d}t, \quad t_f = 5, \\ \dot{x}_1(t) = x_2(t),\ \dot{x}_2(t) = u(t),\ u(t) \in [-1, 1], \\ x(0) = (0, 1). \end{cases}$$

Assume that we know that the optimal structure is composed of two arcs: a first arc with $u = -1$ followed by a singular arc. The pseudo-Hamiltonian in normal form is given by:

$$H(x_1, x_2, p_1, p_2, u) = -(x_1^2 + x_2^2)/2 + p_1 x_2 + p_2 u.$$

The singular arcs are of minimal order and given by the condition $p_2 = 0$. The singular surface is given by $\Sigma_s = \{p_2 = x_2 - p_1 = 0\}$ in the cotagent space $T^*\mathbb{R}^2 \simeq \mathbb{R}^2 \times \mathbb{R}^2$, and the singular control in feedback form is $u = x_1$. We have a competition between three Hamiltonian flows, respectively associated to

$$H_\pm \coloneqq H(x, p, \pm 1), \quad H_s \coloneqq H(x, p, x_1).$$

Knowing the bang-singular optimal structure, we can define the structural indirect multiple shooting function having as unknowns the initial covector $p_0 \in \mathbb{R}^2$ and the switching time $\tau \in \mathbb{R}$. The conditions are given by the transversality condition from Pontryagin Maximum Principle and by Proposition 5:

$$p_1(t_f) = 0, \quad p_2(t_f) = 0, \quad p_2(\tau) = 0, \quad x_2(\tau) - p_1(\tau) = 0.$$

We can notice that the condition $p_2(t_f) = 0$ is redundant since this is automatically checked for a singular extremal. We have three conditions for three unknowns, which defines the shooting function. The solution is given Figure 10.8.

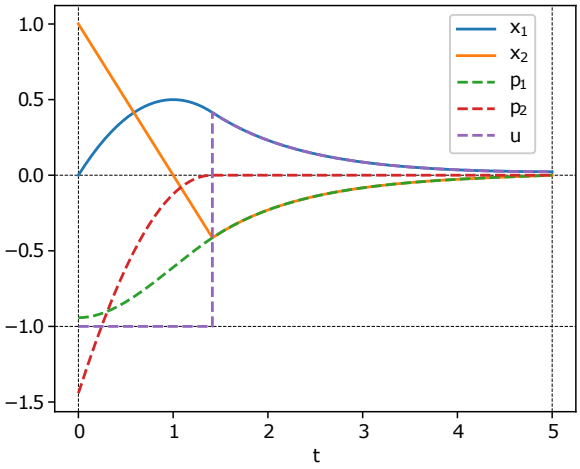


Fig. 10.8: States, covectors and controls for the double integrator problem.

Example 5. We consider the optimal control problem:

$$\begin{cases} \min \displaystyle\int_0^{t_f} x^2(t)\,\mathrm{d}t \\ \dot{x}(t) = u(t), \quad |u(t)| \leq 1, \\ x(0) = 1, \quad x(t_f) = 1/2. \end{cases}$$

with $t_f = 2$. We assume we know the optimal structure, composed of one bang arc $u = -1$, followed by a singular arc with $u = 0$, and then by another bang arc with $u = 1$. The pseudo-Hamiltonian in normal form is: $H(x, p, u) = -x^2 + pu$. The singular arcs are of minimal order, and given by $p = 0$. The singular surface is $\Sigma_s = \{p = x = 0\}$ in the cotangent space $T^*\mathbb{R} \simeq \mathbb{R} \times \mathbb{R}$, and the singular control in feedback form is given by $u = 0$. We have a competition between three Hamiltonian flows, given by $H_\pm := H(x, p, \pm 1)$, $H_s := H(x, p, 0)$. Knowing that the optimal structure is of the form bang-singular-bang, we can define the structural multiple shooting function having as unknowns the initial covector $p_0 \in \mathbb{R}$ and the two switching times $\tau_1, \tau_2 \in \mathbb{R}$. The conditions are given by the final condition $x(t_f) = 1/2$ and by Proposition 5:

$$x(t_f) = 1/2, \quad p(\tau_1) = 0, \quad x(\tau_1) = 0.$$

We have three conditions for three unknowns, which defines the shooting function. The solution is given Figure 10.9.

10.3.4 Homotopy methods

Homotopy methods may bu used to solve families of nonlinear equations. One common use consists in adding an artificial parameter to the set of nonlinear equations, embedding the original problem into a one-parameter family of equations, hoping that for a certain value of the parameter the problem is easy to solve in order to compute a sequence of zeros, modifying step by step the value

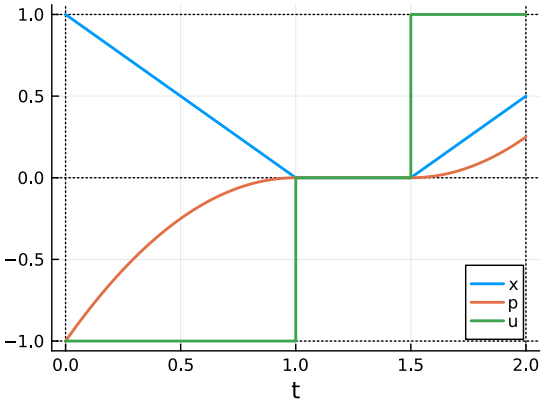


Fig. 10.9: State, covector and control for the second example.

of the parameter, until we get back to the original problem. The homotopic parameter and the nature of the deformation are heuristically chosen in practise, in relation with the constraints of the problem and the physical parameters governing it. This choice is guided on one hand by the simple problem to solve but also by the path of zeros itself, joining the simple problem to the original one, that we wish sufficiently smooth and converging to our target. Another interest of the homotopy is to describe the evolution of solutions with respect to some physical parameters already present inside the problem. In each case, we denote by λ the homotopic parameter, that we can consider in $[0\,,1]$. We talk about continuation when the homotopic parameter is monotone (increasing in our case). The homotopic approach is more general since λ may vary arbitrary. In this paper, we are interested in the homotopy methods in the context of optimal control to compute families of shooting functions or conjugate loci. We present first the homotopy method in the particular frame of geometric control, inspired by [38], and then give some algorithmic tools.

Let consider in a first part, that we have a one-parameter family of optimal control problems parameterized by $\lambda \in [0\,,1]$ of the following form: for a given λ, we minimize the cost

$$J_\lambda(x,u) \coloneqq \int_0^{t_f} f^0(x(t),u(t),\lambda)\,\mathrm{d}t$$

with a fixed final time $t_f > 0$. The state is governed by:

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

The simple limit conditions are given by:

$$x(0) = x_0, \quad x(t_f) = x_f,$$

with x_0 and x_f fixed. We assume that f and f^0 are smooth on $\mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}$. We consider the frame of Section 10.2.2. We can write the problem in its reduced form:

$$\min F_\lambda(u) \coloneqq \pi_{x^0}(\widetilde{E}_\lambda(u)) \text{ under the constraint } E_\lambda(u) = x_f,$$

and we seek u in the open set $\mathscr{U}_\lambda$ (assumed to be non-empty) of the Banach space $L^\infty([0\,,t_f],\mathbb{R}^m)$. The mappings E_λ and $\widetilde{E}_\lambda$ are respectively the endpoint and augmented endpoint mappings. The set $\mathscr{U}_\lambda$ is the admissible set of control laws. From the Lagrange rule, if the control u_λ is optimal, then, there exists $(\psi_\lambda,\psi_\lambda^0)\neq 0\in(\mathbb{R}^n)^*\times\mathbb{R}$ such that

$$\psi_\lambda^0\,F'_\lambda(u_\lambda)+\psi_\lambda\circ E'_\lambda(u_\lambda)=0.$$

Let us assume that there is no minimizing abnormal. Under this assumption, we can fix $\psi_\lambda^0=-1$ since $(\psi_\lambda,\psi_\lambda^0)$ is defined up to a scalar factor. We thus seek a pair (u_λ,ψ_λ) such that $G(\lambda,u_\lambda,\psi_\lambda)=0$, where G is defined by

$$G(\lambda,u,\psi)=\begin{pmatrix}-F'_\lambda(u)+\psi\circ E'_\lambda(u)\\ E_\lambda(u)-x_f\end{pmatrix}=\begin{pmatrix}\partial_u L_\lambda(u,\psi)\\ E_\lambda(u)-x_f\end{pmatrix},$$

where

$$\begin{aligned}L_\lambda(u,\psi)&=-F_\lambda(u)+\psi\cdot E_\lambda(u)\\ &=-\pi_{x^0}(\widetilde{E}_\lambda(u))+\psi\cdot E_\lambda(u)=(\psi,-1)\cdot\widetilde{E}_\lambda(u)\end{aligned}$$

is the Lagrangian. Let $(\bar\lambda,u_{\bar\lambda},\psi_{\bar\lambda})$ be a zero of G. Under our assumptions, G is regular and if the partial derivative of G with respect to (u,ψ) at the point $(\bar\lambda,u_{\bar\lambda},\psi_{\bar\lambda})$ is invertible, then, from the implicit function theorem, we can solve locally the equation $G(\lambda,u_\lambda,\psi_\lambda)=0$, and the solution (u_λ,ψ_λ) depends smoothly on λ. Let us analyze the conditions implying the invertibility of the previously mentioned partial derivative. The Jacobian matrix is

$$\frac{\partial G}{\partial(u,\psi)}(\lambda,u,\psi)=\begin{pmatrix}Q_\lambda & E'_\lambda(u)^*\\ E'_\lambda(u) & 0\end{pmatrix}\tag{10.9}$$

where Q_λ is the quadratic form associated to the augmented system:

$$Q_\lambda=\frac{\partial^2 L_\lambda}{\partial u^2}(u,\psi).$$

The operator matrix (10.9) is invertible if and only if the linear application $E'_\lambda(u)$ is surjective and the quadratic form Q_λ is non-degenerate on $\operatorname{Ker}E'_\lambda(u)$. The surjectivity of $E'_\lambda(u)$ means that the control u is not a singular point of the non-augmented endpoint mapping. But, for this optimal control problem, the singular controls of E_λ are associated to abnormal extremals, see the illustration Figure 10.3. Hence, the absence of minimizing abnormal trajectories is sufficient to ensure the surjectivity of $E'_\lambda(u)$. The fact that Q_λ is non-degenerate on $\operatorname{Ker}E'_\lambda(u)$ is related to the absence of conjugate times. We can conclude that as long as there is no minimizing abnormal trajectories and no conjugate times along the continuation, the continuation process works locally and the solution (u_λ,ψ_λ) is smooth with respect to the parameter.

However, we are interested by homotopic methods which do not restrict the homotopic parameter to be monotone. Besides, we want to present the methods in the Hamiltonian frame and consider techniques in finite dimension. Let us make the following assumption:

(A) For every $\lambda\in[0\,,1]$, there exists a normal Hamilton extremal $(x_\lambda,p_\lambda,u_\lambda)$.

Let $\lambda_0 \in [0\,,1]$. Under our assumptions, the mapping

$$(p_0, \lambda) \mapsto x(t_f, x_0, p_0, \lambda)$$

that maps (p_0, λ) to the value at time t_f of the projection on the state space of the solution $z = (x, p)$ of

$$\dot{z}(t) = \overrightarrow{H}(z(t), \lambda), \quad t \in [0\,, t_f], \quad z(0) = (x_0, p_0),$$

is a smooth implicit function in a neighbourhood of $(p_0(\lambda_0), \lambda_0)$. We define the *homotopic function* (see [21] for instance):

$$h(p_0, \lambda) = x(t_f, x_0, p_0, \lambda) - x_f$$

which is locally smooth. Let us assume that the solution $p_0(\lambda_0)$ of the n-dimensional shooting equation $h(\cdot, \lambda_0) = 0$ gives an extremal along which no Jacobi fields become vertical on $(0\,, t_f]$. In particular, there is no conjugate times on $(0\,, t_f]$ and so we have a solution which is locally $\mathscr{C}^0$-optimal for the optimal control problem with $\lambda = \lambda_0$. Typically, solving the family of problems consists in firstly computing a zero of $h(\cdot, \lambda_0) = 0$ for $\lambda_0 = 0$, then, following the path of zeros of h from $\lambda_0 = 0$ to the given target $\lambda = 1$. We introduce the following frame. Let us assume that the interior of the domain $\Omega \coloneqq h^{-1}(0) \subset \mathbb{R}^n \times [0\,, 1]$, is composed only of regular points of h and that the restriction of h on $\lambda = 0$ is a submersion:

$$\operatorname{rank} h'(p_0, \lambda) = n, \quad (p_0, \lambda) \in \operatorname{Int}(\Omega),$$

$$\operatorname{rank} \frac{\partial h}{\partial p_0}(p_0, \lambda)|_{\lambda=0} = n, \quad p_0 \in \mathbb{R}^n.$$

As a consequence, each connected component of the level line $\{h = 0\}$ is a one-dimensional submanifold of $\mathbb{R}^{n+1}$ called a *path of zeros*, starting from $\lambda = 0$. Each path of zeros is diffeomorphic either to $\mathbb{R}$ or $\mathbb{S}^1$, see [2]. For any $c = (p_0, \lambda) \in \Omega$, $\dim \operatorname{Ker} h'(c) = 1$ hence we can define the tangent vector $T(c)$ as the unique — up to the orientation — unitary vector of the kernel of $h'(c)$. The orientation is chosen such that the following determinant:

$$\det \begin{bmatrix} h'(c) \\ {}^t T(c) \end{bmatrix}$$

which never vanishes, has a constant sign along the path. This gives a parameterization by the arc length and the paths are computed integrating the differential equation:

$$c'(s) = T(c(s)), \quad c(0) = c_0 \in \{h = 0\},$$

with $c_0 = (p_0(0), 0)$ obtained by simple shooting for instance. One difficulty is to give sufficient conditions *à la Smale* which ensure the existence of a branch joining $\lambda = 0$ to $\lambda = 1$. Another difficulty is that for each value $\bar{\lambda}$ of the parameter, we must compare the associated cost for each component of $\{h = 0\} \cap \{\lambda = \bar{\lambda}\}$. This global aspect may be responsible of a lack of regularity of the value function which maps λ to the minimal cost.

For a given branch, there exists several possibilities that prevent the path to reach the target $\lambda = 1$, even if for every point $c = (p_0, \lambda)$ of the branch, c is regular. Since c is regular, the rank of $h'(c)$ is n with

$$h'(c) = \left[\frac{\partial h}{\partial p_0}(c) \; \frac{\partial h}{\partial \lambda}(c) \right] = \left[\frac{\partial x}{\partial p_0}(t_f, x_0, p_0, \lambda) \; \frac{\partial x}{\partial \lambda}(t_f, x_0, p_0, \lambda) \right].$$

Let us assume that for all s, $\lambda'(s) \neq 0$, which is equivalent to

$$\operatorname{rank} \frac{\partial x}{\partial p_0}(t_f, x_0, p_0(s), \lambda(s)) = n.$$

The parameter λ is thus increasing monotone (since $\lambda'(0) > 0$) and the only possibilities that prevent to reach $\lambda = 1$ are that the covector p_0, along the path of zeros, converges towards the boundary of Ω if it is bounded or goes to infinity in norm, see [20]. The path may converge to an abnormal extremal for limit value of λ. The following definition permits to deal with the second case when there exists $\bar{s}$ such that $\lambda'(\bar{s}) = 0$.

Definition 7 (Turning point). *A point $c(\bar{s}) = (p_0(\bar{s}), \lambda(\bar{s})) \in \{h = 0\}$ is a* turning point *if $\lambda'(\bar{s}) = 0$. This is equivalent to*

$$\operatorname{rank} \frac{\partial x}{\partial p_0}(t_f, x_0, p_0(\bar{s}), \lambda(\bar{s})) = n - 1.$$

A turning point is a point such that t_f is a conjugate time for the problem with $\lambda = \lambda(\bar{s})$. At an order 1 turning point, that is such that $\lambda''(\bar{s}) \neq 0$, there is a change in the variation of λ, whence the name. We can relate the first turning point to local optimality.

Definition 8. *We define $\bar{c} = c(\bar{s}) \in \{h = 0\}$ as the* first turning point *along the path starting from $c(0)$ if, $\lambda'(\bar{s}) = 0$, and if for all $s \in [0\,, \bar{s})$, the trajectory $t \mapsto x(t, x_0, p_0(s), \lambda(s))$ has no conjugate times on $(0\,, t_f]$.*

Theorem 5 ([17]). *Let $c(\bar{s}) \in \{h = 0\}$ be the first turning point of order 1. Then, for all $s > \bar{s}$, $|s - \bar{s}|$ small enough, there exists a conjugate time on $(0\,, t_f)$.*

Remark 13. In the proof of Theorem 5, cf. [17], is defined the extended homotopy

$$\widetilde{h}(p_0, \lambda, t_c) = (h(t_f, x_0, p_0, \lambda), \det \frac{\partial x}{\partial p_0}(t_c, x_0, p_0, \lambda)).$$

It is proved that the extended homotopy is well defined and regular in a neighbourhood of $(p_0(\bar{s}), \lambda(\bar{s}), t_f)$. We can thus use homotopy techniques to compute paths of zeros of the extended homotopy which provides the additional information of the first conjugate time.

There exists another difficulty for differential path following. When a path is diffeomorphic to $\mathbb{R}$, the extremities (if any) are singular points of h. The classification of such points starts by the following result which is a simple consequence of Morse lemma.

Proposition 6 ([2]). *Let $\bar{c} \in \{h = 0\}$ be an hyperbolic and non-degenerate singular point of h of corank one. Then, there exists coordinates $d_1, \ldots, d_{n+1}$ such that, in a neighbourhood of $\bar{c}$, $\{h = 0\}$ is given by*

$$d_1^2 - d_2^2 = 0, \quad d_3 = \cdots = d_{n+1} = 0.$$

In this case, the intrinsic second-order derivative writes, up to a scalar,

$$\bar{\mu}\, h''(\bar{c})|_{(\operatorname{Ker} h'(\bar{c}))^2} \in \operatorname{Sym}(2, \mathbb{R}) \subset \mathrm{M}_2(\mathbb{R})$$

where $\bar{\mu} \in (\mathbb{R}^n)^*$ is any non-zero covector with kernel $\operatorname{Im} h'(\bar{c})$. The hyperbolicity means that the symmetric matrix of order 2 is non-degenerate and has two eigenvalues of opposite signs. As a consequence, the path of zeros is locally made of two smooth curves intersecting transversally, resulting in a bifurcation at $\bar{c}$.

10.4 Examples Solved with the Julia `control-toolbox` Package

10.4.1 The Julia `control-toolbox` package

The control-toolbox ecosystem [12] gathers Julia packages for mathematical control and applications. The root package is `OptimalControl.jl` which aims to provide tools to solve optimal control problems by direct and indirect methods. For indirect methods we have developed tools for computing geometric control concepts as the flow of a Hamiltonian system or the Poisson brackets of Hamiltonians. For this we use in particular automatic differentiation (the Julia `ForwardDiff.jl` package).

We suppose here that the `control-toolbox` Julia package has been installed, see [12]. All the numerical experiments are reproducible downloading the codes available online at `https://github.com/control-toolbox/Kupka`.

10.4.2 The surface of revolution of minimum area

The first example is the well-known surface of revolution of minimum area problem which dates back to Euler [34, 31]. This is a problem of calculus of variations for which it is easy to have the analytic solutions. But here, as we use this simple problem to illustrate the use of the `control-toolbox`, we consider the optimal control version:

$$\begin{cases} \min \displaystyle\int_0^1 x(t)\sqrt{1+u^2(t)}\,\mathrm{d}t \\ \dot{x}(t) = u(t), \quad u(t) \in \mathbb{R}, \\ x(0) = 1, \quad x(1) = 2.5. \end{cases}$$

To define this problem in our package we have to type:

```
t0 = 0   # initial time
tf = 1   # final time
x0 = 1   # initial state
xf = 2.5 # final state
@def ocp begin
    t ∈ [ t0, tf ], time
    x ∈ R, state
    u ∈ R, control
    x(t0) == x0
    x(tf) == xf
    ẋ(t) == u(t)
    ∫(x(t)*(1 + u(t)^2)^(1/2)) → min
end
```

Here the maximization of the pseudo-Hamiltonian provides the control with respect to the state and the costate (or covector):

$$u(x,p) = \mathrm{sign}(x)\frac{p}{\sqrt{x^2-p^2}}.$$

Then, we can define the Hamiltonian $\mathbf{H}(x,p) = H(x,p,u(x,p))$ and can compute the flow of the Hamiltonian system by using the `Flow` function of the control-toolbox. At the end we can easily compute and plot this flow for different values of the initial costate, see Figure 10.10.

```
# Control in feedback form
u(x, p) = sign(x) * p / sqrt(x^2-p^2)

# The Flow function computes the Hamiltonian flow
ocp_flow = Flow(ocp, u, reltol=1e-10, abstol=1e-10);
```

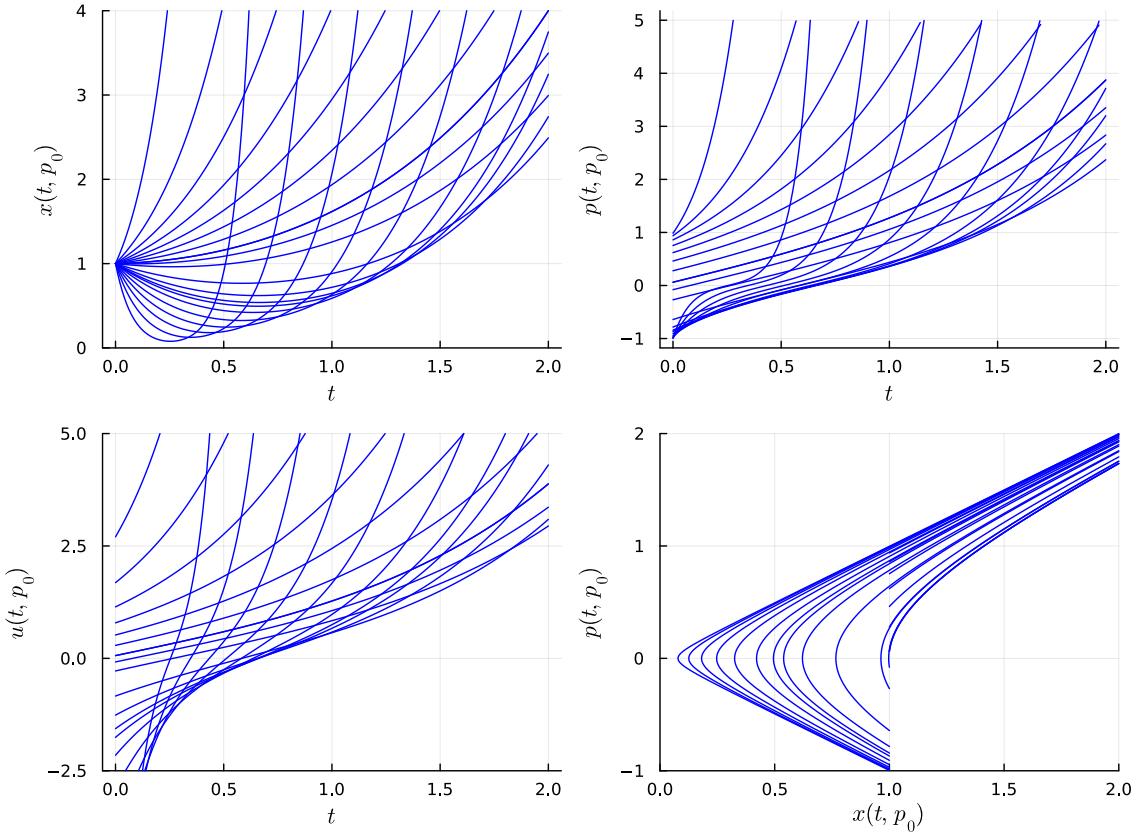


Fig. 10.10: States, costates and controls for $p_0 \in (-1\,,1)$.

Here, the shooting equation given by

$$S(p_0) = \pi(z(t_f, x_0, p_0)) - x_f = 0,$$

with $\pi(x,p) = x$, has two solutions: $p_0 = -0.9851$ and $p_0 = 0.5126$, see Figure 10.11.

```
# Shooting function
pi((x, p)) = x

tf = 1
```

```
xf = 2.5
S(p0) = (pi ∘ ocp_flow)(t0, x0, p0, tf) - xf

# Solve the shooting equation
p0 = -0.985 # First extremal
sol1_p0 = Roots.find_zero(S, (-0.99, -0.97))

p0 = 0.515 # Second extremal
sol2_p0 = Roots.find_zero(S, (0.5, 0.6))
```

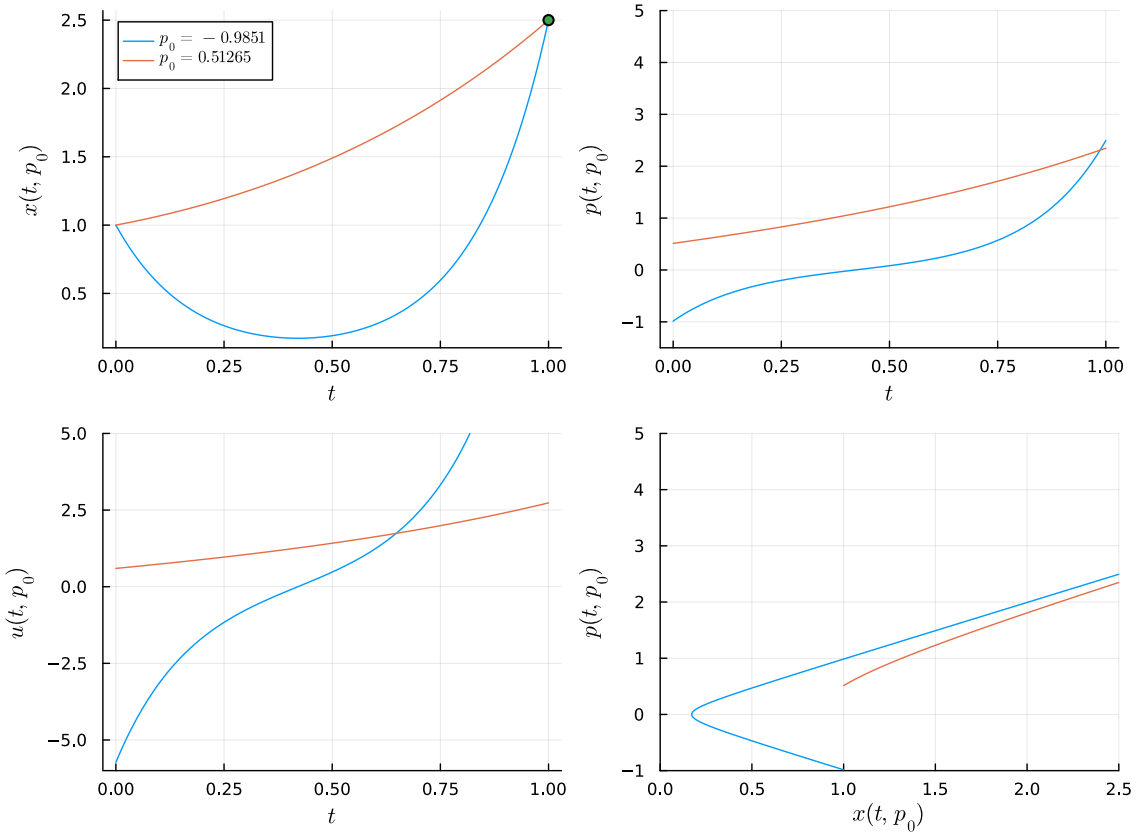


Fig. 10.11: Extremals for the problem of the surface of revolution of minimum area.

Now, we can compute the conjugate points along the two extremals. That's why we have to compute the flow $\delta z(t, p_0)$ of the Jacobi equation with the initial condition $\delta z(0) = (0, 1)$, i.e

$$\delta z(t, p_0) = \frac{\partial}{\partial p_0} z(t, p_0).$$

To compute conjugate points, we only need the first component:

$$\delta z(t, p_0)_1.$$

```
function jacobi_flow(t, p0)
    x(t, p0) = (pi ∘ ocp_flow)(t0, x0, p0, t)
    return ForwardDiff.derivative(p0 -> x(t, p0), p0)
end
```

The first conjugate time is then the first time τ such that

$$\delta x(\tau, p_0) = \delta z(\tau, p_0)_1 = 0,$$

with p_0 fixed. On Figure 10.12, one can see that the first extremal has a conjugate time smaller than $t_f = 1$ while for the second extremal, there is no conjugate time. Thus, the first extremal cannot be optimal.

```
# Compute the first conjugate time
p0 = sol1_p0
tau0 = Roots.find_zero(tau -> jacobi_flow(tau, p0), (0.4, 0.6))
```

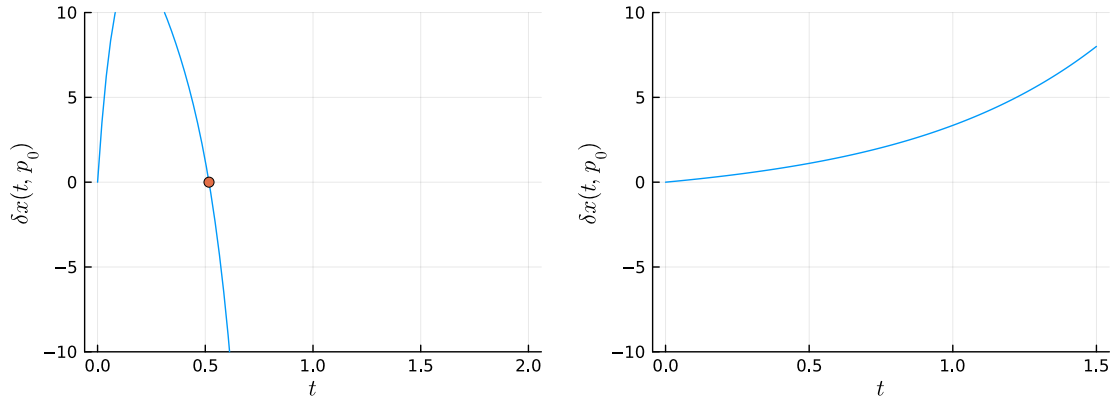


Fig. 10.12: (Left) Conjugate time for $p_0 = -0.9851$. (Right) No conjugate time for $p_0 = -0.51265$.

To conclude on this example, we compute the conjugate locus by using a path following algorithm. Define $F(\tau, p_0) = \delta x(\tau, p_0)$ and suppose that the partial derivative $\partial_\tau F(\tau, p_0)$ is invertible, then, by the implicit function theorem the conjugate time is a function of p_0. So, since here $p_0 \in \mathbb{R}$, we can compute them (see Figure 10.13) by solving the initial value problem for $p_0 \in [\alpha\,, \beta]$:

$$\dot{\tau}(p_0) = -\frac{\partial F}{\partial \tau}(\tau(p_0), p_0)^{-1}\, \frac{\partial F}{\partial p_0}(\tau(p_0), p_0), \quad \tau(\alpha) = \tau_0.$$

For the numerical experiment, we set $\alpha = -0.9995$, $\beta = -0.5$.

```
# conjugate points by path following
function conjugate_times_rhs_path(tau, p0)
    dF = ForwardDiff.gradient(y -> jacobi_flow(y...), [tau, p0])
    return -dF[2]/dF[1]
end
```

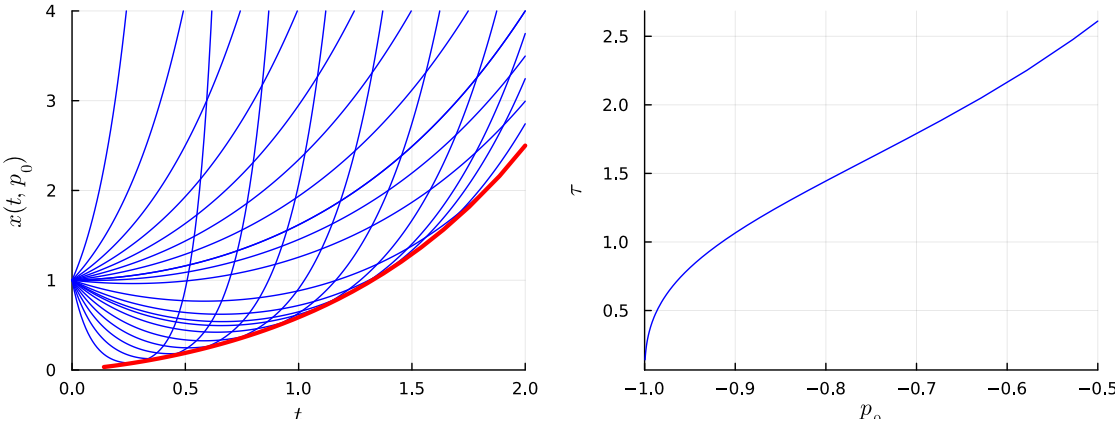


Fig. 10.13: The left graphic represents in blue the geodesic flow for $p_0 \in (-1\,,1)$, and in red the conjugate locus. The right graphic plots the conjugate time with respect to p_0.

10.4.3 Goddard Problem

For this advanced example, we consider the well-known Goddard problem [23, 34] which models the ascent of a rocket through the atmosphere, and we restrict here ourselves to vertical (one dimensional) trajectories. The state variables are the altitude r, speed v and mass m of the rocket during the flight, for a total dimension of 3. The rocket is subject to gravity g, thrust u and drag force D (function of speed and altitude). The final time t_f is free, and the objective is to reach a maximal altitude with a bounded fuel consumption.

We thus want to solve the optimal control problem in Mayer form

$$r(t_f) \to \max$$

subject to the controlled dynamics

$$\dot r = v, \quad \dot v = \frac{T_{\max}\, u - D(r,v)}{m} - g, \quad \dot m = -u,$$

and subject to the control constraint $u(t) \in [0\,,1]$. The initial state is fixed while only the final mass is prescribed. The dynamics may be written in the form: $\dot x(t) = F_0(x(t)) + u(t)\, F_1(x(t))$ with $x = (r, v, m)$. The Julia code to define this problem is simply:

```
t0 = 0      # initial time
r0 = 1      # initial altitude
v0 = 0      # initial speed
m0 = 1      # initial mass
mf = 0.6    # final mass to target

@def ocp begin

    tf, variable           # tf is free
    t ∈ [ t0, tf ], time
    x ∈ R³, state
    u ∈ R, control

    r = x₁
    v = x₂
    m = x₃

    x(t0) == [ r0, v0, m0 ]
    m(tf) == mf,                  (1)
    0 ≤ u(t) ≤ 1
    r(t) ≥ r0

    ẋ(t) == F0(x(t)) + u(t) * F1(x(t))

    r(tf) → max

end

# Dynamics
const Cd = 310
const Tmax = 3.5
const β = 500
const b = 2

F0(x) = begin
    r, v, m = x
    D = Cd * v^2 * exp(-β*(r - 1))
    return [ v, -D/m - 1/r^2, 0 ]
end

F1(x) = begin
    r, v, m = x
    return [ 0, Tmax/m, -b*Tmax ]
end
```

Remark 14. The Hamiltonian is affine with respect to the control, so singular arcs may occur.

We suppose that the optimal solution is composed of a bang arc with maximal control, followed by a singular arc and the final arc is with zero control. Note that the switching function vanishes along the singular. We are in position to solve the problem by an indirect shooting method. We first define the three control laws in feedback form and their associated flows. The control along the *minimal order* singular arcs is obtained as the quotient

$$u_s = -\frac{H_{001}}{H_{101}}$$

of the length three Poisson brackets:

$$H_{001} = \{H_0, \{H_0, H_1\}\}, \quad H_{101} = \{H_1, \{H_0, H_1\}\},$$

see Section 10.2.1 for more details.

Remark 15 (Poisson bracket and Lie derivative). The Poisson bracket $\{H, G\}$ of two Hamiltonians H and G is also given by the Lie derivative of G along the Hamiltonian vector field

$$X_H = (\nabla_p H, -\nabla_x H)$$

of H, that is

$$\{H, G\} = X_H \cdot G$$

which is the reason why we use the "@Lie" macro notation to compute Poisson brackets below.

With the help of the differential geometry primitives from the package `CTBase.jl` these expressions are straightforwardly translated into Julia code:

```
# Controls
u0 = 0                                  # off control
u1 = 1                                  # bang control

H0 = Lift(F0)                           # H0(x, p) = p' * F0(x)
H01  = @Lie { H0, H1 }
H001 = @Lie { H0, H01 }
H101 = @Lie { H1, H01 }
us(x, p) = -H001(x, p) / H101(x, p)     # singular control

# Flows
f0 = Flow(ocp, (x, p, tf) -> u0)
f1 = Flow(ocp, (x, p, tf) -> u1)
fs = Flow(ocp, (x, p, tf) -> us(x, p))
```

Then, we define the shooting function according to the optimal structure we have determined, that is a concatenation of three arcs.

```
x0 = [ r0, v0, m0 ] # initial state

function shoot!(s, p0, t1, t2, tf)

    x1, p1 = f1(t0, x0, p0, t1)
    x2, p2 = fs(t1, x1, p1, t2)
    xf, pf = f0(t2, x2, p2, tf)

    s[1] = constraint(ocp, :eq1)(x0, xf, tf) - mf # constraint (1)
    s[2:3] = pf[1:2] - [ 1, 0 ]                   # transversality
    s[4] = H1(x1, p1)           # H1 = H01 = 0
```

```
    s[5] = H01(x1, p1)          # at the entrance of the singular arc
    s[6] = H0(xf, pf)           # since tf is free
end
```

Finally, with a good initialization we can solve the shooting equations thanks to the MINPACK solver.

```
# auxiliary function with aggregated inputs
nle = (s, y) -> shoot!(s, y[1:3], y[4], y[5], y[6])

y = [ p0 ; t1 ; t2 ; tf ]     # initial guess
indirect_sol = MINPACK.fsolve(nle, y) # resolution of S(y) = 0

# we retrieve the costate solution together with the times
p0 = indirect_sol.x[1:3]
t1 = indirect_sol.x[4]
t2 = indirect_sol.x[5]
tf = indirect_sol.x[6]
```

We plot the solution given by the indirect shooting method on Figure 10.14. To do this, a nice feature of the `control-toolbox` is the concatenation of the flows:

```
f = f1 * (t1, fs) * (t2, f0)   # concatenation of the flows
flow_sol = f((t0, tf), x0, p0) # compute x, p and u solution
plot!(plt, flow_sol)           # plot the solution
```

10.5 Conclusion

We have seen here how it is relatively easy with our Julia `control-toolbox` package to solve optimal control problems and to compute geometric optimal control concepts. The main difficulties for computing the numerical solution of an optimal control problem by indirect methods are to know the optimal control structure and to have a good initial iterate. For the moment, we obtain this by solving the problem by a direct method. But another possibility is to use homotopy methods as described in Section 10.3.4. In the future, we'll develop a path following package for computing the path of zeros of an homotopy $h(z, \lambda) = 0$. Then we'll have in the same environment all the functionalities that the `bocop` [5] and `hampath` [25] softwares provide.

References

1. A. A. Agrachev & Y. L. Sachkov, *Control theory from the geometric viewpoint*, vol **87** of *Encyclopaedia of Mathematical Sciences*, Springer-Verlag, Berlin (2004), 412 pages.
2. E. Allgower & K. Georg, *Introduction to numerical continuation methods*, vol **45** of *Classics in Applied Mathematics*, Soc. for Industrial and Applied Math., Philadelphia, PA, USA, (2003), xxvi+388.
3. V. I. Arnold, *Mathematical methods of classical mechanics*. Translated from the Russian by K. Vogtmann and A. Weinstein. Second edition. Graduate Texts in Mathematics, **60**. Springer-Verlag, New York, 1989, 508 pages.

4. H. G. Bock, *Numerical treatment of inverse problems in chemical reaction kinetics*, vol 18 of *Springer Series in Chemical Physics*, Eds. K. H. Ebert, P. Deufl-hard & W. Jäger, in *Modelling of Chemical Reaction Systems*, Springer, Heidelberg, (1981), 102–125.
5. `www.bocop.org`.
6. A. V. Bolsinov & A. T. Fomenko, *Integrable Hamiltonian Systems, Geometry, Topology, Classification.* Chapman and Hall/CRC, London, 2004, 724 pages.
7. V. G. Boltyanski, *The Maximum Principle — How it came to be?*, Mathematisches Institut, München, Germany, Report No. 526, 1994.
8. B. Bonnard, J.-B. Caillau & E. Trélat, *Second order optimality conditions in the smooth case and applications in optimal control*, ESAIM Control Optim. Calc. Var., **13** (2007), no. 2, 207–236.
9. B. Bonnard & M. Chyba, *Singular trajectories and their role in control theory.* Vol **40** of *Mathematics & Applications*, Springer-Verlag, Berlin (2003), 357 pages.
10. B. Bonnard, O. Cots, Y. Privat & E. Trélat, *Zermelo navigation on the sphere with revolution metrics*, preprint, 2023.
11. B. Bonnard & I. Kupka, *Théorie des singularités de l'application entrée/sortie et optimalité des trajectoires singulières dans le problème du temps minimal.* Forum Math., **5** (1993), no. 2, pp. 111–159.
12. `https://github.com/control-toolbox`
13. L. Cesari, *Optimization-theory and applications: problems with ordinary differential equations*, vol 17 of *Applications of mathematics*, Springer-Verlag, New York, 1983, 542 pages.
14. F. Clarke, *Functional Analysis, Calculus of Variations and Optimal Control*, Graduate Texts in Mathematics 264, Springer-Verlag London 2013.
15. J.-B. Caillau, Z. Chen & Y. Chitour, *L^1-minimization for mechanical systems*, SIAM J. Control Optim. 54 (2016), no. 3, 1245–1265.
16. J.-B. Caillau, O. Cots & J. Gergaud, *Differential continuation for regular optimal control problems*, Optim. Methods Softw., **27** (2012), no 2, 177–196.
17. J.-B. Caillau & B. Daoud, *Minimum time control of the circular restricted three-body problem*, SIAM J. Control Optim., **50** (2012), no. 6, 3178–3202.
18. E. Casas and F. Tröltzsch, *Second order analysis for optimal control problems: Improving results expected from abstract theory*, SIAM J. Optim., **22** (2012), no. 1, 261–279.
19. O. Cots, *Contrôle optimal géométrique : méthodes homotopiques et applications.* Phd thesis, Institut Mathématiques de Bourgogne, Dijon, France, 2012.
20. J. Demailly, *Analyse numérique et équations différentielles*, Collection Grenoble Sciences. EDP Sciences (2006).
21. C. B. García & W. I. Zangwill, *An approach to homotopy and degree theory*, Math. Oper. Res., **4** (1979), no 4, 390–405.
22. J. Gergaud, *Sur la résolution numérique de problèmes de contrôle optimal à solution bang-bang via les méthodes homotopiques*, HDR thesis, Université de Toulouse, 2008.
23. R. H. Goddard, *A Method of Reaching Extreme Altitudes*, volume 71(2) of Smithsonian Miscellaneous Collections, Smithsonian institution, City of Washington (1919).
24. B. S. Goh, *Necessary conditions for singular extremals involving multiple control variables*, SIAM Journal on Control, 1966, vol. 4, no. 4, p. 716–731.
25. `www.hampath.org`.
26. E. Hairer, S. P. Nørsett & G. Wanner, *Solving Ordinary Differential Equations I, Nonstiff Problems*, vol 8 of *Springer Serie in Computational Mathematics*, Springer-Verlag, second edn (1993).
27. A. D. Ioffe & V. M. Tikhomirov, *Theory of extremal problems*, Elsevier, 2009.
28. V. Jurdjevic, *Geometric Control Theory*, Cambridge Studies in Advanced Mathematics, Cambridge University Press, Cambridge, 1996.
29. E. B. Lee & L. Markus, *Foundations of Optimal Control Theory*, Wiley, New York, 1967.
30. A. Lesfari, *Géométrie symplectique, calcul des variations et dynamique hamiltonienne*, Published in Great Britain by ISTE Editions Ltd, London, 2021.
31. D. Liberzon, *Calculus ov Variations and Optimal Control Theory*, Princeton University Press (2012).

32. J. Milnor, *Morse Theory*, Princeton University Press, 1963.
33. L. S. Pontryagin, V. G. Boltyanskiĭ, R. V. Gamkrelidze & 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, 360 pages.
34. H. Schättler & U. Ledzewicz, *Geometric optimal control: theory, methods and examples*, vol **38** of *Interdisciplinary applied mathematics*, Springer Science & Business Media, New York (2012), xiv+640.
35. H. Seywald and E.M. Cliff, *Goddard problem in presence of a dynamic pressure limit*. Journal of Guidance, Control, and Dynamics, 16(4):776–781 (1993).
36. H. J. Sussmann, *Geometry and optimal control*, *Mathematical control theory*, Springer, New York (1999), 140–198.
37. E. Trélat, *Contrôle optimal : théorie et applications*, Coll. Mathématiques concrètes, Vuibert, 2008, 250 pages.
38. E. Trélat, *Optimal control and applications to aerospace: some results and challenges*, J. Optim. Theory Appl., **154** (2012), no 3, 52.
39. R. B. Vinter, *Optimal Control*, Birkhäuser, Boston, 2000.
40. L. C. Young, *Lectures on the Calculus of Variations and Optimal Control Theory*, Saunders, Philadelphia, 1969.

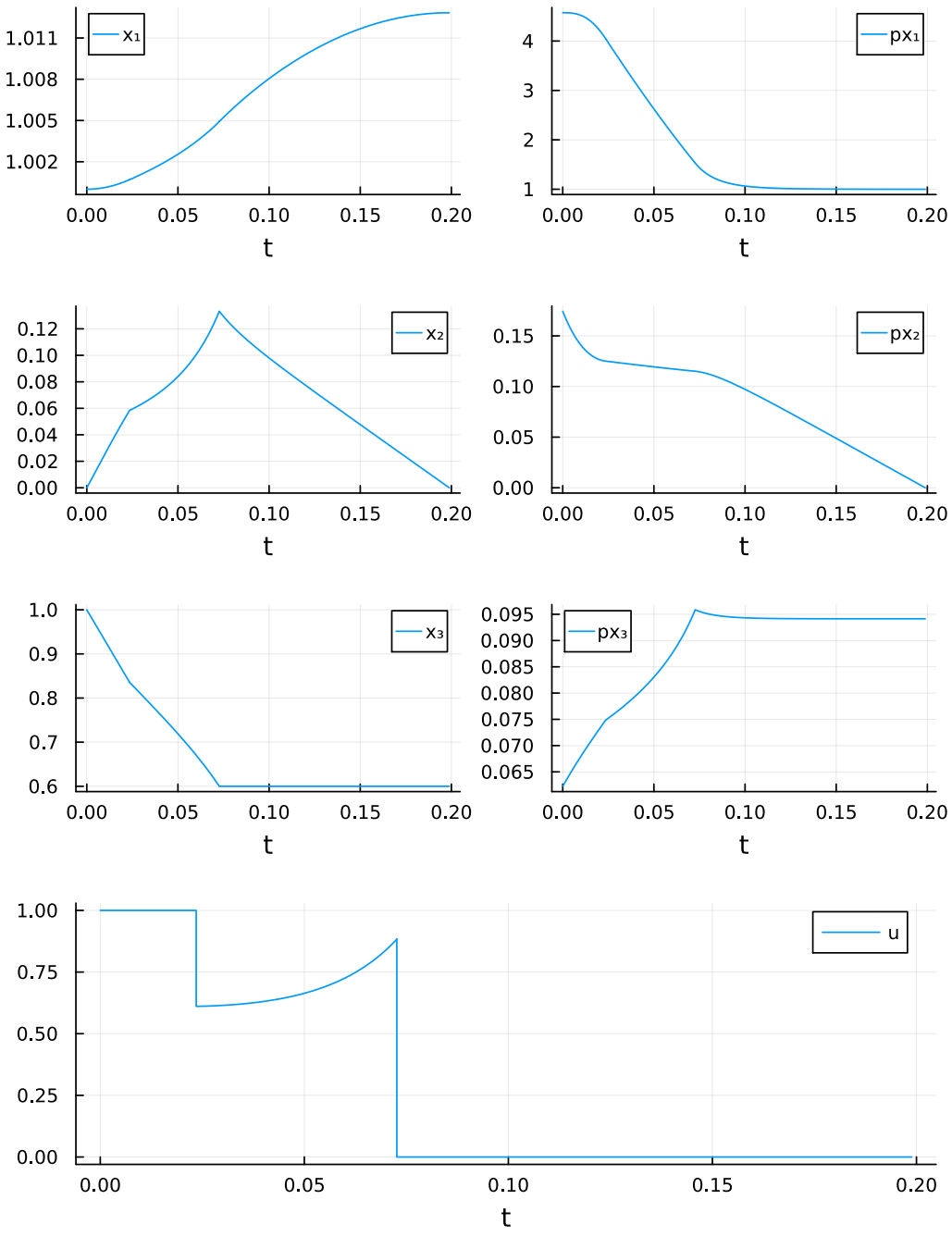


Fig. 10.14: *Solution of the Goddard problem.*

11

On the Reduction of a Spatially Hybrid Optimal Control Problem into a Temporally Hybrid Optimal Control Problem

Térence Bayen[1], Anas Bouali[2], Loïc Bourdin[3], and Olivier Cots[4]

[1] Avignon Université, Laboratoire de Mathématiques d'Avignon (EA 2151) F-84018
terence.bayen@univ-avignon.fr
[2] Avignon Université, Laboratoire de Mathématiques d'Avignon (EA 2151) F-84018
anas.bouali@univ-avignon.fr
[3] Institut de recherche XLIM. UMR CNRS 7252. Université de Limoges, France loic.bourdin@unilim.fr
[4] INP-ENSEEIHT-IRIT. UMR CNRS 5505. Toulouse Université, France olivier.cots@irit.fr

Summary. In this paper we consider a general spatially hybrid optimal control problem, for which a change of dynamics occurs when the state crosses the interface between two strata of a given partition of the state space. Given a (global) solution to this problem, we associate a temporally hybrid optimal control problem, for which changes of dynamics occur at (free) instants of time. We prove the following reduction result: under a strong transverse condition at the interfaces between strata, a (global) solution to the first problem is a L^1–local solution to the second one. As a corollary, we derive a spatially hybrid maximum principle from the application of a temporally hybrid maximum principle. Thanks to an explicit counterexample, we also prove that, when removing the strong transverse condition, the reduction result does not hold true in general, even if a weak transverse condition is satisfied. In addition, the analysis of this example demonstrates that the (global) solutions to the spatially and temporally hybrid optimal control problems are different.

11.1 Introduction

11.1.1 General context

Optimal control theory experienced fundamental advances in the late 1950's, following in particular the proof of the maximum principle by Pontryagin *et al.* (see [24]), which was a breakthrough enabling major advances in many fields of science such as in aerospace. The Pontryagin Maximum Principle (PMP in short) originally addressed optimal control problems (OCPs in short) governed by smooth control systems. It has now been extended to more complex settings, in particular to *hybrid* control systems in which the dynamics can be discontinuous (typically w.r.t. the state in the sense of Fillipov [19] but the nature of the discontinuities can be varied), and arising in many domains such as in nonsmooth mechanics [9], electricity [11], biology [2], viability theory [8], etc. In the literature, a multitude of hybrid settings can be found. Consequently, the generalization of the PMP to such contexts gives rise to different versions of the so-called Hybrid Maximum Principle (HMP in short). For instance, a change of dynamics can be controlled by an automaton [20, 22, 23, 25, 26], leading to the so-called *switched* control systems. In [3, 21], the dynamics depends on the state position

in a given partition of the state space, leading to the so-called *regional* control systems. We also refer to [13, 14] for a very general hybrid framework, called *multiprocesses*. Hence hybrid optimal control theory is very broad due to the diversity of discontinuities in hybrid controls systems. We refer to [3, 5, 13, 14, 15, 16, 17, 20, 21, 22, 23, 25, 26] and references therein.

11.1.2 Temporally and spatially hybrid optimal control problems

In this paper we will focus on two families of hybrid OCPs:

- *Temporally hybrid OCPs*. These are OCPs governed by a hybrid control system in which changes of dynamics occur at (free) instants of time. The number of these so-called *switching times* is fixed in advance, as well as each (smooth) dynamics between two consecutive switching times.
- *Spatially hybrid OCPs*. These are OCPs governed by a hybrid control system in which a change of dynamics occurs when the state crosses the interface between two strata (called *regions*) of a given partition of the state space. The times at which the state goes from one region to another are called *crossing times*.

Some versions of the HMP have been developed for temporally hybrid OCPs, thanks to an augmentation technique (see [15, 16, 17]). This approach consists in reducing the initial temporally hybrid OCP into an augmented classical OCP for which the PMP can be applied. Precisely, through this transformation, one may prove that a (global) solution to the first problem generates a L^1–local solution to the second one: hence the temporally HMP is obtained by, first, applying the PMP to the augmented classical OCP, and then by inverting the augmentation procedure.

For spatially hybrid OCPs, the situation is more intricate, because of the possibility of sliding modes of the trajectories along the interfaces between strata, and thus requires (a priori) the concept of Filippov's solution [19] in order to properly define a solution to the spatially hybrid control system. For this reason, the derivation of a spatially HMP usually requires in the literature the use of transverse conditions that can be classified into two categories:

- by *weak transverse condition* (see, e.g., [5, 7, 21]), we mean that the nominal trajectory is supposed to cross each interface transversally (i.e., not tangentially). This hypothesis involves, locally at each crossing point, the values of the nominal control (only).
- by *strong transverse condition* (see, e.g., [1, 6]), we mean that any admissible trajectory, locally at each crossing point of the nominal trajectory, should cross the interface transversally. This hypothesis involves, not only the values of the nominal control, but also all admissible control values.

We refer to Definition 2 and Remark 1 for mathematical details. Under each one of these transverse conditions, sliding modes are excluded. Nonetheless we refer to [3] where a spatially HMP is addressed in presence of sliding modes.

Now, under transverse conditions, there are essentially two approaches in order to derive a spatially HMP. A first one is based on the sensitivity analysis of the spatially hybrid control system (see [4, 21]), following the standard approach for proving the classical PMP. A second methodology consists in adapting (carefully) the augmentation technique of [17] to the spatially hybrid framework. Indeed a (global) solution to a spatially hybrid OCP may not generate a L^1–local solution to the corresponding augmented classical OCP. We refer to our previous paper [5] for an explicit counterexample, pointing out a possible misconception in the literature concerning the reduction of a spatially hybrid OCP. Actually, as explained in the next subsection, the aim of the present work

is to investigate this issue. Before that, let us mention that a new concept of local solution (weaker than L^1–local solution) was introduced in [5] to establish a spatially HMP through an augmentation technique.

11.1.3 Contributions and organization of this paper

In this paper, given a (global) solution to a general spatially hybrid OCP, we construct a temporally hybrid OCP in such a way that the number of (free) switching times coincides with the number of crossing times of the given (global) solution and, as well, the dynamics between two consecutive switching times coincides with the dynamics followed by the given (global) solution between two consecutive crossing times. The main result of this paper is Theorem 1 in Section 11.3, asserting that, under a strong transverse condition, the given (global) solution to the spatially hybrid OCP is a L^1–local solution to the temporally hybrid OCP. As a corollary, we derive a spatially HMP (Corollary 1) from the application of a temporally HMP.

Next, in Section 11.4, we develop a counterexample showing that Theorem 1 is no longer valid in absence of a strong transverse condition, even if a weak transverse condition is satisfied. Moreover, thanks to the spatially and temporally HMPs applied to this example, together with some numerical simulations, we show that the (global) solutions to the spatially and temporally hybrid OCPs are different.

This paper is organized as follows. In Section 11.2, we give recalls on classical OCPs and on the classical PMP, followed with reminders on temporally hybrid OCPs and on the temporally HMP. In Section 11.3, we introduce a general spatially hybrid OCP, as well as the notions of strong and weak transverse conditions. We prove our main result (Theorem 1) on the reduction of the general spatially hybrid OCP into a temporally hybrid OCP. This section is ended by stating a spatially HMP (Corollary 1). Section 11.4 develops a counterexample, showing that the reduction of Theorem 1 fails in absence of a strong transverse condition. A short conclusion ends this paper in Section 11.5. Finally the proof of the temporally HMP based on an augmentation technique is recalled in Appendix A for the reader's convenience.

11.1.4 Basic notations and functional framework

In this paper, for any positive integer $d \in \mathbb{N}^*$, we denote by $\langle \cdot, \cdot \rangle_{\mathbb{R}^d}$ (resp. $\| \cdot \|_{\mathbb{R}^d}$) the standard inner product (resp. Euclidean norm) of $\mathbb{R}^d$. For any subset $X \subset \mathbb{R}^d$, we denote by ∂X the boundary of X defined by $\partial X := \overline{X} \backslash \mathrm{Int}(X)$, where $\overline{X}$ and $\mathrm{Int}(X)$ stand respectively for the closure and the interior of X, by $\mathbb{1}_X : \mathbb{R}^d \to \mathbb{R}$ the indicator function of X defined by $\mathbb{1}_X(x) := 1$ if $x \in X$ and $\mathbb{1}_X(x) := 0$ otherwise. Given a closed convex subset $Y \subset \mathbb{R}^d$, the normal cone to Y at some point $y \in Y$ is defined by

$$\mathrm{N}_Y[y] := \{y'' \in \mathbb{R}^d \mid \forall y' \in Y,\ \langle y'', y' - y \rangle_{\mathbb{R}^d} \leq 0\}.$$

For any extended-real number $r \in [1, \infty]$ and any real interval $I \subset \mathbb{R}$, we denote by:

- $\mathrm{L}^r(I, \mathbb{R}^d)$ the usual Lebesgue space of r-integrable functions defined on I with values in $\mathbb{R}^d$, endowed with its usual norm $\| \cdot \|_{\mathrm{L}^r}$;
- $\mathrm{C}(I, \mathbb{R}^d)$ the standard space of continuous functions defined on I with values in $\mathbb{R}^d$, endowed with the standard uniform norm $\| \cdot \|_{\mathrm{C}}$;
- $\mathrm{AC}(I, \mathbb{R}^d)$ the subspace of $\mathrm{C}(I, \mathbb{R}^d)$ of absolutely continuous functions.

Now take $I = [0,T]$ for some $T > 0$. Recall that a partition of the interval $[0,T]$ is a finite set $\mathbb{T} = \{\tau_k\}_{k=1,\dots,N-1}$, for some integer $N \geq 2$, such that $0 =: \tau_0 < \tau_1 < \ldots < \tau_{N-1} < \tau_N := T$. In this paper a function $\gamma : [0,T] \to \mathbb{R}^d$ is said to be *piecewise absolutely continuous*, w.r.t. a partition $\mathbb{T} = \{\tau_k\}_{k=1,\dots,N-1}$ of the interval $[0,T]$, if γ is continuous at $\tau_0 := 0$ and $\tau_N := T$ and the restriction of γ over each open interval (τ_{k-1}, τ_k) admits an extension over $[\tau_{k-1}, \tau_k]$ that is absolutely continuous. If so, γ admits left and right limits at each $\tau_k \in (0,T)$, denoted by $\gamma^-(\tau_k)$ and $\gamma^+(\tau_k)$ respectively. In what follows we denote by $\mathrm{PAC}_{\mathbb{T}}([0,T],\mathbb{R}^d)$ the space of all piecewise absolutely continuous functions respecting a given partition $\mathbb{T}$ of $[0,T]$.

For a differentiable map $\psi : \mathbb{R}^d \to \mathbb{R}^{d'}$, with $d' \in \mathbb{N}^*$, we denote by $\nabla\psi(x) := \big(\nabla\psi_1(x) \dots \nabla\psi_{d'}(x)\big) \in \mathbb{R}^{d\times d'}$ the gradient of ψ at some $x \in \mathbb{R}^d$. We say that ψ is *submersive* at $x \in \mathbb{R}^d$ if the differential $D\psi(x) = \nabla\psi(x)^\top \in \mathbb{R}^{d'\times d}$ is surjective. Finally, when $(\mathscr{Z}, \mathrm{d}_{\mathscr{Z}})$ is a metric set, we denote by $\mathrm{B}_{\mathscr{Z}}(z,\nu)$ (resp. $\overline{\mathrm{B}}_{\mathscr{Z}}(z,\nu)$) the standard open (resp. closed) ball of $\mathscr{Z}$ centered at $z \in \mathscr{Z}$ and of radius $\nu > 0$.

11.2 Preliminaries

11.2.1 Reminders on classical OCPs and on the classical PMP

Let n, m, d and $\ell \in \mathbb{N}^*$ be four positive integers and $T > 0$ be a positive real number. In this section we consider a classical Mayer optimal control problem, with parameter and mixed initial-terminal state constraint, given by

$$\begin{array}{ll} \text{minimize} & \phi(x(0),x(T)), \\ \text{subject to} & (x,u,\lambda) \in \mathrm{AC}([0,T],\mathbb{R}^n) \times \mathrm{L}^\infty([0,T],\mathbb{R}^m) \times \mathbb{R}^d, \\ & \dot{x}(t) = f(x(t),u(t),\lambda), \quad \text{a.e. } t \in [0,T], \\ & g(x(0),x(T)) \in \mathrm{S}, \\ & u(t) \in \mathrm{U}, \quad \text{a.e. } t \in [0,T], \\ & \lambda \in \mathit{\Lambda}, \end{array} \tag{CP}$$

where the Mayer cost function $\phi : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$, the dynamics $f : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}^n$ and the constraint function $g : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^\ell$ are of class C^1, where $\mathrm{S} \subset \mathbb{R}^\ell$ and $\mathit{\Lambda} \subset \mathbb{R}^d$ are nonempty closed convex subsets and where $\mathrm{U} \subset \mathbb{R}^m$ is a nonempty subset. For simplicity, in the whole paper, we assume that g is submersive everywhere.

In Problem (CP), as usual in the literature, $x \in \mathrm{AC}([0,T],\mathbb{R}^n)$ is called the *state* (or the *trajectory*), $u \in \mathrm{L}^\infty([0,T],\mathbb{R}^m)$ is called the *control* and $\lambda \in \mathbb{R}^d$ is called the *parameter*. A triplet $(x,u,\lambda) \in \mathrm{AC}([0,T],\mathbb{R}^n) \times \mathrm{L}^\infty([0,T],\mathbb{R}^m) \times \mathbb{R}^d$ is said to be *admissible* for Problem (CP) if it satisfies all the constraints of Problem (CP). An admissible triplet (x^*,u^*,λ^*) is said to be a L^1*–local solution* to Problem (CP) if there exists $\eta > 0$ such that $\phi(x^*(0),x^*(T)) \leq \phi(x(0),x(T))$ for all admissible triplets (x,u,λ) satisfying

$$\|x - x^*\|_{\mathrm{C}} + \|u - u^*\|_{\mathrm{L}^1} + \|\lambda - \lambda^*\|_{\mathbb{R}^d} \leq \eta.$$

Finally the *Hamiltonian* $\mathscr{H} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}$ associated with Problem (CP) is, as usual, defined by $\mathscr{H}(x,u,\lambda,p) := \langle p, f(x,u,\lambda)\rangle_{\mathbb{R}^n}$ for all $(x,u,\lambda,p) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^d \times \mathbb{R}^n$ and the classical PMP [12, 24] can be stated as follows.

Proposition 1 (Classical PMP). *If (x^*, u^*, λ^*) is a L^1–local solution to Problem* (CP)*, then there exists a nontrivial pair $(p, p^0) \in \mathrm{AC}([0,T], \mathbb{R}^n) \times \mathbb{R}_+$ satisfying:*

(i) the Hamiltonian system

$$\dot{x^*}(t) = \nabla_p \mathscr{H}(x^*(t), u^*(t), \lambda^*, p(t)), \quad -\dot{p}(t) = \nabla_x \mathscr{H}(x^*(t), u^*(t), \lambda^*, p(t)),$$

for almost every $t \in [0,T]$;

(ii) the endpoint transversality condition

$$\begin{pmatrix} p(0) \\ -p(T) \end{pmatrix} = p^0 \nabla \phi(x^*(0), x^*(T)) + \nabla g(x^*(0), x^*(T))\xi,$$

for some $\xi \in \mathrm{N}_{\mathrm{S}}[g(x^*(0), x^*(T))]$;

(iii) the Hamiltonian maximization condition

$$u^*(t) \in \arg\max_{\omega \in \mathrm{U}} \mathscr{H}(x^*(t), \omega, \lambda^*, p(t)),$$

for almost every $t \in [0,T]$;

(iv) the averaged Hamiltonian gradient condition

$$\int_0^T \nabla_\lambda \mathscr{H}(x^*(s), u^*(s), \lambda^*, p(s))\, \mathrm{d}s \in \mathrm{N}_\Lambda[\lambda^*];$$

(v) the Hamiltonian constancy condition

$$\mathscr{H}(x^*(t), u^*(t), \lambda^*, p(t)) = c,$$

for almost every $t \in [0,T]$, for some $c \in \mathbb{R}$.

11.2.2 Temporally hybrid OCPs: terminology and temporally HMP

Let n, m, N, and $\ell \in \mathbb{N}^*$ be four positive integers, with $N \geq 2$, and $T > 0$ be a positive real number. In this section we consider a *temporally hybrid* Mayer optimal control problem, with mixed initial-terminal state constraint, given by

$$\begin{array}{ll} \text{minimize} & \phi(x(0), x(T)), \\ \text{subject to} & (x, u, \mathbb{T}) \in \mathrm{AC}([0,T], \mathbb{R}^n) \times \mathrm{L}^\infty([0,T], \mathbb{R}^m) \times \mathbb{R}^{N-1}, \\ & \dot{x}(t) = f_k(x(t), u(t)), \quad \text{a.e. } t \in (\tau_{k-1}, \tau_k), \quad \forall k \in \{1, \ldots, N\}, \\ & g(x(0), x(T)) \in \mathrm{S}, \\ & u(t) \in \mathrm{U}, \quad \text{a.e. } t \in [0,T], \\ & \mathbb{T} = \{\tau_k\}_{k=1,\ldots,N-1} \in \mathit{\Delta}, \\ & F_k(x(\tau_k)) = 0, \quad \forall k \in \{1, \ldots, N-1\}, \end{array} \tag{THP}$$

where $\mathit{\Delta} \subset \mathbb{R}^{N-1}$ is the nonempty closed convex subset defined by

$$\mathit{\Delta} := \{\mathbb{T} = \{\tau_k\}_{k=1,\ldots,N-1} \in \mathbb{R}^{N-1} \mid 0 =: \tau_0 \leq \tau_1 \leq \ldots \leq \tau_{N-1} \leq \tau_N := T\}.$$

The data assumptions and the terminology for Problem (THP) are the same as those for Problem (CP), with the addition that each dynamics $f_k : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ and each function $F_k : \mathbb{R}^n \to \mathbb{R}$ are of class C^1. In the sequel, for simplicity, we assume that each function F_k has no zero gradient.

The above setting is referred to as *hybrid* since, in contrast with Problem (CP), Problem (THP) involves several dynamics f_k. More precisely, this setting is referred to as *temporally hybrid* because the changes of dynamics in Problem (THP) are determined by the time variable t. In the literature, the instants τ_k at which the dynamics changes from f_k to f_{k+1} are usually called the *switching times*.

Finally the *Hamiltonian* $H_1 : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^{N-1} \times [0,T] \times \mathbb{R}^n \to \mathbb{R}$ associated with Problem (THP) is defined by $H_1(x,u,\mathbb{T},t,p) := \langle p, f_0(x,u,\mathbb{T},t)\rangle_{\mathbb{R}^n}$, where $f_0 : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^{N-1} \times [0,T] \to \mathbb{R}^n$ is defined by

$$f_0(x,u,\mathbb{T},t) := \sum_{k=1}^{N} f_k(x,u)\mathbb{1}_{(\tau_{k-1},\tau_k)}(t),$$

for all $(x,u,\mathbb{T},t,p) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^{N-1} \times [0,T] \times \mathbb{R}^n$. A temporally HMP [17] can be stated as follows.

Proposition 2 (Temporally HMP). *If $(x^*,u^*,\mathbb{T}^*)$ is a L^1–local solution to Problem* (THP)*, with $\mathbb{T}^* = \{\tau_k^*\}_{k=1,\dots,N-1} \in \mathrm{Int}(\Delta)$, then there exists a nontrivial pair $(p,p^0) \in \mathrm{PAC}_{\mathbb{T}^*}([0,T],\mathbb{R}^n) \times \mathbb{R}_+$ satisfying:*

(i) the Hamiltonian system

$$\dot{x^*}(t) = \nabla_p H_1(x^*(t),u^*(t),\mathbb{T}^*,t,p(t)),$$
$$-\dot{p}(t) = \nabla_x H_1(x^*(t),u^*(t),\mathbb{T}^*,t,p(t)),$$

for almost every $t \in [0,T]$;

(ii) the endpoint transversality condition

$$\begin{pmatrix} p(0) \\ -p(T) \end{pmatrix} = p^0 \nabla\phi(x^*(0),x^*(T)) + \nabla g(x^*(0),x^*(T))\xi,$$

for some $\xi \in \mathrm{N_S}[g(x^*(0),x^*(T))]$;

(iii) the Hamiltonian maximization condition

$$u^*(t) \in \arg\max_{\omega \in \mathrm{U}} H_1(x^*(t),\omega,\mathbb{T}^*,t,p(t)),$$

for almost every $t \in [0,T]$;

(iv) the discontinuity condition

$$p^+(\tau_k^*) - p^-(\tau_k^*) = \sigma_k \nabla F_k(x^*(\tau_k^*)),$$

for some $\sigma_k \in \mathbb{R}$, for all $k \in \{1,\dots,N-1\}$;

(v) the Hamiltonian constancy condition

$$H_1(x^*(t),u^*(t),\mathbb{T}^*,t,p(t)) = c,$$

for almost every $t \in [0,T]$, for some $c \in \mathbb{R}$.

Proof. For the reader's convenience, the proof of Proposition 2 is recalled in Appendix A. It is very similar to the one developed in [17], based on an augmentation procedure and on the application of the classical PMP recalled in Proposition 1.

11.3 Main Results

11.3.1 A spatially hybrid OCP

Let n, m and $\ell \in \mathbb{N}^*$ be three positive integers and $T > 0$ be a positive real number. In this section we consider a partition of the state space $\mathbb{R}^n$ given by

$$\mathbb{R}^n = \bigcup_{j \in \mathscr{J}} \overline{X_j},$$

where $\mathscr{J}$ is a (possibly infinite) family of indexes and where the nonempty open subsets $X_j \subset \mathbb{R}^n$, called *regions*, are disjoint. In this section we consider a *spatially hybrid* Mayer optimal control problem, with mixed initial-terminal state constraint, given by

$$\begin{array}{ll} \text{minimize} & \phi(x(0), x(T)), \\ \text{subject to} & (x,u) \in \mathrm{AC}([0,T],\mathbb{R}^n) \times \mathrm{L}^\infty([0,T],\mathbb{R}^m), \\ & \dot{x}(t) = h(x(t),u(t)), \quad \text{a.e. } t \in [0,T], \\ & g(x(0),x(T)) \in \mathrm{S}, \\ & u(t) \in \mathrm{U}, \quad \text{a.e. } t \in [0,T], \end{array} \tag{SHP}$$

where the data assumptions and the terminology for Problem (SHP) are the same as those for Problem (CP), with the addition that the dynamics $h : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is defined *regionally* by

$$\forall (x,u) \in \mathbb{R}^n \times \mathbb{R}^m, \quad h(x,u) := h_j(x,u) \quad \text{if } x \in X_j,$$

where the maps $h_j : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ are of class C^1. Note that $h(x,u)$ is not defined when $x \notin \cup_{j \in \mathscr{J}} X_j$ but this fact will have no impact on the rest of this paper thanks to transverse conditions (see Definition 2 and Remark 1 for details).

In contrast with Problem (THP), the above setting is referred to as *spatially hybrid* because the changes of dynamics in Problem (SHP) are determined by the state position $x(t)$ (and not by the time variable t).

11.3.2 Regular solutions to the spatially hybrid control system

Consider the spatially hybrid control system associated with Problem (SHP) given by

$$\dot{x}(t) = h(x(t),u(t)), \quad \text{for a.e. } t \in [0,T]. \tag{SHS}$$

Due to the discontinuities of the dynamics h, we need to precise some notions of *solution* to (SHS).

Definition 1 (Solution to (SHS)**).** *A pair $(x,u) \in \mathrm{AC}([0,T],\mathbb{R}^n) \times \mathrm{L}^\infty([0,T],\mathbb{R}^m)$ is said to be a* solution *to* (SHS) *if there exist a partition $\mathbb{T} = \{\tau_k\}_{k=1,\dots,N-1}$ of the interval $[0,T]$ and a* switching sequence *$j : \{1,\dots,N\} \to \mathscr{J}$ such that:*

(i) It holds that $x(t) \in X_{j(k)}$ for all $t \in (\tau_{k-1},\tau_k)$ and all $k \in \{1,\dots,N\}$, where $j(k) \neq j(k-1)$ for all $k \in \{2,\dots,N\}$;
(ii) It holds that $x(0) \in X_{j(1)}$ and $x(T) \in X_{j(N)}$;
(iii) It holds that $\dot{x}(t) = h_{j(k)}(x(t),u(t))$ for almost every $t \in (\tau_{k-1},\tau_k)$ and all $k \in \{1,\dots,N\}$.

In that case, to ease notation, we set $f_k := h_{j(k)}$ *and* $E_k := X_{j(k)}$ *for all* $k \in \{1, \dots, N\}$*. The times* τ_k *for* $k \in \{1, \dots, N-1\}$*, called* crossing times*, correspond to the instants at which the trajectory* x *goes from the region* E_k *to the region* E_{k+1}*, and thus* $x(\tau_k) \in \partial E_k \cap \partial E_{k+1}$*.*

Definition 2 (Regular solution to (SHS)**).** *Following the notations introduced in Definition 1, a solution* (x, u) *to* (SHS)*, associated with a partition* $\mathbb{T} = \{\tau_k\}_{k=1,\dots,N-1}$*, is said to be* regular *if the following conditions are both satisfied:*

(i) At each crossing time τ_k*, there exists a* C^1 *function* $F_k : \mathbb{R}^n \to \mathbb{R}$ *such that*

$$\exists \nu_k > 0, \quad \forall z \in \overline{\mathrm{B}}_{\mathbb{R}^n}(x(\tau_k), \nu_k), \quad \begin{cases} z \in E_k \Leftrightarrow F_k(z) < 0, \\ z \in \partial E_k \cap \partial E_{k+1} \Leftrightarrow F_k(z) = 0, \\ z \in E_{k+1} \Leftrightarrow F_k(z) > 0. \end{cases}$$

In particular it holds that $F_k(x(\tau_k)) = 0$*.*

(ii) At each crossing time τ_k*, the* U–strong transverse condition

$$\exists \mu_k \in (0, \nu_k],\ \forall (z, \omega) \in \overline{\mathrm{B}}_{\mathbb{R}^n}(x(\tau_k), \mu_k) \times \mathrm{U},\ \begin{cases} \langle \nabla F_k(z), f_k(z, \omega) \rangle_{\mathbb{R}^n} > 0, \\ \langle \nabla F_k(z), f_{k+1}(z, \omega) \rangle_{\mathbb{R}^n} > 0, \end{cases} \tag{TC}$$

is satisfied.

Remark 1. The U–strong transverse condition (TC) implies in particular that, at each crossing time τ_k, it holds that

$$\forall \omega \in \mathrm{U}, \quad \begin{cases} \langle \nabla F_k(x(\tau_k)), f_k(x(\tau_k), \omega) \rangle_{\mathbb{R}^n} > 0, \\ \langle \nabla F_k(x(\tau_k)), f_{k+1}(x(\tau_k), \omega) \rangle_{\mathbb{R}^n} > 0, \end{cases} \tag{TC$'$}$$

which translates into, for any admissible control value $\omega \in \mathrm{U}$, the trajectory x does not cross the boundary $\partial E_k \cap \partial E_{k+1}$ tangentially. In the case where U is compact, one can prove that (TC) and (TC$'$) are equivalent.

Other U–strong transverse conditions can be found in the literature (see [1, 6]). However, weak versions of transverse condition, in the sense that not all values of U are involved (but only the values of the nominal control u), have also been considered in the literature (see [3, 5, 21]). For instance, the weak transverse condition employed in [5] amounts to assume that, at each crossing time τ_k, there exist $\alpha_k > 0$ and $\beta_k > 0$ such that

$$\begin{cases} \langle \nabla F_k(x(\tau_k)), f_k(x(\tau_k), u(t)) \rangle_{\mathbb{R}^n} \geq \beta_k, \quad \text{a.e. } t \in (\tau_k - \alpha_k, \tau_k), \\ \langle \nabla F_k(x(\tau_k)), f_{k+1}(x(\tau_k), u(t)) \rangle_{\mathbb{R}^n} \geq \beta_k, \quad \text{a.e. } t \in (\tau_k, \tau_k + \alpha_k), \end{cases} \tag{TC$''$}$$

whereas the one in [21] requires left and right continuity of the nominal control u at each crossing time τ_k and that, in addition,

$$\begin{cases} \langle \nabla F_k(x(\tau_k)), f_k(x(\tau_k), u^+(\tau_k)) \rangle_{\mathbb{R}^n} > 0, \\ \langle \nabla F_k(x(\tau_k)), f_{k+1}(x(\tau_k), u^-(\tau_k)) \rangle_{\mathbb{R}^n} > 0. \end{cases} \tag{TC$'''$}$$

Note that (TC$''$) is a weaker transverse condition than (TC$'''$).

11.3.3 Reduction into a temporally hybrid OCP

In this section we will establish a correspondence between the *spatially* hybrid optimal control problem (SHP) and a *temporally* hybrid optimal control problem of type (THP). To this aim, let $(x^*, u^*) \in \mathrm{AC}([0,T],\mathbb{R}^n) \times \mathrm{L}^\infty([0,T],\mathbb{R}^m)$ be a (global) solution to Problem (SHP), that is moreover a regular solution to (SHS), associated with a partition $\mathbb{T}^* = \{\tau_k^*\}_{k=1,\ldots,N-1}$. Let us denote by E_k^*, f_k^* and F_k^* the corresponding regions, dynamics and local descriptions of $\partial E_k^* \cap \partial E_{k+1}^*$ (see Definitions 1 and 2). Hence we get that the triplet $(x^*, u^*, \mathbb{T}^*)$ is admissible for the temporally hybrid optimal control problem given by

$$
\begin{array}{ll}
\text{minimize} & \phi(x(0), x(T)), \\
\text{subject to} & (x, u, \mathbb{T}) \in \mathrm{AC}([0,T],\mathbb{R}^n) \times \mathrm{L}^\infty([0,T],\mathbb{R}^m) \times \mathbb{R}^{N-1}, \\
& \dot{x}(t) = f_k^*(x(t), u(t)), \quad \text{a.e. } t \in (\tau_{k-1}, \tau_k), \quad \forall k \in \{1,\ldots,N\}, \\
& g(x(0), x(T)) \in \mathrm{S}, \\
& u(t) \in \mathrm{U}, \quad \text{a.e. } t \in [0,T], \\
& \mathbb{T} = \{\tau_k\}_{k=1,\ldots,N-1} \in \mathit{\Delta}, \\
& F_k^*(x(\tau_k)) = 0, \quad \forall k \in \{1,\ldots,N-1\}.
\end{array}
\tag{THP$'$}
$$

We are now in a position to state and prove our main result, establishing the following correspondence between Problems (SHP) and (THP$'$).

Theorem 1. *If (x^*, u^*) is a (global) solution to Problem* (SHP)*, that is moreover a regular solution to* (SHS)*, associated with a partition $\mathbb{T}^* = \{\tau_k^*\}_{k=1,\ldots,N-1}$, then the triplet $(x^*, u^*, \mathbb{T}^*)$ is a L^1–local solution to Problem* (THP$'$).

Proof. Let us prove that there exists $\eta > 0$ such that $\phi(x^*(0), x^*(T)) \leq \phi(x(0), x(T))$ for any triplet $(x, u, \mathbb{T})$ admissible for Problem (THP$'$) satisfying

$$
\|x - x^*\|_{\mathrm{C}} + \|u - u^*\|_{\mathrm{L}^1} + \|\mathbb{T} - \mathbb{T}^*\|_{\mathbb{R}^{N-1}} \leq \eta. \tag{11.1}
$$

To this aim it is sufficient to prove that the pair (x, u) is admissible for Problem (SHP). First, note that the pair (x, u) satisfies all the constraints of Problem (SHP), except (maybe) the spatially hybrid control system (SHS). Since $\mathbb{T}^* \in \mathrm{Int}(\mathit{\Delta})$, taking $\eta > 0$ sufficiently small, we get that $\mathbb{T} \in \mathrm{Int}(\mathit{\Delta})$ (and thus $0 =: \tau_0 < \tau_1 < \ldots < \tau_{N-1} < \tau_N := T$). Then, since $\dot{x}(t) = f_k^*(x(t), u(t))$ for almost every $t \in (\tau_{k-1}, \tau_k)$ and for all $k \in \{1,\ldots,N\}$, it only remains to prove that $x(t) \in E_1^*$ for all $t \in [0, \tau_1)$, $x(t) \in E_k^*$ for all $t \in (\tau_{k-1}, \tau_k)$ and all $k \in \{2,\ldots,N-1\}$, and $x(t) \in E_N^*$ for all $t \in (\tau_{N-1}, T]$. We will only prove the first statement (in two steps) since the remaining ones can be proved in a similar way.

Step 1. By contradiction assume that, for all $\delta > 0$ and all $\eta_1 > 0$, there exist an admissible triplet $(x, u, \mathbb{T})$ for Problem (THP$'$) and $t' \in [\tau_1 - \delta, \tau_1)$ such that

$$
\begin{cases}
\|x - x^*\|_{\mathrm{C}} + \|u - u^*\|_{\mathrm{L}^1} + \|\mathbb{T} - \mathbb{T}^*\|_{\mathbb{R}^{N-1}} \leq \eta_1, \\
x(t') \notin E_1^*.
\end{cases}
$$

From the Lipschitz continuity of x^*, one can easily obtain that, taking $\delta > 0$ and $\eta_1 > 0$ small enough, it holds that $x(s) \in \overline{\mathrm{B}}_{\mathbb{R}^n}(x^*(\tau_1^*), \mu_1^*)$ for all $s \in [t', \tau_1)$. In particular, since $x(t') \notin E_1^*$, we get that $F_1^*(x(t')) \geq 0$. Since $F_1^*(x(\tau_1)) = 0$, we obtain

$$\int_{t'}^{\tau_1} \langle \nabla F_1^*(x(s)), f_1^*(x(s),u(s)) \rangle_{\mathbb{R}^n} \, \mathrm{d}s \leq 0.$$

Since $u(s) \in \mathrm{U}$ for almost every $s \in [0,T]$, the U–strong transverse condition (TC) contradicts the above inequality. Hence we deduce that there exist $0 < \delta < \tau_1^*$ and $\eta_1 > 0$ such that, for all triplets $(x,u,\mathbb{T})$ admissible for Problem (THP$'$) satisfying (11.1) with $\eta = \eta_1 > 0$, we have $x(t) \in E_1^*$ for all $t \in [\tau_1 - \delta, \tau_1)$.

Step 2. Since $x^*(t) \in E_1^*$ for all $t \in [0,\tau_1^*)$ and $E_1^* \subset \mathbb{R}^n$ is an open subset, we get from the continuity of x^* that there exists a (uniform) $\sigma^* > 0$ such that $\overline{\mathrm{B}}_{\mathbb{R}^n}(x^*(t),\sigma^*) \subset E_1^*$ for all $t \in [0, \tau_1^* - \frac{\delta}{2}]$. Then it is clear that there exists $0 < \eta \leq \eta_1$ sufficiently small so that, for any triplet $(x,u,\mathbb{T})$ admissible for Problem (THP$'$) satisfying (11.1), it holds that $\tau_1 - \delta < \tau_1^* - \frac{\delta}{2}$ and $x(t) \in \overline{\mathrm{B}}_{\mathbb{R}^n}(x^*(t),\sigma^*) \subset E_1^*$ for all $t \in [0, \tau_1^* - \frac{\delta}{2}]$ and thus for all $t \in [0, \tau_1 - \delta]$. This completes the proof.

Remark 2. We emphasize that the U–strong transverse condition (TC) plays a crucial role in the proof of Theorem 1. This latter reduces, in some sense, the *spatially* hybrid optimal control problem (SHP) into a *temporally* hybrid optimal control problem (THP$'$). Note that, even under a weak version of transverse condition (as the ones evoked in the second part of Remark 1), such a reduction is no longer possible in general. Section 11.4 is devoted to a counterexample that emphasizes this issue.

11.3.4 A spatially HMP as corollary

We are now in a position to state a spatially HMP for Problem (SHP), for which the corresponding Hamiltonian $H : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ is defined by $H(x,u,p) := \langle p, h(x,u) \rangle$ for all $(x,u,p) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n$, and whose proof follows directly from the application of Theorem 1 and Proposition 2.

Corollary 1 (Spatially HMP). *If (x^*,u^*) is a (global) solution to Problem* (SHP)*, that is moreover a regular solution to* (SHS)*, associated with a partition $\mathbb{T}^* = \{\tau_k^*\}_{k=1,\dots,N-1}$, then there exists a nontrivial pair $(p,p^0) \in \mathrm{PAC}_{\mathbb{T}^*}([0,T],\mathbb{R}^n) \times \mathbb{R}_+$ satisfying:*

(i) the Hamiltonian system

$$\dot{x^*}(t) = \nabla_p H(x^*(t),u^*(t),p(t)), \quad -\dot{p}(t) = \nabla_x H(x^*(t),u^*(t),p(t)),$$

for almost every $t \in [0,T]$;

(ii) the endpoint transversality condition

$$\begin{pmatrix} p(0) \\ -p(T) \end{pmatrix} = p^0 \nabla \phi(x^*(0),x^*(T)) + \nabla g(x^*(0),x^*(T))\xi,$$

for some $\xi \in \mathrm{N_S}[g(x^*(0),x^*(T))]$;

(iii) the Hamiltonian maximization condition

$$u^*(t) \in \arg\max_{\omega \in \mathrm{U}} H(x^*(t),\omega,p(t)),$$

for almost every $t \in [0,T]$;

(iv) the discontinuity condition

$$p^+(\tau_k^*) - p^-(\tau_k^*) = \sigma_k \nabla F_k^*(x^*(\tau_k^*)),$$

for some $\sigma_k \in \mathbb{R}$*, for all* $k \in \{1, \dots, N-1\}$*;*
(v) the Hamiltonian constancy condition

$$H(x^*(t), u^*(t), p(t)) = c,$$

for almost every $t \in [0, T]$*, for some* $c \in \mathbb{R}$*.*

Remark 3. As evoked in Remark 2 and showed in the next section with a counterexample, Theorem 1 is not valid when replacing the U–strong transverse condition (TC) by a weak transverse condition in general. Nevertheless, under the weak transverse condition (TC″), Corollary 1 remains valid (see [5]). In that context, its proof (which cannot be based on Theorem 1) follows a different approach.

11.4 Failure of Reduction: An Example

The aim of this section is to highlight, by means of an explicit counterexample, that Theorem 1 fails to hold in general without the U-strong transverse condition, even if the weak transverse condition (TC″) is satisfied. This example also emphasizes the differences between temporally and spatially hybrid OCPs, showing that their (global) solutions may be different.

11.4.1 A (global) solution to an explicit spatially hybrid OCP

Let $\rho > 0$ be a fixed parameter and, for all $a \in \mathbb{R}$, denote by $a_+ := \max(a, 0)$. In this section we study the explicit spatially hybrid OCP given by

$$\begin{array}{ll} \text{minimize} & -(x_1(2) - 2)^3 - \rho x_2(2), \\ \text{subject to} & (x, u) \in \mathrm{AC}([0,2], \mathbb{R}^2) \times \mathrm{L}^\infty([0,2], \mathbb{R}), \\ & \dot{x}(t) = h(x(t), u(t)), \quad \text{a.e. } t \in [0,2], \\ & x(0) = 0_{\mathbb{R}^2}, \\ & u(t) \in [-1, 1], \quad \text{a.e. } t \in [0,2], \end{array} \tag{$\mathrm{SHP}_{\mathrm{ex}}$}$$

where the spatially hybrid dynamics $h : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^2$ is defined regionally by

$$\forall (x, u) \in \mathbb{R}^2 \times \mathbb{R}, \quad h(x, u) := \begin{cases} (u + 2, (1 - x_1)_+^2) & \text{if } x \in X_1, \\ (u, (1 - x_1)_+^2) & \text{if } x \in X_2, \end{cases}$$

where $X_1 := (-\infty, 1) \times \mathbb{R}$ and $X_2 := (1, +\infty) \times \mathbb{R}$.

Note that the bidimensional-state Problem ($\mathrm{SHP}_{\mathrm{ex}}$) with Mayer cost can be rewritten as a unidimensional-state problem with Bolza cost, by replacing $x_2(2)$ by the Lagrange cost $\int_0^2 (1 - x_1(s))_+^2 \, \mathrm{d}s$.

Proposition 3. *Problem* ($\mathrm{SHP}_{\mathrm{ex}}$) *admits a unique (global) solution* (x^*, u^*)*, depicted in Figure 11.1, given by*

$$x^*(t) = \begin{cases} (t, \frac{1}{3}(1-(1-t)^3), & \forall t \in [0, \tau^*], \\ (t, \frac{1}{3}), & \forall t \in [\tau^*, 2], \end{cases} \qquad u^*(t) = \begin{cases} -1, & \text{a.e. } t \in [0, \tau^*], \\ +1, & \text{a.e. } t \in [\tau^*, 2], \end{cases}$$

with a unique crossing time $\tau^* := 1$ *at which the weak transverse condition* (TC'') *is satisfied, while the* U*-strong transverse condition* (TC) *with* $\mathrm{U} = [-1, 1]$ *is not. The corresponding cost is* $\mathscr{C}^* = -\frac{\rho}{3}$.

Proof. Due to the control system in the region X_1 and the control constraints set $\mathrm{U} = [-1, 1]$, it is clear that any admissible pair (x, u) of Problem ($\mathrm{SHP}_{\mathrm{ex}}$) is necessarily such that x intersects the interface $\Sigma := \{1\} \times \mathbb{R}$ at some (first) time $t_c \in [\frac{1}{3}, 1]$, and that $x(t) \in X_2 \cup \Sigma$ for all $t \in [t_c, 2]$. Thus $x_1(t) < 1$ for all $t \in [0, t_c)$, $x_1(t_c) = 1$ and $x_1(t) \geq 1$ for all $t \in (t_c, 2]$. In particular we get that $x_2(2) = x_2(t_c) = \int_0^{t_c} (1 - x_1(s))_+^2 \mathrm{d}s$. Hence, on the interval $[t_c, 2]$, the remaining objective is to maximize $x_1(2)$ and this can only be done by taking $u(t) = +1$ over $[t_c, 2]$, which gives $x_1(t) = x_1(t_c) + t - t_c$ for all $t \in [t_c, 2]$ and thus $x_1(2) = 3 - t_c$. Hence, to solve Problem ($\mathrm{SHP}_{\mathrm{ex}}$), we only need to solve the classical unidimensional-state optimal control problem (with free final time t_c and Bolza cost) given by

$$\begin{array}{ll} \text{minimize} & -(1-t_c)^3 - \rho \int_0^{t_c} (1 - x_1(s))_+^2 \, \mathrm{d}s, \\ \text{subject to} & (x_1, u, t_c) \in \mathrm{AC}([0, t_c], \mathbb{R}) \times \mathrm{L}^\infty([0, t_c], \mathbb{R}) \times [\frac{1}{3}, 1], \\ & \dot{x}_1(t) = u(t) + 2, \quad \text{a.e. } t \in [0, t_c], \\ & x_1(0) = 0, \\ & x_1(t_c) = 1, \\ & u(t) \in [-1, 1], \quad \text{a.e. } t \in [0, t_c]. \end{array}$$

Since the two terms in the objective are not in competition, it is clear that the unique (global) solution (x_1^*, u^*, t_c^*) to the above problem is given by $x_1^*(t) = t$ and $u^*(t) = -1$ over $[0, t_c^*]$ with $t_c^* = 1$. This concludes the proof. Indeed the fact that (TC'') is satisfied, while (TC) is not, and the value of the optimal cost $\mathscr{C}^*$ can be obtained with simple computations.

11.4.2 Failure of Theorem 1

At this step we have seen that the unique (global) solution (x^*, u^*) to Problem ($\mathrm{SHP}_{\mathrm{ex}}$) has a unique crossing time $\tau^* := 1$ at which the weak transverse condition (TC'') is satisfied, but not the U-strong transverse condition (TC) with $\mathrm{U} = [-1, 1]$. We now introduce the temporally hybrid OCP given by

$$\begin{array}{ll} \text{minimize} & -(x_1(2) - 2)^3 - \rho x_2(2), \\ \text{subject to} & (x, u, \tau) \in \mathrm{AC}([0, 2], \mathbb{R}^2) \times \mathrm{L}^\infty([0, 2], \mathbb{R}) \times \mathbb{R}, \\ & \dot{x}(t) = f_1(x(t), u(t)), \quad \text{a.e. } t \in (0, \tau), \\ & \dot{x}(t) = f_2(x(t), u(t)), \quad \text{a.e. } t \in (\tau, 2), \\ & x(0) = 0_{\mathbb{R}^2}, \\ & u(t) \in [-1, 1], \quad \text{a.e. } t \in [0, 2], \\ & \tau \in [0, 2], \\ & x(\tau) \in \{1\} \times \mathbb{R}, \end{array} \tag{$\mathrm{THP}_{\mathrm{ex}}$}$$

where the dynamics $f_k : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^2$ are defined by

$$f_1(x,u) := (u+2,(1-x_1)_+^2) \quad \text{and} \quad f_2(x,u) := (u,(1-x_1)_+^2).$$

for all $(x,u) \in \mathbb{R}^2 \times \mathbb{R}$.

Proposition 4. *For $\rho > 96$, the triplet (x^*,u^*,τ^*) is not a L^1–local solution to Problem* ($\mathrm{THP_{ex}}$).

Proof. For every $\varepsilon > 0$ small enough, take $\tau^\varepsilon := 1$ and consider the admissible control u^ε defined by

$$u^\varepsilon(t) := \begin{cases} +1, & \forall t \in [0,1+\varepsilon], \\ -1, & \forall t \in [1+\varepsilon,1+3\varepsilon], \\ +1, & \forall t \in [1+3\varepsilon,2]. \end{cases}$$

Then denote by x^ε the corresponding trajectory generated by the control system of Problem ($\mathrm{THP_{ex}}$). With basic computations, one can easily obtain that the corresponding cost satisfies

$$\mathscr{C}^\varepsilon = -\frac{\rho}{3} + \varepsilon^3\left(64 - \frac{2\rho}{3}\right) < \mathscr{C}^* = -\frac{\rho}{3},$$

since $\rho > 96$, and prove that

$$\lim_{\varepsilon \to 0} \|x^\varepsilon - x^*\|_{\mathrm{C}} + \|u^\varepsilon - u^*\|_{\mathrm{L}^1} + |\tau^\varepsilon - \tau^*| = 0,$$

which concludes the proof.

11.4.3 The (global) solution to Problem ($\mathrm{THP_{ex}}$)

In this section we want to synthesize a (global) solution to Problem ($\mathrm{THP_{ex}}$). We will see that, for $\rho > 0$ large enough, the (global) solution to Problem ($\mathrm{THP_{ex}}$) is the concatenation of three bang arcs and so, global solutions to Problem ($\mathrm{SHP_{ex}}$) and Problem ($\mathrm{THP_{ex}}$) are different. An intuitive reason for obtaining a different (global) solution to Problem ($\mathrm{THP_{ex}}$) is that, in contrast with Problem ($\mathrm{SHP_{ex}}$) where, after the crossing time, one cannot improve the cost $-\rho x_2(2)$ (since admissible trajectories cannot visit the region X_1 a second time), in Problem ($\mathrm{THP_{ex}}$), one can improve the cost $-\rho x_2(2)$.

We now turn to the resolution of Problem ($\mathrm{THP_{ex}}$). The Hamiltonian associated with Problem ($\mathrm{THP_{ex}}$) is defined by

$$H_1(x,u,\tau,t,p) := p_1(2+u)\mathbb{1}_{(0,\tau)}(t) + p_1 u \mathbb{1}_{(\tau,2)}(t) + p_2(1-x_1)_+^2,$$

for all $(x,u,\tau,t,p) \in \mathbb{R}^2 \times \mathbb{R} \times \mathbb{R} \times [0,T] \times \mathbb{R}^2$. Let $(\hat{x},\hat{u},\hat{\tau})$ be a (global) solution to Problem ($\mathrm{THP_{ex}}$) that we omit the proof of existence for brevity. Consider the nontrivial pair (p,p^0) provided by the temporally HMP (Proposition 2) that we consider normal ($p^0 \neq 0$) and that we renormalize so that $p^0 = 1$. The adjoint equation and transversality condition write then

$$\begin{cases} \dot{p}_1(t) = 2p_2(t)(1-\hat{x}_1(t))_+, & \text{a.e. } t \in [0,2], \\ \dot{p}_2(t) = 0, & \text{a.e. } t \in [0,2], \end{cases}$$

and

$$p_1(2) = 3(\hat{x}_1(2)-2)^2, \quad p_2(2) = \rho,$$

and the costate p can have a jump only at time $t = \hat{\tau}$. The outward unit normal vector to X_1 being the vector $(1, 0)$, we deduce that p_2 is continuous at $t = \hat{\tau}$, so, one has $p_2 \equiv \rho$ over $[0, 2]$ and only p_1 may have a jump at $t = \hat{\tau}$. Finally, from the Hamiltonian maximization condition, we get $\hat{u}(t) \in \mathrm{sign}(p_1(t))$ for almost every $t \in (0, 2)$.

In the sequel, we consider that $\hat{u}$ takes no intermediate value in $(-1, 1)$ and $B_\pm$ denotes a bang arc corresponding to $\hat{u} = \pm 1$ over some time interval. As usual in the literature (see [10]), a *switching time* of $\hat{u}$ stands for an instant $t_s \in (0, 2)$ such that $\hat{u}$ is non-constant in every neighborhood of t_s. To prevent confusion between a "switching time" corresponding to a change of dynamics and a "switching time" associated with a change in the control value, we employ distinct notations for each concept. Specifically, we denote by $\tau \in (0, 2)$ a switching time for change in the dynamics and by $t_s \in (0, 2)$ a switching time for change in the control value.

Proposition 5. *The optimal control $\hat{u}$ is one of the following three types:*
- $\{B_-B_+\}$ *with a single switching time at time $\hat{\tau} = 1$ and $\hat{u} = u^*$;*
- $\{B_+B_-B_+\}$ *with two consecutive switching times at time $\hat{\tau} = \frac{1}{3}$ and at some instant $t'_s \in (\frac{2}{3}, 2)$;*
- $\{B_-B_+B_-B_+\}$ *with three consecutive switching times at some instant $t_s \in (\frac{1}{3}, 1)$, at the switching time $\hat{\tau} \in (t_s, 1)$, and at some instant $t'_s \in (\frac{2}{3}, 2)$. Additionally, in this case, $\hat{x}_1(t_s) = \hat{x}_1(t'_s)$.*

Proof. First case: Suppose that $\hat{x}_1(2) \geq 2$. By a similar reasoning as in the proof of Proposition 3, we obtain that the only candidate for optimality is the triplet (x^*, u^*, τ^*). This gives the first item of Proposition 5.

Second case: Suppose that $\hat{x}_1(2) < 2$. Note that $\hat{x}_1(0) = 0$, that $\hat{\tau} \in [\frac{1}{3}, 1]$, and that $p_1(2) = 3(\hat{x}_1(2) - 2)^2 > 0$. Hence we necessarily have $\hat{u} = +1$ in a left neighborhood of $t = 2$. Now, $\hat{u}$ necessarily switches at some instant from -1 to the last bang arc $\hat{u} = +1$ (otherwise, we would have a contradiction with $\hat{x}_1(0) = 0$). Since p_1 is constant whenever $\hat{x}_1 \geq 1$ and $\dot{p}_1 > 0$ whenever $\hat{x}_1 < 1$, we deduce that there is a unique $t'_s \in (\frac{2}{3}, 2)$ such that $\hat{u}$ switches from -1 to $+1$ at time t'_s. Hence, one has $\hat{u}(t) = -1$ for almost every $t < t'_s$ sufficiently close to t'_s. The only possibility for the trajectory to reach the origin at time $t = 0$ (by reasoning backward in time) is to switch from $u = +1$ to $u = -1$ at the switching time $\hat{\tau}$ for which $\hat{x}_1(\hat{\tau}) = 1$ (otherwise, we would have $\hat{u}(t) = -1$ for a.e. $t < \hat{\tau}$ contradicting $\hat{x}_1(0) = 0$). Now, by a similar reasoning backward in time from $t = \hat{\tau}$, the monotonicity property of the switching function implies that the control possesses at most one switching time $t_s \in (0, \hat{\tau})$. It follows that, if t_s does not exist, then $\hat{u}$ is of type $\{B_+B_-B_+\}$ and that $\hat{\tau} = \frac{1}{3}$. Otherwise, if the optimal control $\hat{u}$ has a switching time $t_s \in (0, \hat{\tau})$, then it is of type $\{B_-B_+B_-B_+\}$ and using the constancy of H_1 at $t = t_s$ and at $t = t'_s$, we get that $\hat{x}_1(t_s) = \hat{x}_1(t'_s)$. This ends the proof.

At this step, u^* is a candidate for optimality for Problem ($\mathrm{THP}_{\mathrm{ex}}$) but, additionally, an optimal control can also be a sequence of three or four bang arcs. From the temporally HMP, we can completely characterize any optimal control for Problem ($\mathrm{THP}_{\mathrm{ex}}$) of type $\{B_+B_-B_+\}$ as we show in the next lemma.

Lemma 1. *Suppose that $\rho > \frac{96}{25}$ and consider a (global) solution $(\hat{x}, \hat{u}, \hat{\tau})$ to Problem ($\mathrm{THP}_{\mathrm{ex}}$) of type $\{B_+B_-B_+\}$. Then, one has $\hat{\tau} = \frac{1}{3}$ and $t'_s = \frac{5}{3} - \frac{\alpha}{2}$ where*

$$\alpha = \frac{2(2\rho + 36 - \sqrt{25\rho^2 - 96\rho})}{9\rho + 36}.$$

Proof. Setting $\alpha := \hat{x}_1(2)$, $\hat{x}_1$ can be expressed as follows:

$$\hat{x}_1(t) = \begin{cases} 3t & \text{if } t \in [0, \frac{1}{3}], \\ -t + \frac{4}{3} & \text{if } t \in [\frac{1}{3}, t'_s], \\ t + \alpha - 2 & \text{if } t \in [t'_s, 2], \end{cases}$$

where $t'_s = \frac{5}{3} - \frac{\alpha}{2}$. Using the constancy of H_1 at $t = 2$ and at $t = t'_s$, we find that α is a solution to the algebraic equation $3(\alpha-2)^2+\rho(1-\alpha)_+^2 = \rho(1-x_1(t'_s))^2$, or equivalently $3(\alpha-2)^2+\rho(1-\alpha)_+^2 = \rho\left(\frac{4}{3} - \frac{\alpha}{2}\right)^2$. Solving this equation gives us the desired value of α which ends the proof.

By using a similar argumentation, one can also characterize any optimal control for Problem ($\mathrm{THP}_{\mathrm{ex}}$) of type $\{B_-B_+B_-B_+\}$ as follows.

Lemma 2. *Suppose that $\rho > \frac{48}{9}$ and consider a (global) solution $(\hat{x}, \hat{u}, \hat{\tau})$ to Problem ($\mathrm{THP}_{\mathrm{ex}}$) of type $\{B_-B_+B_-B_+\}$. Then, one has $t_s = \frac{3\alpha}{4} - \frac{1}{2}$, $\hat{\tau} = \frac{\alpha}{2}$ and $t'_s = \frac{3}{2} - \frac{\alpha}{4}$ where*

$$\alpha = \frac{2(-\rho + 48 + 2\sqrt{9\rho^2 - 48\rho})}{7\rho + 48}.$$

Proof. For brevity, we omit the proof which is analogous to the proof of the previous lemma.

Thanks to the preceding results, for $\rho > \frac{48}{9}$, there are three types of candidate optimal control for Problem ($\mathrm{THP}_{\mathrm{ex}}$). We end up by numerical simulations to determine which one is optimal:

- For small values of $\rho > 0$ (typically, $\rho = 0.1$), then the optimal control for Problem ($\mathrm{THP}_{\mathrm{ex}}$) is u^* (see Figure 11.1) ;
- For larger values of $\rho > 0$ (typically, $\rho = 10, 30, 100$), then the optimal control for Problem ($\mathrm{THP}_{\mathrm{ex}}$) is of type $\{B_+B_-B_+\}$ (see Figure 11.2 for $\rho = 10$, Figure 11.3 for $\rho = 30$ and Figure 11.4 for $\rho = 30$).

In Table 11.1, we indicate the cost $\widehat{\mathscr{C}}$ associated with $\hat{u}$ in Problem ($\mathrm{THP}_{\mathrm{ex}}$), as well as the optimal cost $\mathscr{C}^* = -\frac{\rho}{3}$ of Problem ($\mathrm{SHP}_{\mathrm{ex}}$). In particular, when $\rho > 0$ increases, note that $|\widehat{\mathscr{C}} - \mathscr{C}^*|$ also increases.

ρ	0.1	10	30	100
$\widehat{\mathscr{C}}$	-0.33	-6.13	-34.55	-148.02
$\hat{u}$	$\{B_-B_+\}$	$\{B_+B_-B_+\}$	$\{B_+B_-B_+\}$	$\{B_+B_-B_+\}$
$\mathscr{C}^*$	-0.33	-3.33	-10	-33.33

Table 11.1: Comparison of the optimal costs of Problem ($\mathrm{THP}_{\mathrm{ex}}$) and Problem ($\mathrm{SHP}_{\mathrm{ex}}$). For large values of $\rho > 0$, both problems have different (global) solutions.

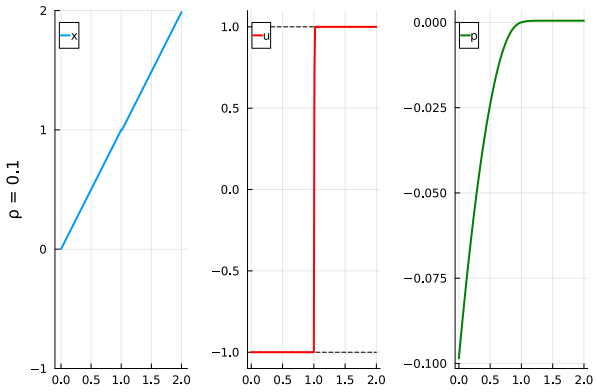


Fig. 11.1: Global solution to Problem ($\mathrm{THP}_{\mathrm{ex}}$) and Problem ($\mathrm{SHP}_{\mathrm{ex}}$) for $\rho = 0.1$ (plot of trajectory x_1, control u and costate p_1).

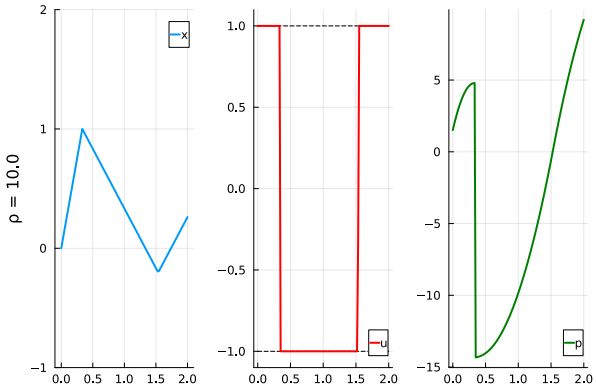


Fig. 11.2: Global solution $\{B_+B_-B_+\}$ to Problem ($\mathrm{THP}_{\mathrm{ex}}$) for $\rho = 10$ obtained by a direct numerical method (plot of trajectory x_1, control u and costate p_1 with a jump at $\hat{\tau} = \frac{1}{3}$).

11.5 Conclusion and Further Comments

Thanks to Theorem 1, we expect to have clarified the connection between spatially and temporally hybrid OCPs under a strong transverse condition. However, the example developed in Section 11.4 shows that, in general (even under a weak transverse condition), a (global) solution to a spatially hybrid OCP is not a (global) solution, and not even a L^1–local solution, to the corresponding temporally hybrid OCP, and the value functions of the two problems may widely differ. This corroborates the fact that these two hybrid frameworks are different. In particular, the trajectories generally differ due to the presence of strata in the spatially hybrid case.

We would like to insist on the fact that, even if the necessary optimality conditions are the same in both the spatially HMP and temporally HMP (Hamiltonian maximization condition, transver-

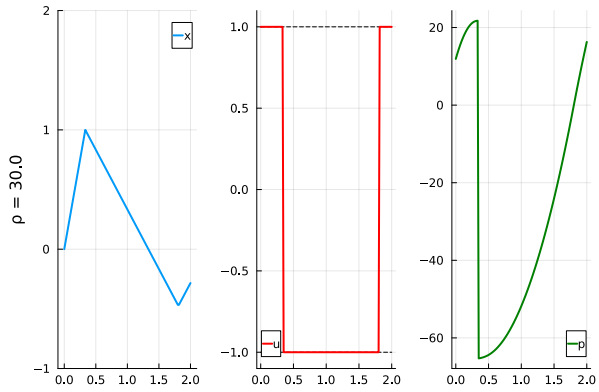


Fig. 11.3: Global solution $\{B_+B_-B_+\}$ to Problem $(\mathrm{THP}_{\mathrm{ex}})$ for $\rho = 30$ obtained by a direct numerical method (plot of trajectory x_1, control u and costate p_1 with a jump at $\hat{\tau} = \frac{1}{3}$).

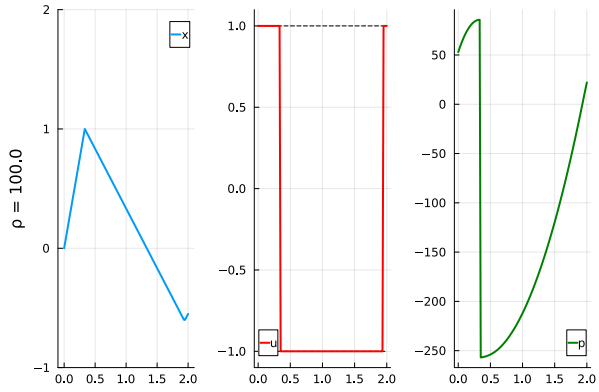


Fig. 11.4: Global solution $\{B_+B_-B_+\}$ to Problem $(\mathrm{THP}_{\mathrm{ex}})$ for $\rho = 100$ obtained by a direct numerical method (plot of trajectory, control, and costate p_1 with a jump at $\hat{\tau} = \frac{1}{3}$).

sality condition, discontinuity condition, etc.), the frameworks being different, they do not describe the same set of extremal solutions in general.

Future works could investigate the possibility of deriving a spatially HMP from the application of a version of the PMP that handles running state constraints (for which a Borel measure is involved, that could make it possible to retrieve the discontinuity condition of the costate).

A Proof of Proposition 2

Let $(x^*, u^*, \mathbb{T}^*) \in \mathrm{AC}([0,T],\mathbb{R}^n) \times \mathrm{L}^\infty([0,T],\mathbb{R}^m) \times \mathbb{R}^{N-1}$ be a L^1–local solution to Problem (THP) with $\mathbb{T}^* = \{\tau_k^*\}_{k=1,\dots,N-1} \in \mathrm{Int}(\Delta)$. In particular it holds that $0 =: \tau_0^* < \tau_1^* < \ldots < \tau_{N-1}^* <$

$\tau_N^* := T$ and there exists $\eta_0 > 0$ such that $\phi(x^*(0), x^*(T)) \leq \phi(x(0), x(T))$ for all triplets $(x, u, \mathbb{T})$ admissible for Problem (THP) satisfying

$$\|x - x^*\|_{\mathrm{C}} + \|u - u^*\|_{\mathrm{L}^1} + \|\mathbb{T} - \mathbb{T}^*\|_{\mathbb{R}^{N-1}} \leq \eta_0. \tag{11.2}$$

The proof of Proposition 2 is done in four steps.

Step 1: augmentation procedure. Roughly speaking, the goal here is to reduce Problem (THP) into a classical optimal control problem of type (CP). To this aim we introduce

$$y_k^*(s) := x^*(\tau_{k-1}^* + (\tau_k^* - \tau_{k-1}^*)s) \ \text{ and } \ v_k^*(s) := u^*(\tau_{k-1}^* + (\tau_k^* - \tau_{k-1}^*)s),$$

for all $s \in [0,1]$ and all $k \in \{1, \dots, N\}$. We get that the triplet $(y^*, v^*, \mathbb{T}^*)$ is admissible for the classical optimal control problem given by

$$\begin{array}{ll} \text{minimize} & \phi^*(y(0), y(1)), \\ \text{subject to} & (y, v, \mathbb{T}) \in \mathrm{AC}([0,1], \mathbb{R}^{nN}) \times \mathrm{L}^\infty([0,1], \mathbb{R}^{mN}) \times \mathbb{R}^{N-1}, \\ & \dot{y}(s) = f^*(y(s), v(s), \mathbb{T}), \quad \text{a.e. } s \in [0,1], \\ & g^*(y(0), y(1)) \in \mathrm{S}^*, \\ & v(s) \in \mathrm{U}^N, \quad \text{a.e. } s \in [0,1], \\ & \mathbb{T} \in \Delta, \end{array} \tag{CP'}$$

where $\phi^* : \mathbb{R}^{nN} \times \mathbb{R}^{nN} \to \mathbb{R}$, $f^* : \mathbb{R}^{nN} \times \mathbb{R}^{mN} \times \mathbb{R}^{N-1} \to \mathbb{R}^{nN}$ and $g^* : \mathbb{R}^{nN} \times \mathbb{R}^{nN} \to \mathbb{R}^{\ell^*}$ are defined by $\phi^*(y^0, y^1) := \phi(y_1^0, y_N^1)$,

$$f^*(y, v, \mathbb{T}) := \Big((\tau_1 - \tau_0) f_1(y_1, v_1), \dots, (\tau_N - \tau_{N-1}) f_N(y_N, v_N) \Big),$$

and

$$g^*(y^0, y^1) := \Big(g(y_1^0, y_N^1), y_2^0 - y_1^1, \dots, y_N^0 - y_{N-1}^1, F_1(y_1^1), \dots, F_{N-1}(y_{N-1}^1) \Big),$$

for all $y^0 = (y_1^0, \dots, y_N^0)$, $y^1 = (y_1^1, \dots, y_N^1) \in \mathbb{R}^{nN}$, $y = (y_1, \dots, y_N) \in \mathbb{R}^{nN}$, $v = (v_1, \dots, v_N) \in \mathbb{R}^{mN}$ and $\mathbb{T} = \{\tau_k\}_{k=1,\dots,N-1} \in \mathbb{R}^{N-1}$, and where $\mathrm{S}^* := \mathrm{S} \times \{0_{\mathbb{R}^n}\}^{N-1} \times \{0\}^{N-1} \subset \mathbb{R}^{\ell^*}$ with $\ell^* := \ell + n(N-1) + (N-1)$. One can easily prove that, since g is submersive everywhere and each function F_k has no zero gradient, then g^* is submersive everywhere.

Step 2: $(y^*, v^*, \mathbb{T}^*)$ is a L^1–local solution to Problem (CP'). Let us prove that there exists $\eta > 0$ such that $\phi^*(y^*(0), y^*(1)) \leq \phi^*(y(0), y(1))$ for all triplets $(y, v, \mathbb{T})$ admissible for Problem (CP') satisfying

$$\|y - y^*\|_{\mathrm{C}} + \|v - v^*\|_{\mathrm{L}^1} + \|\mathbb{T} - \mathbb{T}^*\|_{\mathbb{R}^{N-1}} \leq \eta. \tag{11.3}$$

To this aim, let $\eta > 0$ (to be reduced later) and let $(y, v, \mathbb{T})$ be an admissible triplet for Problem (CP') satisfying (11.3). In the sequel one should note that each reduction of $\eta > 0$ will be made independently of the triplet $(y, v, \mathbb{T})$ (and only in function of the triplet $(y^*, v^*, \mathbb{T}^*)$).

(i) First, since $\mathbb{T}^* \in \mathrm{Int}(\Delta)$, we can reduce $\eta > 0$ to get that $\mathbb{T} \in \mathrm{Int}(\Delta)$, and thus $0 =: \tau_0 < \tau_1 < \dots < \tau_{N-1} < \tau_N := T$. We introduce

$$x(t) := y_k\left(\frac{t - \tau_{k-1}}{\tau_k - \tau_{k-1}}\right) \quad \text{and} \quad u(t) := v_k\left(\frac{t - \tau_{k-1}}{\tau_k - \tau_{k-1}}\right),$$

for all $t \in [\tau_{k-1}, \tau_k]$ and all $k \in \{1, \dots, N\}$. Since $(y, v, \mathbb{T})$ is admissible for Problem (CP'), one can easily verify that $(x, u, \mathbb{T})$ is admissible for Problem (THP).

(ii) Now our objective is to reduce $\eta > 0$ to guarantee from (11.3) that $(x, u, \mathbb{T})$ satisfies (11.2). If so, we deduce that $\phi(x^*(0), x^*(T)) \leq \phi(x(0), x(T))$ and thus $\phi^*(y^*(0), y^*(1)) \leq \phi^*(y(0), y(1))$, which concludes Step 2. Due to the presence of two different partitions ($\mathbb{T}^*$ and $\mathbb{T}$), we underline that this step is not as trivial as it looks like. First, consider a Lipschitz continuous function $\varphi^* : [0, T] \to \mathbb{R}^m$ such that $\|u^* - \varphi^*\|_{\mathrm{L}^1}$ is small enough (chosen later), and denote by $M_1^* > 0$ and $M_2^* > 0$ the Lipschitz constants of x^* and φ^*, respectively.

– Here we want to prove that $\eta > 0$ can be reduced to get $\|x - x^*\|_{\mathrm{C}}$ as small as desired. Take $t \in [\tau_{k-1}, \tau_k]$ for some $k \in \{1, \dots, N\}$. Then

$$\begin{aligned}
\|x(t) - x^*(t)\|_{\mathbb{R}^n} &\leq \left\| y_k\left(\frac{t-\tau_{k-1}}{\tau_k - \tau_{k-1}}\right) - y_k^*\left(\frac{t-\tau_{k-1}}{\tau_k - \tau_{k-1}}\right)\right\|_{\mathbb{R}^n} + \left\| y_k^*\left(\frac{t-\tau_{k-1}}{\tau_k - \tau_{k-1}}\right) - x^*(t)\right\|_{\mathbb{R}^n} \\
&\leq \eta + \left\| x^*\left(\tau_{k-1}^* + (\tau_k^* - \tau_{k-1}^*)\frac{t-\tau_{k-1}}{\tau_k - \tau_{k-1}}\right) - x^*(t)\right\|_{\mathbb{R}^n} \\
&\leq \eta + M_1^* \left| \tau_{k-1}^* + (\tau_k^* - \tau_{k-1}^*)\frac{t-\tau_{k-1}}{\tau_k - \tau_{k-1}} - t\right| \\
&\leq \eta + M_1^* \left(T \left| \frac{\tau_k^* - \tau_{k-1}^*}{\tau_k - \tau_{k-1}} - 1\right| + \left| \tau_{k-1}^* - \tau_{k-1}\frac{\tau_k^* - \tau_{k-1}^*}{\tau_k - \tau_{k-1}}\right|\right),
\end{aligned}$$

and $\eta > 0$ can be reduced to reduce $\|\mathbb{T} - \mathbb{T}^*\|_{\mathbb{R}^{N-1}}$ sufficiently to get the above term as small as desired.

– Now we want to prove that $\eta > 0$ can be reduced to get $\|u - u^*\|_{\mathrm{L}^1}$ as small as desired. Take $k \in \{1, \dots, N\}$. Then

$$\begin{aligned}
&\int_{\tau_{k-1}}^{\tau_k} \|u(s) - u^*(s)\|_{\mathbb{R}^m}\, \mathrm{d}s \\
&\leq \int_{\tau_{k-1}}^{\tau_k} \left\| v_k\left(\frac{s-\tau_{k-1}}{\tau_k - \tau_{k-1}}\right) - v_k^*\left(\frac{s-\tau_{k-1}}{\tau_k - \tau_{k-1}}\right)\right\|_{\mathbb{R}^m} + \left\| v_k^*\left(\frac{s-\tau_{k-1}}{\tau_k - \tau_{k-1}}\right) - u^*(s)\right\|_{\mathbb{R}^m} \mathrm{d}s \\
&\leq \eta + \int_{\tau_{k-1}}^{\tau_k} \left\| u^*\left(\tau_{k-1}^* + (\tau_k^* - \tau_{k-1}^*)\frac{s-\tau_{k-1}}{\tau_k - \tau_{k-1}}\right) - u^*(s)\right\|_{\mathbb{R}^m} \mathrm{d}s \\
&\leq \eta + \int_{\tau_{k-1}}^{\tau_k} \left\| u^*\left(\tau_{k-1}^* + (\tau_k^* - \tau_{k-1}^*)\frac{s-\tau_{k-1}}{\tau_k - \tau_{k-1}}\right) - \varphi^*\left(\tau_{k-1}^* + (\tau_k^* - \tau_{k-1}^*)\frac{s-\tau_{k-1}}{\tau_k - \tau_{k-1}}\right)\right\|_{\mathbb{R}^m} \mathrm{d}s \\
&\quad + \int_{\tau_{k-1}}^{\tau_k} \left\| \varphi^*\left(\tau_{k-1}^* + (\tau_k^* - \tau_{k-1}^*)\frac{s-\tau_{k-1}}{\tau_k - \tau_{k-1}}\right) - \varphi^*(s)\right\|_{\mathbb{R}^m} \mathrm{d}s \\
&\quad + \int_{\tau_{k-1}}^{\tau_k} \|\varphi^*(s) - u^*(s)\|_{\mathbb{R}^m}\, \mathrm{d}s \\
&\leq \eta + 2\|u^* - \varphi^*\|_{\mathrm{L}^1} + M_2^* T \left(T \left| \frac{\tau_k^* - \tau_{k-1}^*}{\tau_k - \tau_{k-1}} - 1\right| + \left| \tau_{k-1}^* - \tau_{k-1}\frac{\tau_k^* - \tau_{k-1}^*}{\tau_k - \tau_{k-1}}\right|\right),
\end{aligned}$$

and $\|u^* - \varphi^*\|_{\mathrm{L}^1}$ can be reduced sufficiently (which fixes M_2^*) and then $\eta > 0$ can be reduced to reduce $\|\mathbb{T} - \mathbb{T}^*\|_{\mathbb{R}^{N-1}}$ sufficiently to get the above term as small as desired.

Step 3: application of Proposition 1. The Hamiltonian $\mathscr{H} : \mathbb{R}^{nN} \times \mathbb{R}^{mN} \times \mathbb{R}^{N-1} \times \mathbb{R}^{nN} \to \mathbb{R}$ associated with Problem (CP') is given by

$$\mathscr{H}(y, v, \mathbb{T}, q) := \langle q, f^*(y, v, \mathbb{T})\rangle_{\mathbb{R}^{nN}} = \sum_{k=1}^{N} (\tau_k - \tau_{k-1}) \langle q_k, f_k(y_k, v_k)\rangle_{\mathbb{R}^n},$$

for all $(y, v, \mathbb{T}, q) \in \mathbb{R}^{nN} \times \mathbb{R}^{mN} \times \mathbb{R}^{N-1} \times \mathbb{R}^{nN}$. From Proposition 1 applied to the triplet $(y^*, v^*, \mathbb{T}^*)$, there exists a nontrivial pair $(q, q^0) \in \mathrm{AC}([0,1], \mathbb{R}^{nN}) \times \mathbb{R}_+$ such that:

(i) it holds that

$$\dot{y^*}(s) = \nabla_q \mathscr{H}(y^*(s), v^*(s), \mathbb{T}^*, q(s)),$$
$$-\dot{q}(s) = \nabla_y \mathscr{H}(y^*(s), v^*(s), \mathbb{T}^*, q(s)),$$

for almost every $s \in [0,1]$;

(ii) it holds that

$$\begin{pmatrix} q(0) \\ -q(1) \end{pmatrix} = q^0 \nabla \phi^*(y^*(0), y^*(1)) + \nabla g^*(y^*(0), y^*(1))\tilde{\xi},$$

for some $\tilde{\xi} \in \mathrm{N}_{\mathrm{S}^*}[g^*(y^*(0), y^*(1))]$;

(iii) it holds that

$$v^*(s) \in \arg\max_{\omega \in \mathrm{U}^N} \mathscr{H}(y^*(s), \omega, \mathbb{T}^*, q(s)),$$

for almost every $s \in [0,1]$;

(iv) it holds that

$$\int_0^1 \nabla_{\mathbb{T}} \mathscr{H}(y^*(s), v^*(s), \mathbb{T}^*, q(s)) \, \mathrm{d}s \in \mathrm{N}_{\Delta}[\mathbb{T}^*],$$

with $\mathrm{N}_{\Delta}[\mathbb{T}^*] = \{0_{\mathbb{R}^{N-1}}\}$ since $\mathbb{T}^* \in \mathrm{Int}(\Delta)$;

(v) it holds that

$$\mathscr{H}(y^*(s), v^*(s), \mathbb{T}^*, q(s)) = c,$$

for almost every $s \in [0,1]$, for some $c \in \mathbb{R}$.

Step 4: inverting the augmentation procedure. Let us define $p^0 := q^0 \in \mathbb{R}_+$ and $p \in \mathrm{PAC}_{\mathbb{T}^*}([0,T], \mathbb{R}^n)$ by $p(0) := q_1(0)$, $p(T) := q_N(1)$ and by

$$p(t) := q_k\left(\frac{t - \tau_{k-1}^*}{\tau_k^* - \tau_{k-1}^*}\right) \text{ for all } t \in (\tau_{k-1}^*, \tau_k^*) \text{ and all } k \in \{1, \dots, N\}.$$

Note that the nontriviality of the pair (q, q^0) implies the nontriviality of the pair (p, p^0). We now prove the items of Proposition 2 one by one.

• *Hamiltonian system of Proposition 2.* It follows directly from the above Item (i).

• *Endpoint transversality condition of Proposition 2.* It follows from the above Item (ii) and from the definitions of ϕ^*, g^* and S^*. First we can write $\tilde{\xi} := (\xi, \xi^2, \xi^3) \in \mathbb{R}^\ell \times \mathbb{R}^{n(N-1)} \times \mathbb{R}^{N-1}$ with $\xi \in \mathrm{N}_{\mathrm{S}}[g(y_1^*(0), y_N^*(1))]$. Second, since $(y_1^*(0), y_N^*(1)) = (x^*(0), x^*(T))$, we get that

$$p(0) = q_1(0) = q^0 \nabla_1 \phi(y_1^*(0), y_N^*(1)) + \nabla_1 g(y_1^*(0), y_N^*(1))\xi,$$
$$= p^0 \nabla_1 \phi(x^*(0), x^*(T)) + \nabla_1 g(x^*(0), x^*(T))\xi.$$

Additionally we obtain

$$-p(T) = -q_N(1) = q^0 \nabla_2 \phi(y_1^*(0), y_N^*(1)) + \nabla_2 g(y_1^*(0), y_N^*(1))\xi,$$

$$= p^0 \nabla_2 \phi(x^*(0), x^*(T)) + \nabla_2 g(x^*(0), x^*(T))\xi,$$

which proves the endpoint transversality condition of Proposition 2.

• *Discontinuity condition of Proposition 2.* It follows from the above Item (ii) and the definition of g^*. Precisely it holds that

$$\forall k \in \{2, \dots, N-1\},\ q_k(0) = \xi^2_{k-1}, \qquad \forall k \in \{1, \dots, N-1\},\ -q_k(1) = -\xi^2_k + \xi^3_k \nabla F_k(y^*_k(1)).$$

We deduce that

$$p^+(\tau^*_k) - p^-(\tau^*_k) = q_{k+1}(0) - q_k(1) = \xi^3_k \nabla F_k(x^*(\tau^*_k)),$$

for all $k \in \{1, \dots, N-1\}$. By taking $\sigma_k := \xi^3_k$ for all $k \in \{1, \dots, N-1\}$, the discontinuity condition of Proposition 2 is proved.

• *Hamiltonian maximization condition of Proposition 2.* It follows straightforwardly from the above Item (iii).

• *Hamiltonian constancy condition of Proposition 2.* From the Hamiltonian system and the Hamiltonian maximization condition of Proposition 2, and applying [18, Theorem 2.6.1] on each interval (τ^*_{k-1}, τ^*_k) (on which the Hamiltonian H_1 is autonomous), we obtain that

$$\forall k \in \{1, \dots, N\}, \quad \exists c_k \in \mathbb{R}, \quad \langle p(t), f_k(x^*(t), u^*(t)) \rangle_{\mathbb{R}^n} = c_k,$$

for almost every $t \in (\tau^*_{k-1}, \tau^*_k)$. On the other hand, the above Item (iv) implies that

$$\int_0^1 \langle q_{k+1}(s), f_{k+1}(y^*_{k+1}(s), v^*_{k+1}(s)) \rangle_{\mathbb{R}^n} \, \mathrm{d}s = \int_0^1 \langle q_k(s), f_k(y^*_k(s), v^*_k(s)) \rangle_{\mathbb{R}^n} \, \mathrm{d}s,$$

for all $k \in \{1, \dots, N-1\}$. Now, inverting the change of time variable, we obtain the equality

$$\frac{1}{\tau^*_{k+1} - \tau^*_k} \int_{\tau^*_k}^{\tau^*_{k+1}} \langle p(t), f_{k+1}(x^*(t), u^*(t)) \rangle_{\mathbb{R}^n} \, \mathrm{d}t = \frac{1}{\tau^*_k - \tau^*_{k-1}} \int_{\tau^*_{k-1}}^{\tau^*_k} \langle p(t), f_k(x^*(t), u^*(t)) \rangle_{\mathbb{R}^n} \, \mathrm{d}t,$$

for all $k \in \{1, \dots, N-1\}$. We deduce that $c_{k+1} = c_k$ for all $k \in \{1, \dots, N-1\}$. This proves the Hamiltonian constancy condition of Proposition 2.

References

1. L. T. Ashchepkov, D. V. Dolgy, T. Kim and R. P. Agarwal, General optimal control problem, in *Optimal Control*, Springer, 2022, 141–164.
2. N. Augier and A. G. Yabo, Time-optimal control of piecewise affine bistable gene-regulatory networks, *International Journal of Robust and Nonlinear Control*, **33** (2023), 4967–4988.
3. G. Barles, A. Briani and E. Trélat, Value function for regional control problems via dynamic programming and Pontryagin maximum principle, *Math. Control Relat. Fields*, **8** (2018), 509–533.
4. T. Bayen, A. Bouali and L. Bourdin, Hybrid maximum principle with regionally switching parameter, *(in revision), hal-03638701.*

5. T. Bayen, A. Bouali and L. Bourdin, The hybrid maximum principle for optimal control problems with spatially heterogeneous dynamics is a consequence of a pontryagin maximum principle for $L^1_\square$–local solutions, *(in revision), hal-03985420*.
6. T. Bayen, A. Bouali and L. Bourdin, Minimum time problem for the double integrator with a loss control region, *submitted, hal-03928967v2*.
7. T. Bayen, K. Boumaza and A. Rapaport, Necessary optimality condition for the minimal time crisis relaxing transverse condition via regularization, *ESAIM Control Optim. Calc. Var.*, **27** (2021), Paper No. 105, 30pp.
8. T. Bayen and A. Rapaport, Minimal time crisis versus minimum time to reach a viability kernel: a case study in the prey-predator model, *Optim. Control Appl.*, **40** (2019), 330–350.
9. A. Blumentals, B. Brogliato and F. Bertails-Descoubes, The contact problem in lagrangian systems subject to bilateral and unilateral constraints, with or without sliding coulomb's friction: a tutorial, *Multibody System Dynamics*, **38** (2016), 43–76.
10. U. Boscain and B. Piccoli, *Optimal syntheses for control systems on 2-D manifolds*, vol. 43, Springer Science & Business Media, 2003.
11. B. Brogliato and L. Thibault, Existence and uniqueness of solutions for non-autonomous complementarity dynamical systems, *Journal of Convex Analysis*, **17** (2010), 961–990.
12. L. Cesari, *Lagrange and Bolza Problems of optimal control and other problems*, 196–205, Springer, New York, 1983.
13. F. H. Clarke and R. B. Vinter, Applications of optimal multiprocesses, *SIAM Journal on Control and Optimization*, **27** (1989), 1048–1071.
14. F. H. Clarke and R. B. Vinter, Optimal multiprocesses, *SIAM Journal on Control and Optimization*, **27** (1989), 1072–1091.
15. A. V. Dmitruk and A. Kaganovich, Maximum principle for optimal control problems with intermediate constraints, *Computational Mathematics and Modeling*, **22** (2011), 180–215.
16. A. Dmitruk, Maximum principle for the general optimal control problem with phase and regular mixed constraints, *Computational Mathematics and Modeling*, **4** (1993), 364–377.
17. A. Dmitruk and A. Kaganovich, The hybrid maximum principle is a consequence of Pontryagin maximum principle, *Systems Control Lett.*, **57** (2008), 964–970.
18. H. O. Fattorini, *Infinite dimensional optimization and control theory*, Cambridge University Press, 1999.
19. A. F. Filippov, Differential equations with discontinuous right-hand side, *Matematicheskii sbornik*, **93** (1960), 99–128.
20. M. Garavello and B. Piccoli, Hybrid necessary principle, *SIAM J. Control Optim.*, **43** (2005), 1867–1887.
21. T. Haberkorn and E. Trélat, Convergence results for smooth regularizations of hybrid nonlinear optimal control problems, *SIAM J. Control Optim.*, **49** (2011), 1498–1522.
22. A. Pakniyat and P. E. Caines, On the hybrid minimum principle: the Hamiltonian and adjoint boundary conditions, *IEEE Trans. Automat. Contr.*, **66** (2020), 1246–1253.
23. A. Pakniyat and P. E. Caines, The hybrid minimum principle in the presence of switching costs, in *52nd IEEE Conference on Decision and Control*, IEEE, 2013, 3831–3836.
24. L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze and E. F. Mishchenko, *The mathematical theory of optimal processes*, Translated by D. E. Brown, A Pergamon Press Book. The Macmillan Co., New York, 1964.
25. H. J. Sussmann, A nonsmooth hybrid maximum principle, in *Stability and stabilization of nonlinear systems*, vol. 246, Springer London, 1999, 325–354.
26. H. J. Sussmann, A maximum principle for hybrid optimal control problems, in *Proc. 38th IEEE Conf. Decis. Control*, vol. 1, Ieee, 1999, 425–430.

12

About Optimal Control Problem Under Action Duration Constraint and Infimum-gap

Dan Goreac[1,2] and Alain Rapaport[3]

[1] School of Mathematics and Statistics, Shandong University, Weihai, China
[2] LAMA, Univ. Eiffel, UPEM, Univ. Paris Est Creteil, CNRS, Marne-la-Vallée, France
dan.goreac@u-pem.fr
[3] MISTEA, Univ. Montpellier, INRAE, Institut Agro, Montpellier, France alain.rapaport@inrae.fr

How to pay a right tribute to a mathematical genius of such vast curiosity and knowledge? Ivan Kupka contributed brilliantly to the theory of optimal control, notably on the geometric theory of extremals and generic properties of singular trajectories. We modestly dedicate this note on some optimal control problems, which we believe he would have been interested... to his memory.

Summary. We consider optimal control problems with scalar control in $[0, 1]$ under the constraint that the length of time during which the control is non-null is bounded by a prescribed value. We show that such problems present an infimum-gap when the solution of the relaxed problem is not bang-bang. Then, we show that the gap can be closed if one considers the infimum over all survival probability functions dominated by the survival probability function of the uniform law as a constraint on the control function.

12.1 Introduction

The Covid-19 pandemic has led to a surge in activity among researchers in applied mathematics about epidemiological modeling, and in particular for optimization and decision support issues, which are found to be of general interest in epidemiology beyond the Sars diseases.

Thus, the problems of minimizing the epidemic peak or maximizing the final size of the susceptible sub-population under duration constraints on the interventions have recently been tackled in the literature [13, 12, 2, 3] for the well-known SIR model [9]. Optimal solutions have been mathematically demonstrated for this model. The contributions of these theoretical analyses versus purely numerical solutions are to provide explicit structures of the optimal solution in terms of feedback strategies. Let us stress that these criteria are not standard, i.e. not in the usual Mayer, Lagrange or Bolza form, or over infinite horizon with moreover unconventional constraints on the control variable. This deserves special interests because one cannot apply straightforwardly usual tools such as the Maximum Principle of Pontryagin to prove the structure of the optimal solution. In the three contributions [13, 12, 2], the controlled SIR model that is considered is as follows

$$\dot{S} = -\beta(1-u)SI$$
$$\dot{I} = \beta(1-u)SI - \gamma I$$

$$\dot{R} = \gamma I$$

where the control $u \in [0, u_{max}]$ with $u_{max} \leq 1$ represents actions or interventions (such as lock-downs and curfew) which reduce contacts between infected and susceptible individuals. In [12] the problem consists in minimizing the L^∞ norm of the variable I (the so-called "epidemic peak")

$$\inf_{u(\cdot)} \max_{t \geq 0} I(t), \tag{12.1}$$

under a L_1 budget constraint on the action

$$\int_0^{+\infty} u(t)dt \leq Q. \tag{12.2}$$

Alternatively, the authors in [11, 1, 5] have considered the "dual" problem which consists in minimizing the L^1 norm of the control

$$\inf_{u(\cdot)} \int_0^{+\infty} u(t)dt,$$

under the state constraints

$$I(t) \leq \bar{I}, \quad t \in [0, +\infty).$$

The optimal strategies of these two problems turn out to be identical (bang-singular-bang), as this has indeed been proved to be true in a more general framework [7]. In [13], the same criterion (12.1) has been considered but for the class of controls $u(\cdot)$ that verify a duration constraint

$$u(t) = 0,\ t \notin [t_i, t_i + \tau], \quad u(t) > 0,\ t \in [t_i, t_i + \tau], \tag{12.3}$$

where τ is fixed and t_i has to be chosen. The structure of the optimal solution (band-singular-bang-bang) is different.

In [2, 3] , the problem of maximizing the final size

$$\sup_{u(\cdot)} \lim_{t \to +\infty} S(t) \tag{12.4}$$

has been investigated for the same class of controls (12.3), and the optimal solutions has been proved to be bang-bang i.e. $u(t) = u_{max}$, for $t \in [t_i, t_i + \tau]$, where t_i has to be optimized. In [4] the authors have considered the same criterion (12.4) but for the class of controls in L_1 with a budget constraint (12.2), and have shown that the same bang-bang control on a single time interval of interventions was optimal, when $u_{max}\tau = Q$. Here, this means that the optimal solution for the L_1 constraint is not only optimal for the class of controls (12.3), but also for the more general class of controls such that

$$\int_0^{+\infty} I_{\mathbb{R}_\star^+}(u(t)dt \leq \tau, \tag{12.5}$$

where $I_{\mathbb{R}_\star^+}$ denotes the indicator function of positive numbers, that is measurable controls for which the occupation measure of $\mathbb{R}_\star^+$ is bounded by τ.

More generally, when the optimal solution under a L^1 constraint of the control is not bang-bang, as for problem (12.1) for instance, the problems with duration constraints are no longer equivalent. Motivated by these observations, the objective of the present note is to investigate problems for scalar controls under the "action duration constraint" (12.5). The use of occupation

measures in control theory has already been considered to deal with state or mixed constraints to reformulate nonlinear optimal control problems as infinite dimensional linear programming problems on spaces of occupational measures generated by control-state trajectories [10, 6]. Here, we consider the occupation measure to define the constraint on the control function, which has not been yet considered in this way up to our knowledge.

The following is organized as follows. In the next section, we propose an equivalent reformulation of optimal control problems with action duration constraint ad show that a infimum-gap occurs when bang-bang controls are not optimal. Then, in Section we generalize the constraint (12.5) for different measures on the control and show under which conditions an infimum-gap is avoided.

12.2 Reformulation with Extended Velocity Set and Relaxation

We consider a control system in $\mathbb{R}^n$

$$\begin{cases} \dot{x} = f(x) + g(x)u, \\ x(0) = x_0 \end{cases} \qquad u \in U := [0,1], \tag{12.6}$$

where f, g are C^1 maps with linear growth, and the set of control functions

$$\mathscr{U} := \big\{u : [0,T] \to U \text{ Borel mesurable}\big\}.$$

Define the Mayer problem for $\Phi \in C^1$

$$(\mathscr{P}_1) : \quad \inf_{u(\cdot)\in\mathscr{U}_\tau} \Phi(x(T)),$$

under a constraint on the action duration, that is for

$$\mathscr{U}_\tau := \big\{u(\cdot) \in \mathscr{U} : \text{ meas}\, E(u) \le \tau\big\} \text{ where } E(u) := \{t \in [0,T];\ u(t) > 0\}.$$

As recalled in the introduction, such a constraint is not classical in optimal control theory.

Alternatively, we consider the extended dynamics

$$\begin{cases} \dot{x} = f(x) + g(x)u,\ x(0) = x_0, \\ \dot{z} = v,\ z(0) = 0, \end{cases} \tag{12.7}$$

where

$$w := (u,v) \in W := \{(u,v) \in [0,1]^2,\ uv = 0\},$$

and the (more classical) optimal control problem with a target condition

$$(\mathscr{P}_2) : \quad \inf_{w(\cdot)\in\mathscr{W}} \Phi(x(T)) \text{ s.t. } z(T) \ge T - \tau,$$

where $\mathscr{W} := \big\{w : [0,T] \to W \text{ Borel mesurable}\big\}$.

Lemma 1. *Problems* $(\mathscr{P}_1)$ *and* $(\mathscr{P}_2)$ *are equivalent.*

Proof. Take $\epsilon > 0$ and let $u_\epsilon(\cdot) \in \mathscr{U}_T$ be such that the corresponding solution $x_\epsilon(\cdot)$ satisfies

$$\Phi(x_\epsilon(T)) < \inf_{u(\cdot)\in\mathscr{U}_\tau} \Phi(x(T)) + \epsilon.$$

Let

$$v_\epsilon(t) = \begin{cases} 0 & t \in E(u_\epsilon), \\ 1 & t \notin E(u_\epsilon). \end{cases}$$

Clearly (u_ϵ, v_ϵ) belongs to $\mathscr{W}$ and the corresponding solution $z_\epsilon(\cdot)$ verifies

$$z_\epsilon(T) = \int_0^T v_\epsilon(t)dt = T - E(u_\epsilon) \geq T - \tau.$$

Then one gets

$$\Phi(x_\epsilon(T)) \geq \inf_{w(\cdot)\in\mathscr{W}} \Phi(x(T)) \text{ s.t. } z(T) \geq T - \tau,$$

and, thus,

$$\inf_{u(\cdot)\in\mathscr{U}_\tau} \Phi(x(T)) > \left(\inf_{w(\cdot)\in\mathscr{W}} \Phi(x(T)) \text{ s.t. } z(T) \geq T - \tau \right) - \epsilon.$$

As $\epsilon > 0$ is arbitrary, we obtain

$$\inf_{u(\cdot)\in\mathscr{U}_\tau} \Phi(x(T)) \geq \inf_{w(\cdot)\in\mathscr{W}} \Phi(x(T)) \text{ s.t. } z(T) \geq T - \tau. \tag{12.8}$$

Conversely, let $(u_\epsilon, v_\epsilon) \in \mathscr{W}$ such that the corresponding solution $(x_\epsilon(\cdot), z_\epsilon(\cdot)$ verifies $z_\epsilon(T) \geq T - \tau$ and

$$\Phi(x_\epsilon(T)) < \left(\inf_{w(\cdot)\in\mathscr{W}} \Phi(x(T)) \text{ s.t. } z(T) \geq T - \tau \right) + \epsilon.$$

One has

$$T - \tau \leq z_\epsilon(T) = \int_0^T v_\epsilon(t)dt = \int_{t\notin E(u_\epsilon)} v_\epsilon(t)dt \leq T - \text{meas}\, E(u_\epsilon),$$

and, thus, $u_\epsilon(\cdot) \in \mathscr{U}_\tau$. This allows to write

$$\Phi(x_\epsilon(T)) \geq \inf_{u(\cdot)\in\mathscr{U}_\tau} \Phi(x(T)),$$

from which we get

$$\left(\inf_{w(\cdot)\in\mathscr{W}} \Phi(x(T)) \text{ s.t. } z(T) \geq T - \tau \right) > \inf_{u(\cdot)\in\mathscr{U}_\tau} \Phi(x(T)) - \epsilon.$$

Letting ϵ be arbitrary small, one obtains

$$\left(\inf_{w(\cdot)\in\mathscr{W}} \Phi(x(T)) \text{ s.t. } z(T) \geq T - \tau \right) \geq \inf_{u(\cdot)\in\mathscr{U}_\tau} \Phi(x(T)). \tag{12.9}$$

Inequalities (12.8) and (12.9) give the equivalence of problems $(\mathscr{P}_1)$ and $(\mathscr{P}_2)$.

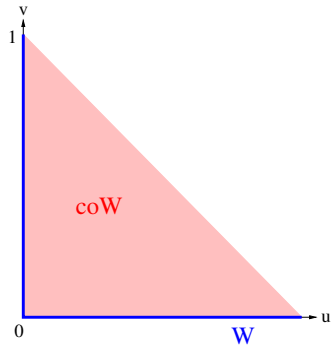


Fig. 12.1: The set W and its convexification $\operatorname{co} W$

Note that the control set W is not convex and thus existence of optimal solution is not guaranteed. However, one has

$$\overline{\operatorname{co}}\, W = \{(u,v) \in [0,1]^2,\ u+v \leq 1\},$$

(where $\overline{co}$ denotes the closed convex hull, see Figure 12.1) and one can consider the "convexified" or relaxed problem

$$(\overline{\mathscr{P}_2}): \quad \inf_{w(\cdot)\in\overline{\mathscr{W}}} \Phi(x(T)) \text{ s.t. } z(T) \geq T-\tau,$$

where $\overline{\mathscr{W}} := \big\{w : [0,T] \to \overline{\operatorname{co}}\, W \text{ mesurable}\big\}$. Let us also consider the problem with L^1 constraint on the control, that is

$$(\mathscr{P}_3): \quad \inf_{u(\cdot)\in\mathscr{U}_\tau^1} \Phi(x(T)),$$

where

$$\mathscr{U}_\tau^1 := \big\{u(\cdot) \in \mathscr{U} \text{ s.t. } ||u||_1 \leq \tau\big\}, \text{ where } ||u||_1 := \int_0^T u(t)dt.$$

Note that problems $(\overline{\mathscr{P}_2})$ and $(\mathscr{P}_3)$ fulfill the usual convexity assumption that guarantees the existence of optimal solutions.

Lemma 2. *Problems* $(\overline{\mathscr{P}_2})$ *and* $(\mathscr{P}_3)$ *are equivalent.*

Proof. Take $(u(\cdot), v(\cdot)$ in $\overline{\mathscr{W}}$ such that the corresponding solution satisfies $z(T) \geq T-\tau$. Then, one can write

$$\int_0^T u(t)dt \leq \int_0^T 1-v(t)dt = T - z(T) \leq \tau,$$

that is $u(\cdot)$ belongs to $\mathscr{U}_\tau^1$ and one has

$$\inf_{u(\cdot)\in\mathscr{U}_\tau^1} \Phi(x(T)) \leq \inf_{w(\cdot)\in\overline{\mathscr{W}}} \Phi(x(T)) \text{ s.t. } z(T) \geq T-\tau. \tag{12.10}$$

Conversely, let $u(\cdot)$ belong to $\mathscr{U}_\tau^1$ and posit $v(t) = 1-u(t)$ for any $t \in [0,T]$. Clearly, $(u(\cdot), v(\cdot)$ belongs to $\overline{\mathscr{W}}$ and one has

$$z(T) = \int_0^T v(t)dt = \int_0^T 1-u(t)dt = T - \int_0^T u(t)dt \geq T-\tau.$$

Therefore, one has

$$\inf_{u(\cdot)\in\mathscr{U}_\tau^1}\Phi(x(T))\geq\inf_{w(\cdot)\in\overline{\mathscr{W}}}\Phi(x(T))\text{ s.t. }z(T)\geq T-\tau. \tag{12.11}$$

Finally, inequalities (12.10), (12.11) show the equivalence of problems $(\overline{\mathscr{P}_2})$ and $(\mathscr{P}_3)$.

From Lemmas 1 and 2, we immediately deduce the following property.

Proposition 1. *Problem $(\mathscr{P}_3)$ is equivalent to problem $(\mathscr{P}_1)$ with relaxed controls.*

Remark 1. If problem $(\mathscr{P}_3)$ admits an optimal solution with a control $u(\cdot)$ that takes values 0 or 1 only, then it is optimal for problem $(\mathscr{P}_1)$, as $||u||_1=\text{meas }E(u)$ in this case.

Finally, we obtain the following result about existence of infimum-gap in the original problem.

Proposition 2. *If any optimal solution of problem $(\mathscr{P}_3)$ saturates the L^1 constraint and possesses a singular arc, then there is an infimum-gap between between problem $(\mathscr{P}_1)$ and problem $(\mathscr{P}_3)$, that is*

$$\inf_{u(\cdot)\in\mathscr{U}_\tau}\Phi(x(T))>\min_{u(\cdot)\in\mathscr{U}_\tau^1}\Phi(x(T)).$$

Proof. Assume by contradiction that

$$\inf_{u(\cdot)\in\mathscr{U}_\tau}\Phi(x(T))=J^\star:=\min_{u(\cdot)\in\mathscr{U}_\tau^1}\Phi(x(T)).$$

Then, for any $n\in\mathbb{N}$, there exists a control $u_n(\cdot)\in\mathscr{U}_\tau$ such that

$$\Phi(x_n(T))<J^\star+\frac{1}{n},$$

where $x_n(\cdot)$ denotes the solution of (12.6) associated to $u_n(\cdot)$. Thanks to the convexity of the velocity set of the dynamics (12.6) and the usual regularity assumptions, the theorem of compactness of solutions of (12.6) applies and there exists a sub-sequence, also denoted x_n, such that $x_n(\cdot)$ converges pointwise to a certain $x^\star(\cdot)$ solution of (12.6) a certain $u^\star\in\mathscr{U}$, and $\dot{x}_n(\cdot)$ converges weakly to $\dot{x}^\star(\cdot)$, that is $u_n(\cdot)$ converges weakly to $u^\star(\cdot)$. Moreover, we obtain passing at the limit $\Phi(x^\star(T))=J^\star$.

On another hand, $\mathbb{P}(u):=\frac{1}{T}\text{meas }E(u)$ is the occupation probability for the open set $(0,+\infty)$ for any measurable function $u(\cdot)$. By the Portmanteau Theorem (see for instance [8]), we get

$$\liminf_n\mathbb{P}(u_n)\geq\mathbb{P}(u^\star).$$

and as meas $E(u_n)\leq\tau$ for any n, we deduce that

$$\tau\geq\text{meas }E(u^\star)\geq||u^\star||_1.$$

Therefore $u^\star$ belongs to $\mathscr{U}_\tau^1$ with $\Phi(x^\star(T))=J^\star$. The control $u^\star$ is thus optimal for problem $(\mathscr{P}_1)$ with meas $E(u^\star)=\tau$. Then, having $||u^\star||_1=\text{meas }E(u^\star)$ implies that $u^\star(t)$ is equal to 0 or 1 for a.e. $t\in[0,T]$, which contradicts the presence of a singular arc.

12.3 Generalization of the Action Duration Constraint

The action duration constraint defined by the set $\mathscr{U}_\tau$ in Section 12.2 can be seen as a particular measure or "energy" imposed on the control functions $u(\cdot)$, where energy and time spend by an action are the same. If one draws analogy with mechanics, this means that any action count the same one unit of energy. From a Lagrangian view point, this amounts to measure the energy to go from one point to another simply by the euclidean distance. One may consider other ways to define energy of actions $u(\cdot)$ with a non euclidean geometry. For pure bang-bang controls, there is no need to distinguish how counts each possible action but this is no longer the case with singular arcs, as revealed by Proposition 2. For such arcs, the energy needs to be computed taking into account the geometry.

This intuition roughly corresponds to considering some measure on the control set $[0,1]$, or, alternatively, some affine transformation of a cumulative distribution function (c.d.f.) $u \mapsto \gamma(u)$ and compute the action energy (instead of the simple duration) $\int_0^T \gamma(u(t))dt$, or, even more generally $\int_0^T \gamma(t,u(t))dF(t)$ with F being a c.d.f. of a $[0,T]$-supported random variable.

12.3.1 A control point of view for measuring

Having a look at the constraint defining the set W in Section 12.2, i.e. $uv = 0$ and the ensuing formulation for $\overline{co}\, W$, one may consider a similar approach for any constraint $\phi(u,v) = 0$, provided that the kernel of ϕ is included in $\overline{co}\ W$ and $\phi(0,v) = \phi(u,0) = 0$ for any $(u,v) \in [0,1]^2$. In particular, constraints that are expressed as $v \in \Gamma(u)$ possess these properties, provided that the set-valued map Γ verifies

$$0 \in \Gamma(u) \subset [0, 1-u], \text{ for all } u \in [0,1] \text{ and } \Gamma(0) = [0,1]\,. \tag{12.12}$$

For ensuring the existence of optimal solutions, we shall require the set-valued map Γ to take convex compact values and to be upper hemicontinuous, which basically amounts to asking $\Gamma(u) = [0, \theta(u)]$, where θ is a $[0,1]$-valued upper semicontinuous function such that $\theta(0) = 1$ and $\theta(u) \leq 1 - u$, for every $u \in [0,1]$. Indeed, we are going to offer a slight generalization by considering a function

$$h : [0,1] \mapsto [0,1] \text{ convex continuous s.t. } h(u) = 0 \Leftrightarrow u = 0. \tag{12.13}$$

Having fixed such a function h, we then define the set of functions

$$\begin{aligned} \Theta(h) := \Big\{ & \theta : [0,1] \longrightarrow [0,1]\,;\ \theta \text{ is } u.s.c., \\ & \theta(u) \leq 1 - h(u),\ \forall u \in [0,1] \text{ and } \theta(0) = 1 \Big\}. \end{aligned} \tag{12.14}$$

Let $\theta \in \Theta(h)$ and consider the control set

$$W_\theta := \left\{ (u,v) \in [0,1]^2 :\ v \leq \theta(u) \right\}.$$

For F being the cummulative distribution function (c.d.f.) of a $[0,T]$-supported random variable, we define the following optimal control problem

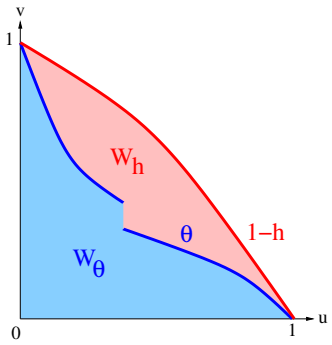


Fig. 12.2: Example of sets W_θ and $\overline{W}_h$

$$(\mathscr{P}_{1,F,\theta}) \ : \inf_{u\in\mathscr{U}_{F,\tau}(\theta)} \Phi(x(T)),$$

for the dynamics (12.6), where

$$\mathscr{U}_{F,\tau}(\theta) := \left\{ u(\cdot)\in\mathscr{U} \,:\, \mathbb{E}_F\left[1-\theta(u(\cdot))\right] := \int_0^T (1-\theta(u(t)))\, dF(t) \le \frac{\tau}{T} \right\}.$$

As in Section 12.2, we consider for the extended dynamics

$$\begin{cases} \dot{x} = f(x) + g(x)u,\ x(0) = x_0, \\ dz_F = v\ dF,\ z_F(0) = 0, \end{cases} \tag{12.15}$$

with control $w = (u, v) \in W_\theta$ the optimization problem

$$(\mathscr{P}_{2,F,\theta}) : \quad \inf_{w(\cdot)\in\mathscr{W}_\theta} \phi(x(T)) \text{ s.t. } z_F(T) \ge 1 - \frac{\tau}{T},$$

for the family of Borel measurable functions $\mathscr{W}_\theta := \mathbb{L}^0\left([0,T]\,; W_\theta\right)$.

Remark 2. The initial problems $(\mathscr{P}_1)$ and $(\mathscr{P}_2)$ in Section 12.2 are obtained for F corresponding to the uniformly distributed r.v. on $[0,T]$ (i.e. $F(u) = \frac{u}{T}$, on $[0,T]$) and the indicator function $\theta = \mathbf{1}_0$.

Lemma 3. *Problems $(\mathscr{P}_{1,F,\theta})$ and $(\mathscr{P}_{2,F,\theta})$ are equivalent.*

Proof. The proof is quite similar to the one provided in Section 12.2. For convenience, let us denote $V_{j,F,\theta}(x_0)$ the value functions for the problems $\mathscr{P}_{j,F,\theta}$, with $j \in \{1,2\}$ and $x_0 \in \mathbb{R}^n$.
For $\varepsilon > 0$ and u_ε being an ε-optimal control for the problem $(\mathscr{P}_{1,F,\theta})$, we set $v_\varepsilon := \theta\left(u_\varepsilon\right)$. It is clear that the associated solution $z_F^{v_\varepsilon}$ satisfies

$$z_F^{v_\varepsilon}(T) = \int_0^T v_\varepsilon(t) dF(t) = 1 - \int_0^T (1-\theta(u_\varepsilon(t))\, dF(t) \ge 1 - \frac{\tau}{T},$$

which, by invoking the arbitrariness of $\varepsilon > 0$, leads to

$$V_{1,F,\theta}(x_0) \ge V_{2,F,\theta}(x_0).$$

For the converse, again with $\varepsilon > 0$ and $w_\varepsilon = (u_\varepsilon, v_\varepsilon)$ being an ε-optimal control for problem $\mathscr{P}_{2,F,\theta}$, i.e., $\Phi\left(x_\varepsilon(T)\right) \leq V_{2,F,\theta} + \varepsilon$, one has

$$1 - \frac{\tau}{T} \leq z_F^{v_\varepsilon}(T) = \int_0^T v_\varepsilon(t) dF(t) \leq \int_0^T \theta(u_\varepsilon(t)) dF(t),$$

leading to u_ε being admissible for problem $\mathscr{P}_{1,F,\theta}$, i.e., $u_\varepsilon \in \mathscr{U}_{F,\tau}(\theta)$.As a consequence, one has

$$V_{2,F,\theta}(x_0) + \varepsilon \geq \Phi\left(x_\varepsilon(T)\right) \geq V_{1,F,\theta}(x_0).$$

Again, we invoke the arbitrariness of $\varepsilon > 0$ to complete the proof of our assertions.

A simple glance at the argument developed at the beginning shows that, having fixed h and $\theta \in \Theta(h)$, one has

$$\overline{co}\, W_\theta \subset \overline{W}_h := \left\{(u, v) \in [0,1]^2 \ : \ h(u) + v \leq 1\right\} \tag{12.16}$$

(see Figure 12.2 as an illustration). Note that we have equality in the particular case of the identity function $h = Id$.

We now consider the relaxed control problems that are defined independently of $\theta \in \Theta(h)$ (but may still depend on h)

$$\left(\overline{\mathscr{P}}_{2,F,h}\right): \quad \inf_{w(\cdot) \in \overline{\mathscr{W}}_h} \phi(x(T)) \text{ s.t. } z_F(T) \geq 1 - \frac{\tau}{T},$$

where $\overline{\mathscr{W}}_h := \mathbb{L}^0\left([0,T]; \overline{W}_h\right)$, and

$$\left(\mathscr{P}_{3,F,h}\right): \quad \inf_{u \in \mathscr{U}^1_{F,\tau}(h)} \Phi(x(T)),$$

where

$$\mathscr{U}^1_{F,\tau}(h) := \left\{u(\cdot) \in \mathscr{U} \text{ s.t. } \|h(u(\cdot))\|_{\mathbb{L}^1([0,T], dF; \mathbb{R}_+)} := \int_0^T h(u(t)) dF(t) \leq \frac{\tau}{T}\right\}.$$

With an argument identical (up to replacing in the inequalities u by $h(u)$ and dt by dF and recalling that the total mass of dF is 1) to the one employed in Lemma 2, one establishes the following equivalence result.

Lemma 4. *The problems $\left(\mathscr{P}_{3,F,h}\right)$ and $\left(\overline{\mathscr{P}}_{2,F,h}\right)$ are equivalent.*

We emphasize that this equivalence is provided for every cummulative distribution function F associated to some $[0,T]$-supported real-valued random variables.

12.3.2 A non infimum-gap result

We are now ready to state and prove the following result stating a no-gap property between problems $\left(\mathscr{P}_{3,F,h}\right)$ and a minimization of $\left(\mathscr{P}_{1,F,\theta}\right)$ over $\theta \in \Theta(h)$.

Proposition 3. *Let h be a function satisfying the property* (12.13) *and F be a cumulative distribution function (c.d.f.) of a real-valued random variable taking its values in $[0,T]$. Then, the following equality holds true.*

$$\inf_{\theta \in \Theta(h)} \inf_{u(\cdot) \in \mathscr{U}_{F,\tau}(\theta)} \Phi(x(T)) = \inf_{u(\cdot) \in \mathscr{U}^1_{F,\tau}(h)} \Phi(x(T)).$$

Proof. By Lemmas 3 and 4, it follows that, for every $\theta \in \Theta$, the optimal value of problem $(\mathscr{P}_{1,F,\theta})$ is no lower than the one of problem $(\mathscr{P}_{3,F,h})$. By taking the infimum over $\theta \in \Theta(h)$, it follows

$$\inf_{\theta\in\Theta(h)} \inf_{u(\cdot)\in\mathscr{U}_{F,\tau}(\theta)} \Phi(x(T)) \geq \min_{u(\cdot)\in\mathscr{U}^1_{F,\tau}(h)} \Phi(x(T)).$$

Let $u_\varepsilon \in \mathscr{U}^1_{F,\tau}(h)$ be an optimal control for $(\mathscr{P}_{3,F,h})$. By definition, one has

$$\int_0^T h\left(u_\varepsilon(t)\right) dF(t) \leq \frac{\tau}{T}.$$

Then, by simply taking $\theta^*(u) := 1 - h(u)$, we complete the proof since $\theta^* \in \Theta(h)$ and

$$\mathbb{E}_F\left[1 - \theta^*(u_\varepsilon(\cdot))\right] = \int_0^T h\left(u_\varepsilon(t)\right) dF(t),$$

thus showing that $u_\varepsilon \in \mathscr{U}_{F,\tau}(\theta^*)$.

Remark 3. For $F(u) = \min\Big(\max\big(0, \frac{u}{T}\big), 1\Big)$ (i.e., F corresponding to the uniform law on $[0,T]$), and $h(u) = u$), let us we drop for the dependence on F and h and simply write $(\mathscr{P}_{1,\theta})$, $(\mathscr{P}_{2,\theta})$, $\big(\overline{\mathscr{P}}_2\big)$, $(\mathscr{P}_3)$ for simplicity. The result states that the control problem $(\mathscr{P}_3)$ of minimizing the final cost under the integral constraint on the control is equivalent to minimizing $(\mathscr{P}_{1,\theta})$ over all survival probability functions θ of r.v. on $[0,1]$ and dominated by the survival probability function of the uniform law. This closes the gap and refers to the mechanical heuristic depicted at the beginning of this section: it is not only the action/non action that need to be taken into account (as for the action duration constraint of Section 12.2) but one has to minimize over all "energy" curves with $\theta(0) = 1$ and $\theta(1) = 0$.

12.4 Conclusion

In the present work, we show the benefit of extending the dynamics with an additional control to reformulate the optimal control problem with action duration constraint. This allows us to show that the relaxed problem is indeed the problem with the simple constraint on the L_1 norm on the control. An infimum-gap appears then when the optimal control under the L_1 constraint presents a singular arc. We generalize this approach to "energy" constraints and show that there is no infimum-gap when taking the infimum over the family of optimal control problems with control subject to a probability measure dominated by the one defining the energy constraint.

Acknowledgments.

This research was funded in whole or in part by the French National Research Agency (ANR) under the NOCIME project (ANR-23-CE48-0004-03). D.G. acknowledges support from the NSF of Shandong Province (NO. ZR202306020015), the National Key R and D Program of China (NO. 2018YFA0703900), and the NSF of P.R. China (NO. 12031009).

References

1. Avram, F., Freddi, L. and Goreac, G.: Optimal control of a SIR epidemic with ICU constraints and target objectives. Applied Mathematics and Computation, 418:126816 (2022)
2. Bliman, P.-A. and Duprez, M.: How best can finite-time social distancing reduce epidemic final size? Journal of Theoretical Biology, 511:110557 (2021)
3. Bliman, P.-A., Duprez, M., Privat, Y. and Vauchelet, N.: Optimal immunity control and final size minimization by social distancing for the SIR epidemic model. Journal of Optimization Theory and Applications, 189(2):408–436 (2021)
4. Bliman, P.-A. and Rapaport, A.: On the problem of minimizing the epidemic final size for SIR model by social distancing. Proceedings of the 22nd IFAC World Congress, Yokohama, Japan (2023)
5. Freddi, L., Goreac, D., Li, J. and Xu, B.: SIR Epidemics with State-Dependent Costs and ICU Constraints: A Hamilton–Jacobi Verification Argument and Dual LP Algorithms. Applied Mathematics & Optimization, 86(2):23 (2022)
6. Gaitsgory V. and Quincampoix, Q.: On sets of occupational measures gener- ated by a deterministic control system on an infinite time horizon. Nonlinear Analysis: Theory, Methods & Applications, 88:27–41 (2013)
7. Goreac D. and Rapaport, A.: L^∞/L^1 duality results in optimal control problems. Preprint arXiv:2305.02585, under revision for IEEE Transactions on Automatic Control (2023)
8. Kallenberg, O.: Foundations of Modern Probability. Springer, series Probability and its applications, 2nd Edition (2001)
9. Kermack, W. O. and McKendrick, A. G.: Contributions to the mathematical theory of epidemics - Proceedings of the Royal Society, 115A:700–721 (1927)
10. Lasserre, J.-B., Henrion, D., Prieur, C. and Trélat, E.: Nonlinear optimal control via occupation measures and LMI-relaxations. SIAM Journal on Control and Optimization, 47(4):1643–1666 (2008)
11. Miclo, L., Spiro, D. and Weibull, J.: Optimal epidemic suppression under an ICU constraint: An analytical solution. Journal of Mathematical Economics, 102669 (2022)
12. Molina E. and Rapaport, A.: An optimal feedback control that minimizes the epidemic peak in the SIR model under a budget constraint. Automatica, 146:110596 (2022)
13. Morris, D., Rossine, F., Plotkin, J. and Levin, S.: Optimal, near-optimal, and robust epidemic control. Communications Physics, 4(1):1–8 (2021)

13

Optimal Control Synchronization of a Complex Network of Predator-Prey Systems

Cristiana J. Silva[1] and Guillaume Cantin[2]

[1] Department of Mathematics, ISTA, Iscte - Instituto Universitário de Lisboa, Av. das Forças Armadas, 1649-026 Lisboa, Portugal and Center for Research and Development in Mathematics and Applications (CIDMA), Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal.
cristiana.joao.silva@iscte-iul.pt

[2] Laboratoire des Sciences du Numérique de Nantes, Nantes Université, CNRS UMR 6004, France.
guillaume.cantin@univ-nantes.fr

Summary. In this work, we consider a complex network of predator-prey systems, modeling the ecological dynamics of interacting species living in a fragmented environment. We consider non-identical instances of a Lotka-Volterra model with Holling type II functional response. We study optimal control problems, for the minimization of the default of synchronization in the complex network, where the controls reproduce the implementation of ecological corridors. The main goal is to restore biodiversity of life species in a heterogeneous habitat by reaching at least a global coexistence equilibrium, or in a better scenario, a global limit cycle which would guarantee biological oscillations, which means rich life dynamics.

13.1 Introduction

Optimizing the biodiversity restoration of life species in a fragmented habitat through the implementation of ecological corridors between each component of the fragmented environment, while maintaining human activity at a reasonable level, is a challenge that we wish to study in this work. We assume that the geographical habitat of the species is perturbed by the anthropic extension, so that it is fragmented in several patches. This fragmentation is likely to alter the equilibrium of the ecological system. In order to model such a fragmented environment, we consider a complex network of predator-prey models, first proposed in [16], which reproduces the heterogeneous natural environment, that is perturbed by fragmentation, by coupling several patches on which interacting wild species are living. To construct the complex network model, on each patch, the ecological inter-species dynamics are modeled by a Lotka-Volterra predator-prey model with Holling type II functional response, which is able to describe several biological dynamics, such as extinction, coexistence or ecological cycles (see e.g. [21, 31, 33]). Here, each patch can admit its own dynamic, that is, the local components of the network can for instance exhibit an extinction equilibrium on some places, whereas other places can present cycles [16]. Moreover, migrations of biological individuals in space, between each component of the fragmented environment, are taken into account by coupling the patches of the network (see Figure 13.1, where the disks model the patches of a fragmented environment, where the inter-species dynamics of Lotka-Volterra type occur, and the arrows model the ecological corridors which can be implemented between these patches, so as to increase the migrations in space of the species between each patch).

In [16], sufficient conditions of synchronization of the local dynamics, under a variation of the couplings, are proved, namely a theorem for near-synchronization is proved, which guarantees that the complex network remains in a neighborhood of a synchronization state, provided the coupling strength is strong enough, even if the local behaviors are non-identical. This result improves the sufficient conditions of synchronization for the particular case of identical dynamics, proved in [15]. The relevance of synchronization in complex networks has been highlighted by several studies in different areas, such that, coupled oscillators, networks of chemical reactions, neural networks or meta-populations models (see for instance [3, 5, 8, 29] and the references therein).

The main goal of this paper is to optimize the synchronization of the complex network, through optimal control theory. The possibility to reach synchronization through an optimal control process has been studied in [14], with an application to an epidemic model, or in [13], with an application to a panic model. Meanwhile, the dynamics of Lotka-Volterra type models have been widely analyzed (see for instance [7] or [24]) and the optimal control of such models has been studied in [18, 22], but not in the framework of complex networks. On the optimal control of periodic solutions, the non-existence of limit cycle was proved in [9], and periodic optimal control problems have been analyzed in [6, 20, 27, 38].

Focusing on Lotka Volterra models, in [32], a fish population optimal control problem is studied considering the Lotka Volterra model

$$\begin{cases} \dot{x}_1(t) = x_1(t) - x_1(t)x_2(t) \\ \dot{x}_2(t) = -x_2(t) + x_1(t)x_2(t) \\ x_i(t_0) = x_{i0}\,, \quad i = 1, 2\,, \end{cases} \tag{13.1}$$

where $x_1(t)$ and $x_2(t)$ represent the biomass of the prey and predator species, respectively, with initial state conditions $x_i(t_0) = x_{i0}$, $i = 1, 2$.

The main goal in [32] is to bring the control system (13.1) close to a steady state to avoid the high fluctuations that cause economical problems. More precisely, the authors choose to vary the fishing quota for a certain time span $T - t_0$. Adding an objective functional that punishes deviation from the steady state $\tilde{x} = (1, 1)^T$ for $u(t) = 0$, and $\tilde{x} = (1 + c_2, 1 - c_1)^T$ for $u(t) = 1$, respectively. The following optimal control problem is analyzed

$$\min_u \int_{t_0}^{T} (x_1(t) - 1)^2 + (x_2(t) - 1)^2 \, dt$$

such that

$$\begin{cases} \dot{x}_1(t) = x_1(t) - x_1(t)x_2(t) - c_1 x_1(t)u(t) \\ \dot{x}_2(t) = -x_2(t) + x_1(t)x_2(t) - c_2 x_2(t)u(t) \\ x_i(t_0) = x_{i0}\,, \quad u(t) \in [0, 1]\,. \end{cases} \tag{13.2}$$

and where control function $u(t)$ describes the percentage of the fleet that is actually fished at time t. The parameters c_1 and c_2 indicate how many fish would be caught by the entire fleet. For optimization methods to solve the previous optimal control problem see, e.g. [32] and also [40].

In [23], the turnpike phenomenon is illustrated by studying an optimal control problem, considering the control system (13.2) with $c_1 = 0.4$ and $c_2 = 0.2$, initial conditions $(x_1(0), x_2(0)) = (0.5; 0.7)$, final time $T = 60$, and with the following cost functional:

$$\min_u \frac{1}{2}\int_0^T (x_1(t)-1)^2 + x_2(t)^2 + u(t)^2\, dt\,.$$

An analogous control system is also considered in [17], where optimal strategies for reaching fixed steady states, namely co-existence of both species are studied. General turnpike results for optimal control problems have been established in [35, 36, 37].

In [4], a Mayer-type optimal control problem is studied for Lotka-Volterra systems with a hunter population, where the goal is to maximize the population of both species at the final time, that is, $x_1(T) + x_2(T)$ and the control u represents the hunting proportionality factor. In [39] the authors analyze analytically a Mayer-type optimal control problem applied to a two dimensional Lotka Volterra system. In [10], the shooting method is applied to a minimal time optimal control problem with the control system from [23, 32].

Recently, advances on geometrical optimal control theory of Generalized Lotka-Volterra systems applied to the intestinal microbiome have been developed in [11, 12].

In this paper, we aim to study a more general problem than the ones studied in [23, 32] reaching at least a global coexistence equilibrium, or in a better scenario, a global limit cycle, instead of a fixed steady state. The optimal control of limit cycles in medical models applied to diabetes and heart attack problem was studied in [19] and [27], respectively. Moreover, the nonexistence of limit cycle for an optimal control problem applied to a diabetes model was proved in [9]. In this chapter, we first consider a controlled complex network of Lotka-Volterra systems, where the strength of the migrations of biological individuals in system (13.6) is replaced by control functions, reproducing the implementation of ecological corridors We prove that a solution of the controlled complex network can reach a near-synchronization state, under sufficient conditions which highlight the importance to consider a positive lower on the controls functions. After, we study optimal control problems where the main goal is the minimization of the default of synchronization in the complex network. We consider different cost functionals taking into account that the dynamics of the controlled complex network ensure the conservation of both species, namely, our goal is to impose synchronization or synchronization of limit cycles. Therefore, the solutions of the optimal control problems lead to a restoration of the biodiversity of life species in a heterogeneous habitat by reaching at least a global coexistence equilibrium, or in a better scenario, a global limit cycle which would guarantee biological oscillations, which means rich life dynamics.

This chapter is organized as follows. In Section 13.2, we recall the construction of the uncontrolled complex network of Lotka-Volterra systems and the near synchronization results, from [16]. In Section 13.3, we propose a controlled complex network, where the strength of the migrations of biological individuals in the Lotka-Volterra systems is replaced by control functions, and prove a sufficient condition for the near-synchronization of the solutions of the controlled system. In Section 13.4, we consider optimal control problems, in order to exert a command on the global behavior of the controlled complex network. To model the goal of restoring biodiversity and biological dynamics in a fragmented environment, we define appropriate cost functionals where the conservation of species is guaranteed by imposing synchronization or synchronization of limit cycles. We end this chapter with Section 13.5 with some conclusions and future work.

13.2 Setting of the Complex Network of Lotka-Volterra Systems

Based on the previous work [16], we present the construction of a complex network of Lotka-Volterra systems, which describes the dynamics of interacting species living in a fragmented environment, and recall important near synchronization results, proved in [16], for the uncontrolled complex network.

13.2.1 Lotka-Volterra predator-prey model with Holling type II functional response

Let us consider a biological environment in which two species interact. We assume that the densities of the species are determined by a predator-prey model of Lotka-Volterra type, which can be written by:

$$\begin{cases} \dot{x} = rx(1-x) - \dfrac{cxy}{\alpha + x}, \\ \dot{y} = -dy + \dfrac{cxy}{\alpha + x}. \end{cases} \tag{13.3}$$

Here, x and y denote the prey and predator density, respectively; $\dot{x}$ and $\dot{y}$ denote their derivatives with respect to the time variable t. The parameters r, c, d and α are positive coefficients; r is the birth rate of the preys, d is the mortality rate of predators, and c, α determine the non-linear interaction between preys and predators (see, for instance [24], for a deep study on the dynamics of predator-prey system (13.3)). Depending on the values of the parameters r, c, d, α, the solutions of system (13.3) can be attracted to a coexistence equilibrium, to an extinction equilibrium or to a limit cycle. The extinction equilibrium is denoted $E_0 = (0, 0)$. The coexistence equilibrium E_1, which implies persistence of each specie, is given, for $c \neq d$, by

$$E_1 = \left(\frac{\alpha d}{c-d}, \frac{r\alpha}{c-d} \left(1 - \frac{\alpha d}{c-d} \right) \right). \tag{13.4}$$

System (13.3) also admits the equilibrium $E_2 = (1, 0)$. Let us introduce the critical value α_0 given by

$$\alpha_0 = \frac{c-d}{c+d}. \tag{13.5}$$

It is well-known (see for instance [24], Chapter 3 or [7], Section 3.4.1) that system (13.3) undergoes a Hopf bifurcation at $\alpha = \alpha_0$. For $\alpha < \alpha_0$, a stable limit cycle bifurcates from the persistence equilibrium E_1. Therefore, for α small enough, system (13.3) presents oscillations, which are interpreted as healthy ecological cycles.

13.2.2 Complex network of predator-prey models for a fragmented environment

Next, we assume that the geographical habitat of the species is perturbed by the anthropic extension, so that it is fragmented in several patches. This fragmentation is likely to alter the equilibrium of the ecological system. In order to model such a fragmented environment, we construct a complex network of predator-prey models as follows.

First, let $n > 0$ denote the number of patches on the fragmented environment. On each patch $i \in \{1, \ldots, n\}$, we denote by (x_i, y_i) the densities of preys and predators respectively. We assume that each patch $i \in \{1, \ldots, n\}$ can be connected to other patches and we denote by $\mathcal{N}_i \subset \{1, \ldots, n\}$ the

set of patches which are connected to patch i. We assume that migrations of biological individuals can occur between two connected patches, at rates σ_1 for preys and σ_2 for predators. In this way, the dynamics of the fragmented environment are determined by the following complex network:

$$\begin{cases} \dot{x}_i = r_i x_i(1-x_i) - \dfrac{c_i x_i y_i}{\alpha_i + x_i} - \sigma_1 \displaystyle\sum_{j \in \mathscr{N}_i} (x_i - x_j), \\ \dot{y}_i = -d_i y_i + \dfrac{c_i x_i y_i}{\alpha_i + x_i} - \sigma_2 \displaystyle\sum_{j \in \mathscr{N}_i} (y_i - y_j), \end{cases} \tag{13.6}$$

for $1 \leq i \leq n$, with $\sigma_1 \geq 0$ and $\sigma_2 \geq 0$.

We emphasize that the parameters r_i, c_i, d_i, α_i can differ from one patch to another, which means that the ecological dynamics are non-identical within the fragmented environment. For instance, some patches could present limit cycles, whereas other patches could exhibit an extinction of both species. Note also that the couplings are symmetric, which means that if the species x_i, y_i of patch i can move towards some patch j, then the species x_j, y_j of patch j can conversely move towards patch i.

One remarkable case of fragmented environment is that of a complete graph topology, for which we have $\mathscr{N}_i = \{1, \ldots, n\} \setminus \{i\}$; this situation means that each patch is connected to all other patches. At the opposite, if the coupling parameters σ_1, σ_2 are equal to 0, then no migration of individuals occur in the network.

Let us now introduce some notations. Let $X = \big((x_1, y_1), \ldots, (x_n, y_n)\big)^\top \in \mathbb{R}^{2n}$. For each $i \in \{1, \ldots, n\}$, we denote

$$\begin{aligned} &\lambda_i = (r_i, c_i, d_i, \alpha_i)^\top \in \mathbb{R}^4, \\ &f_1(x_i, y_i, \lambda_i) = r_i x_i(1-x_i) - \frac{c_i x_i y_i}{\alpha_i + x_i}, \\ &f_2(x_i, y_i, \lambda_i) = -d_i y_i + \frac{c_i x_i y_i}{\alpha_i + x_i}, \\ &g_1(x_i, X, \sigma_1) = -\sigma_1 \sum_{j \in \mathscr{N}_i} (x_i - x_j), \\ &g_2(y_i, X, \sigma_2) = -\sigma_2 \sum_{j \in \mathscr{N}_i} (y_i - y_j). \end{aligned} \tag{13.7}$$

We also denote $\sigma = (\sigma_1, \sigma_2)^\top \in \mathbb{R}^2$ and

$$\begin{aligned} &\Lambda = (\lambda_1, \ldots, \lambda_n)^\top \in \mathbb{R}^{4n}, \\ &F(X, \Lambda) = \Big(f_1(x_1, y_1, \lambda_1), f_2(x_1, y_1, \lambda_1), \ldots, f_1(x_n, y_n, \lambda_n), f_2(x_n, y_n, \lambda_n)\Big)^\top \in \mathbb{R}^{2n}, \\ &G(X, \sigma) = \Big(g_1(x_1, X, \sigma_1), g_2(y_1, X, \sigma_2), \ldots, g_1(x_n, X, \sigma_1), g_2(y_n, X, \sigma_2)\Big)^\top \in \mathbb{R}^{2n}. \end{aligned} \tag{13.8}$$

With these notations, the complex network (13.6) can be written under the following short form

$$\dot{X} = F(X, \Lambda) + G(X, \sigma). \tag{13.9}$$

13.2.3 Review of known results

In this section, we recall recent results obtained in [16], that motivate the controlled system and the optimal control problem studied in the present work.

The following theorem guarantees that the complex network problem determined by system (13.9) admits global solutions.

Theorem 1 ([16]). *Let $X_0 \in (\mathbb{R}^+)^{2n}$. Then the complex network problem determined by* (13.9) *and $X(0) = X_0$ admits a unique global solution $X(t, X_0)$ defined on $[0, +\infty)$, whose components are non-negative.*

Furthermore, the flow induced by Equation (13.9) *admits a positively invariant region Θ which is compact in $(\mathbb{R}^+)^{2n}$.*

One remarkable property of complex network is the *synchronization* property. The following definition is classical.

Definition 1 (Synchronization). *Let $i, j \in \{1, \ldots, n\}$ such that $i \neq j$. We say that the patches i and j of the complex network* (13.9) *synchronize in Θ if, for any initial condition $X_0 \in \Theta$, the solution of* (13.9) *starting from X_0 satisfies*

$$\lim_{t \to +\infty} \Big(|x_i(t) - x_j(t)|^2 + |y_i(t) - y_j(t)|^2 \Big) = 0.$$

We say that the complex network (13.9) *synchronizes in Θ if every pair (i, j) of patches synchronizes in Θ.*

In the case of a complex network of nonidentical systems, it is not always possible to prove that a synchronization state is reached. Therefore, we are led to introduce a relaxed definition of synchronization, called *near-synchronisation*.

Definition 2 (Near-synchronization). *Let $i, j \in \{1, \ldots, n\}$ such that $i \neq j$. We say that the patches i and j of the complex network* (13.9) *nearly synchronize in Θ with respect to $\tilde{\sigma}$ if, for any initial condition $X_0 \in \Theta$, and for any $\varepsilon > 0$, the solution of* (13.9) *starting from X_0 satisfies*

$$0 \leq \lim_{t \to +\infty} \Big(|x_i(t) - x_j(t)|^2 + |y_i(t) - y_j(t)|^2 \Big) < \varepsilon,$$

for $\tilde{\sigma}$ sufficiently large.

We say that the complex network (13.9) *nearly synchronizes in Θ if every pair (i, j) of patches nearly synchronizes in Θ.*

In [16], sufficient conditions of near-synchronization have been established for the complex network (13.9) with non-identical systems. We recall below these results. For the sake of simplicity, it is assumed that the graph underlying the complex network (13.9) is complete, that is, each patch is connected to all other patches; equivalently, we have $\mathcal{N}_i = \{1, \ldots, n\} \setminus \{i\}$ for $1 \leq i \leq n$, where $\mathcal{N}_i$ denotes the finite set of patches which are connected to patch i. For all $i, j \in \{1, \ldots, n\}$, we introduce the energy function $E_{i,j}$ defined along the trajectories of the complex network by

$$E_{i,j}(t) = \frac{1}{2} \Big[|x_i(t) - x_j(t)|^2 + |y_i(t) - y_j(t)|^2 \Big], \tag{13.10}$$

and for $\lambda_i = (r_i, d_i, c_i, \alpha_i), \lambda_j = (r_j, d_j, c_j, \alpha_j) \in \mathbb{R}^4$, denote

$$\|\lambda_i - \lambda_j\|_\infty = \max \Big\{ |r_i - r_j|, |d_i - d_j|, |c_i - c_j|, |\alpha_i - \alpha_j| \Big\}.$$

The next theorem establishes an estimate of the energy functions $E_{i,j}$ defined by (13.10).

Theorem 2 (Near-synchronization of the uncontrolled complex network predator-prey model, [16]). *There exist positive constants η, δ such that, for any initial condition $X_0 \in \Theta$, the solution of the complex network* (13.9) *starting from X_0 satisfies*

$$\dot{E}_{i,j}(t) \leq \eta \left\|\lambda_i - \lambda_j\right|_\infty E_{i,j}^{1/2}(t) + \Big[\delta - 2n\tilde{\sigma}\Big] E_{i,j}(t), \quad t > 0, \tag{13.11}$$

where $\tilde{\sigma} = \min\{\sigma_1, \sigma_2\}$.[3]

Furthermore, the constants η and δ do not depend on the coupling parameters σ_1, σ_2.

We now recall important consequences of Theorem 2.

Corollary 1 ([16]). *Assume that $\lambda_i = \lambda_j$ for some $i, j \in \{1, \ldots, n\}$. Then the patches i and j synchronize if the following condition is fulfilled:*

$$2n\tilde{\sigma} > \delta. \tag{13.12}$$

If $\lambda_i = \lambda_j$ for all $i, j \in \{1, \ldots, n\}$, then obviously the whole network synchronizes under condition (13.12). Next, since the constant δ does not depend on the coupling parameters σ_1, σ_2, the sufficient condition (13.12) can easily be satisfied, provided the number n of patches in the network is sufficiently large, or provided the minimum coupling strength $\tilde{\sigma} = \min\{\sigma_1, \sigma_2\}$ is sufficiently large.

From the ecological point of view, increasing the number n of patches in the network would correspond to a worse fragmentation of the habitat, which is not a reasonable strategy for our purposes. However, increasing the minimum coupling strength $\tilde{\sigma}$ can be realized by providing wider ecological corridors.

The non trivial case of Theorem 2 corresponds to a complex network of non-identical patches, for which we have $\lambda_i \neq \lambda_j$ for at least one pair $(i, j) \in \{1, \ldots, n\}^2$. In that case, the synchronization state $\{(x_i, y_i) = (x_j, y_j)\}$ is likely to present a soft loss of stability. Indeed, it is well-known that the solution w of the Bernoulli equation

$$\dot{w}(t) = \eta \left\|\lambda_i - \lambda_j\right|_\infty w^{1/2}(t) + \Big[\delta - 2n\tilde{\sigma}\Big] w(t), \quad t > 0, \tag{13.13}$$

converges towards a positive limit given by

$$\lim_{t \to +\infty} w(t) = \left(\frac{\eta \left\|\lambda_i - \lambda_j\right|_\infty}{\delta - 2n\tilde{\sigma}} \right)^2,$$

provided $w(0) > 0$. We obtain the following corollaries.

Corollary 2 ([16]). *The energy function $E_{i,j}$ defined by* (13.10) *along a solution of the complex network* (13.9) *starting from $X_0 \in \Theta$, satisfies*

$$0 \leq \limsup E_{i,j}(t) \leq \left(\frac{\eta \left\|\lambda_i - \lambda_j\right|_\infty}{\delta - 2n\tilde{\sigma}} \right)^2. \tag{13.14}$$

complex

[3] In this paper, we correct a misprint of [16], since the quantity $2(n-1)\tilde{\sigma}$ in [16] should be $2n\tilde{\sigma}$.

Corollary 3 ([16]). *The complex network* (13.9) *nearly synchronizes in* Θ *with respect to the the minimum coupling strength* $\tilde{\sigma}$.

Remark 1. Note that near-synchronization can occur in the complex network (13.9) without imposing any particular asymptotic dynamics; for example, the complex network could synchronize towards a global dynamic of extinction, towards a global dynamic of coexistence, or towards a global dynamic of limit cycles (it could even happen that a new dynamic emerges from the complex network structure).

In the next section, we construct a controlled complex network of Lotka-Volterra systems, where the strength of the migrations of biological individuals in system (13.6) is replaced by control functions. We prove that a solution of the controlled system can reach a near-synchronization state, under sufficient conditions which highlight the importance to consider a positive lower bound on the controls functions.

13.3 Controlled Synchronization

In this section, we present a controlled complex network of Lotka-Volterra systems, where the strength of the migrations of biological individuals in system (13.6) is replaced by control functions $u_{i,j}(\cdot) \in L^{\infty}(0,T)$, $1 \leq i,j \leq n$.

The main goal is to restore biodiversity and biological dynamics in a fragmented environment. Our aim is to reach at least a global coexistence equilibrium, or better, a global limit cycle which would guaranty biological oscillations, which means rich life dynamics.

13.3.1 Setting of the control system

We consider the control system, given by

$$\begin{cases} \dot{x}_i = r_i x_i(1-x_i) - \dfrac{c_i x_i y_i}{\alpha_i + x_i} - \displaystyle\sum_{j \in \mathscr{N}_i} u_{i,j}(t)(x_i - x_j), \\ \dot{y}_i = -d_i y_i + \dfrac{c_i x_i y_i}{\alpha_i + x_i} - \displaystyle\sum_{j \in \mathscr{N}_i} u_{i,j}(t)(y_i - y_j), \end{cases} \tag{13.15}$$

for $1 \leq i \leq n$, with the following control constraints

$$u_{\min} \leq u_{i,j}(t) \leq u_{\max} \quad \forall t \in [0,T], \quad \text{for} \quad \text{all} \quad (i,j) \in \{1,\ldots,n\}^2, \tag{13.16}$$

with $u_{\min} > 0$. Hence, the set of admissible control functions is given by

$$\Omega = \left\{ u_{i,j}(\cdot) \in L^{\infty}(0,T) \mid u_{\min} \leq u_{i,j}(t) \leq u_{\max} \quad \forall t \in [0,T], \forall \quad (i,j) \in \{1,\ldots,n\}^2 \right\}.$$

Moreover, we consider fixed initial conditions $X(0) = X_0 \in (\mathbb{R}^+)^{2n}$.

Analogously to equation (13.9), we can write the controlled system (13.15) in the form

$$\dot{X} = F(X,\Lambda) + G(X, \{u_{i,j}\}_{1 \leq i,j \leq n}).$$

The following theorem guarantees the existence of a positively invariant region for a solution of the controlled system (13.15).

Let $a_0 = \sum_{i=1}^{n} a_i$, $b_0 = \min_{1\leq i\leq n} b_i$, $d_0 = \min_{1\leq i\leq n} d_i$,$c_0 = \min\{b_0, d_0\}$, where the coefficients a_i, b_i are chosen such that

$$r_i s(1-s) \leq a_i - b_i s,$$

for all $s \in \mathbb{R}$.

Theorem 3 (Positively invariant region). *The region Θ defined by*

$$\Theta = \left\{ X = (x_i, y_i)_{1\leq i\leq n} \in (\mathbb{R}^+)^{2n} \;\Big|\; \sum_{1\leq i\leq n} (x_i + y_i) \leq \frac{a_0}{c_0} \right\} \tag{13.17}$$

is positively invariant for the flow induced by the controlled system (13.15).

Proof. Let $P(t) = \sum_{1\leq i\leq n} \big(x_i(t) + y_i(t)\big)$. Summing the equations of the complex network problem, we easily prove that

$$\dot{P} + c_0 P \leq a_0,$$

since the sum of the control couplings vanishes. Applying Gronwall lemma finishes the proof.

13.3.2 Near-synchronization of the controlled system

In this section, we prove that a solution of the controlled system (13.15) can reach a near-synchronization state, under sufficient conditions which highlight the importance to consider a positive lower bound on the controls functions. The following theorem establishes an estimate of the energy function corresponding to a solution of the control system (13.15).

Theorem 4 (Energy estimate the controlled system). *Assume that the graph $\mathscr{G}$ underlying the complex network* (13.15) *is a complete graph. Then the energy functions $E_{i,j}$ defined by*

$$E_{i,j}(t) = \frac{1}{2}\big[(x_i - x_j)^2 + (y_i - y_j)^2\big]$$

satisfy the following estimate:

$$0 \leq \limsup_{t\to+\infty} E_{i,j}(t) \leq \left(\frac{\eta \left\|\lambda_i - \lambda_j\right|_\infty + \tilde{K}(n-2)(u_{\max} - u_{\min})}{\delta - 2nu_{\min}} \right)^2. \tag{13.18}$$

Proof. We compute

$$\begin{aligned}
\frac{dE_{i,j}}{dt} &= (\dot{x}_i - \dot{x}_j)(x_i - x_j) + (\dot{y}_i - \dot{y}_j)(y_i - y_j) \\
&= \big[f_i(x_i, y_j) - f_j(x_j, y_j)\big](x_i - x_j) \\
&\quad + \left[-\sum_{k\in\mathscr{N}_i} u_{i,k}(t)(x_i - x_k) + \sum_{k\in\mathscr{N}_j} u_{j,k}(t)(x_j - x_k) \right](x_i - x_j) \\
&\quad + \big[g_i(x_i, y_j) - g_j(x_j, y_j)\big](y_i - y_j) \\
&\quad + \left[-\sum_{k\in\mathscr{N}_i} u_{i,k}(t)(y_i - y_k) + \sum_{k\in\mathscr{N}_j} u_{j,k}(t)(y_j - y_k) \right](y_i - y_j).
\end{aligned}$$

Next, we write

$$\sum_{k\in\mathscr{N}_i} u_{i,k}(t)(x_i-x_k)-\sum_{k\in\mathscr{N}_j} u_{j,k}(t)(x_j-x_k)=u_{i,j}(x_i-x_j)-u_{j,i}(x_j-x_i)$$
$$+\sum_{k\in\mathscr{N}_i\setminus\{j\}} u_{i,k}(t)(x_i-x_k)-\sum_{k\in\mathscr{N}_j\setminus\{i\}} u_{j,k}(t)(x_j-x_k).$$

If the graph if complete, then we have $\mathscr{N}_i\setminus\{j\}=\mathscr{N}_j\setminus\{i\}$. Moreover, we have $u_{i,j}=u_{j,i}$. We obtain

$$\sum_{k\in\mathscr{N}_i} u_{i,k}(t)(x_i-x_k)-\sum_{k\in\mathscr{N}_j} u_{j,k}(t)(x_j-x_k)=2u_{i,j}(x_i-x_j)$$
$$+\sum_{k\in\mathscr{N}_i\setminus\{j\}} u_{i,k}(t)(x_i-x_k)-\sum_{k\in\mathscr{N}_j\setminus\{i\}} u_{j,k}(t)(x_j-x_k)$$

We introduce $\mathscr{S}_{i,j}=\mathscr{N}_i\setminus\{j\}=\mathscr{N}_j\setminus\{i\}$ and we observe that

$$\sum_{k\in\mathscr{S}_{i,j}} u_{i,k}(x_i-x_k)-\sum_{k\in\mathscr{S}_{i,j}} u_{j,k}(x_j-x_k)=\sum_{k\in\mathscr{S}_{i,j}}(u_{i,k}x_i-u_{j,k}x_j)-\sum_{k\in\mathscr{S}_{i,j}}(u_{i,k}-u_{j,k}x_k).$$

We write

$$\begin{aligned}(u_{i,k}x_i-u_{j,k}x_j)(x_i-x_j)&=(u_{i,k}x_i-u_{i,k}x_j)(x_i-x_j)+(u_{i,k}x_j-u_{j,k}x_j)(x_i-x_j)\\&=u_{i,k}(x_i-x_j)^2+x_j(u_{i,k}-u_{j,k})(x_i-x_j)\\&\geq u_{\min}(x_i-x_j)^2+x_j(u_{i,k}-u_{j,k})(x_i-x_j).\end{aligned}$$

Similarly, we have

$$(u_{i,k}x_i-u_{j,k}x_j)(x_i-x_j)\geq u_{\min}(x_i-x_j)^2+x_i(u_{i,k}-u_{j,k})(x_i-x_j).$$

It follows that

$$(u_{i,k}x_i-u_{j,k}x_j)(x_i-x_j)\geq u_{\min}(x_i-x_j)^2+\frac{x_i+x_j}{2}(u_{i,k}-u_{j,k})(x_i-x_j).$$

We can deduce

$$(x_i-x_j)\sum_{k\in\mathscr{S}_{i,j}}(u_{i,k}x_i-u_{j,k}x_j)\geq u_{\min}(n-2)(x_i-x_j)^2$$
$$+\frac{(x_i+x_j)(x_i-x_j)}{2}\sum_{k\in\mathscr{S}_{i,j}}(u_{i,k}-u_{j,k}),$$

since the set $\mathscr{S}_{i,j}$ contains $(n-2)$ elements. We obtain

$$\begin{aligned}-(x_i-x_j)\Big[\sum_{k\in\mathscr{S}_{i,j}} u_{i,k}(x_i-x_k)-\sum_{k\in\mathscr{S}_{i,j}} u_{j,k}(x_j-x_k)\Big]&\\\leq -2u_{\min}(n-2)E^1_{i,j}&\\-\frac{(x_i+x_j)(x_i-x_j)}{2}\sum_{k\in\mathscr{S}_{i,j}}(u_{i,k}-u_{j,k})&\\+(x_i-x_j)\sum_{k\in\mathscr{S}_{i,j}}(u_{i,k}-u_{j,k})x_k,&\end{aligned}$$

with $E^1_{i,j} = \frac{1}{2}(x_i - x_j)^2$. Next, we estimate

$$\begin{aligned}
-\frac{(x_i + x_j)(x_i - x_j)}{2} &\sum_{k\in\mathscr{S}_{i,j}} (u_{i,k} - u_{j,k}) + (x_i - x_j) \sum_{k\in\mathscr{S}_{i,j}} (u_{i,k} - u_{j,k})x_k \\
&\le \left|\frac{x_i + x_j}{2}\right| |x_i - x_j| \sum_{k\in\mathscr{S}_{i,j}} |u_{i,k} - u_{j,k}| \\
&+ |x_i - x_j| \sum_{k\in\mathscr{S}_{i,j}} |u_{i,k} - u_{j,k}|\,|x_k| \\
&\le K\,|x_i - x_j| \sum_{k\in\mathscr{S}_{i,j}} (u_{\max} - u_{\min}) \\
&+ |x_i - x_j| \sum_{k\in\mathscr{S}_{i,j}} (u_{\max} - u_{\min})K,
\end{aligned}$$

where K is a positive constant such that $|x_i| \le K$ for all i, whose existence is guaranteed by Theorem 3. We obtain

$$\begin{aligned}
-(x_i - x_j)\Big[\sum_{k\in\mathscr{S}_{i,j}} u_{i,k}(x_i - x_k) - \sum_{k\in\mathscr{S}_{i,j}} u_{j,k}(x_j - x_k)\Big] \\
\le -2u_{\min}(n-2)E^1_{i,j} + 2K(n-2)\,|x_i - x_j|\,(u_{\max} - u_{\min}).
\end{aligned}$$

Similarly, we estimate

$$\begin{aligned}
-(y_i - y_j)\Big[\sum_{k\in\mathscr{S}_{i,j}} u_{i,k}(y_i - y_k) - \sum_{k\in\mathscr{S}_{i,j}} u_{j,k}(y_j - y_k)\Big] \\
\le -2u_{\min}(n-2)E^2_{i,j} + 2K(n-2)\,|y_i - y_j|\,(u_{\max} - u_{\min}),
\end{aligned}$$

where $E^2_{i,j} = \frac{1}{2}(y_i - x_j)^2$. Now we come back to estimate $\frac{dE_{i,j}}{dt}$:

$$\begin{aligned}
\frac{dE_{i,j}}{dt} &\le \big[f_i(x_i, y_i) - f_j(x_j, y_j)\big](x_i - x_j) + \big[g_i(x_i, y_i) - g_j(x_j, y_j)\big](y_i - y_j) \\
&- 2u_{\min} n E_{i,j} + 2K(n-2)\big[\,|x_i - x_j| + |y_i - y_j|\,\big] \\
&\le \big[\eta\,\|\lambda_i - \lambda_j|_\infty + \tilde{K}(n-2)(u_{\max} - u_{\min})\big]E^{1/2}_{i,j} + \big[\delta - 2nu_{\min}\big]E_{i,j}.
\end{aligned}$$

Applying Gronwall Lemma and comparing with the solution of the Bernoulli equation (13.13) finishes the proof.

Corollary 4. *The controlled system* (13.15) *nearly synchronizes with respect to* $u_{\min}$*, provided* $(u_{\max} - u_{\min})$ *is uniformly bounded.*

Remark 2 (Generalization to other models). We can consider a complex network of ecological systems of the general form

$$\dot{x} = xM(x, y, \lambda), \qquad \dot{y} = yN(x, y, \lambda), \tag{13.19}$$

where M and N are regular functions defined in $\mathbb{R}\times\mathbb{R}\times\mathbb{R}^p$ (see notably [7], [28]), by considering the following system:

$$\begin{cases} \dot{x}_i = xM_i(x_i, y_i) - \sum\limits_{j \in \mathcal{N}_i} u_{i,j}(x_i - x_j), \\ \dot{y}_i = yN_i(x_i, y_i) - \sum\limits_{j \in \mathcal{N}_i} u_{i,j}(y_i - y_j), \end{cases} \tag{13.20}$$

with $1 \leq i \leq n$ and control functions $u_{i,j}$. Here, the functions M_i and N_i are non identical instances of the functions M, N defined in (13.19), that is

$$M_i(x_i, y_i) = M(x_i, y_i, \lambda_i), \qquad N_i(x_i, y_i) = N(x_i, y_i, \lambda_i),$$

with $\lambda_1, \lambda_2, \ldots, \lambda_n \in \mathbb{R}^p$. Theorem 2 and its corollaries can easily be generalized to this setting, with a complex network of the form (13.20), under the single assumption that there exists constants $\eta > 0$ and $\delta > 0$ such that

$$\begin{pmatrix} f_i(x_i, y_i) - f_j(x_j, y_j) \\ g_i(x_i, y_i) - g_j(x_j, y_j) \end{pmatrix} \cdot \begin{pmatrix} x_i - x_j \\ y_i - y_j \end{pmatrix} \leq \eta \left\| \lambda_i - \lambda_j \right|_\infty E_{i,j}^{1/2} + \delta E_{i,j}. \tag{13.21}$$

Remark 3 (Non complete graph topologies). Recent results have been obtained (see [2]) for synchronization in non complete graph topologies.

The previous result motivates the setting of an optimal control problem, so as to exert a command on the dynamics of the complex network (13.9) and to reach a synchronization state, even in the case of non-identical patches.

13.4 Optimal Control Synchronization

Considering the controlled complex network of Lotka-Volterra systems (13.15), we propose an optimal control problem, in order to exert a command on the global behavior of this complex network. To model the goal of restoring biodiversity and biological dynamics in a fragmented environment, we need define an appropriate cost functional.

The choice of the cost functional has an important role on the optimal synchronization of the complex network (see e.g. [25] for a study on the role of the objective functional in optimal control problems applied to compartmental models for biomedical therapies). Here, our main focus on the choice of the cost functional is not only on the properties of the optimal control solution but mainly on the dynamics of the state functions x_i, y_i ensuring the conservation of both species.

In what follows, we will consider different cost functionals where the conservation of species is guaranteed by:

i. imposing synchronization;
ii. synchronization of limit cycles.

Impose synchronization – optimal solutions may kill *limit cycles or damped oscillations*

For the scenario "*imposing synchronization*", we consider the optimal control problem of determining $X^*(\cdot)$ associated to the admissible controls $u_{i,j}^*(\cdot) \in \Omega$ on the time interval $[0, T]$, satisfying the controlled system $\dot{X} = F(X, \Lambda) + G(X, \{u_{i,j}\}_{1 \leq i,j \leq n})$, given by (13.15), the fixed initial conditions $X(0) = X_0 \in (\mathbb{R}^+)^{2n}$ and minimizing the one of the following cost functionals:

$$J_1 = \int_0^T \sum_{i \neq j} \Big[(x_i(t) - x_j(t))^2 + (y_i(t) - y_j(t))^2 \Big] \, dt. \tag{13.22}$$

or

$$J_2 = \sum_{i \neq j} \Big[(x_i(T) - x_j(T))^2 + (y_i(T) - y_j(T))^2 \Big]. \tag{13.23}$$

In this case, we emphasize that maximizing synchronization can *kill* limit cycles or damped oscillations possibly occurring in the case of constant couplings (σ_1, σ_2). Indeed, let us consider as a simple example a four nodes network, associated with a complete graph topology (as shown in Figure 13.1). The parameters of each patch are given in Table 13.1, and the initial conditions were randomly generated between 0 and 1. If the couplings are fixed to $\sigma_1 = \sigma_2 = 1$, then the local dynamics are synchronized towards the same damped oscillations (when $\alpha_i = 0.4$, $1 \leq i \leq 4$, see Figure 13.2) or towards the same limit cycle (when $\alpha_i = 0.3$, $1 \leq i \leq 4$, see Figure 13.3). However, when the couplings are determined by a control associated with the functionals J_1, J_2 given by (13.22), (13.23), then we observe that the oscillations vanish, as depicted in Figure 13.4. In other words, the optimality criterion can lead to an *unexpected* emergent behavior.

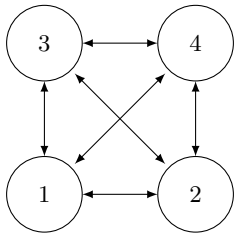


Fig. 13.1: Simple 4 nodes complex network with a complete graph topology.

Table 13.1: Values of the parameters for the example of a 4 nodes network.

Parameter	**Value**	**Parameter**	**Value**
r_1	0.8	r_3	0.9
d_1	0.98	d_3	0.7
c_1	1.6	c_3	1.6
α_1	0.3, 0.4	α_3	0.3, 0.4
r_2	0.8	r_4	0.8
d_2	0.6	d_4	0.75
c_2	1.6	c_4	1.6
α_2	0.3, 0.4	α_4	0.3, 0.4

Remark 4. The non nonexistence of a limit cycle in an optimal control problem applied to a diabetes model was proved in [9].

This lead us to consider an alternative cost functional for which the optimal solution is attracted to a local limit cycle.

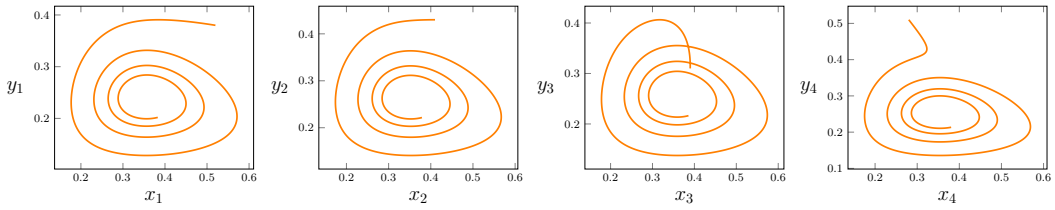


Fig. 13.2: Synchronization of damped oscillations in a four nodes network with constant couplings. Even if the initial conditions and parameters are distinct from one patch to another, the local dynamics are synchronized towards the same damped oscillations.

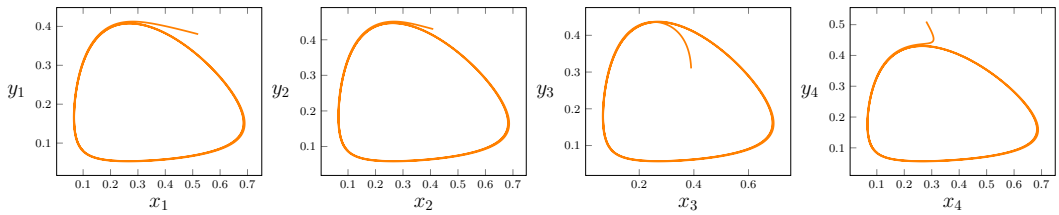


Fig. 13.3: Synchronization of limit cycles in a four nodes network with constant couplings.

How to reach a limit cycle?

Suppose that without couplings, the local dynamics are attracted to local limit cycles. Moreover, suppose that we can synchronize (or near synchronize) these local limit cycles with a constant coupling strength.

We denote by $\gamma(t) = \big(\zeta(t), \xi(t)\big)_{0\leq t\leq \phi}$ a parametrization of the global limit cycle (with period ϕ), obtained by synchronization with a constant coupling strength.

Then we can try to preserve and reach this cycle in an optimal control problem by minimizing the cost functional

$$J_\gamma\left(x_i, y_i\right) = \sum_{k=0}^{k^*} \int_{T+k\phi}^{T+(k+1)\phi} \sum_{i=1}^{n} \Big[\left(x_i(t) - \zeta(t)\right)^2 + \left(y_i(t) - \xi(t)\right)^2 \Big] \, dt \,, \tag{13.24}$$

where T is a positive time such that the transitional dynamics occurs in $[0, T]$, and k^* is the number of periods of oscillations around the limit cycles. Considering again the four nodes network given in Figure 13.1 and the parameters given in Table 13.1, we show in Figure 13.5 a first simulation of the control problem determined by the functional (13.24). We observe that controls can be found to maintain oscillations. It is now our aim to analyze the properties of these controls in a rigorous framework.

13.4.1 Optimal control problem

Our goal is to find the optimal control solution that reaches a limit cycle γ, ensuring the preservation of the ecological biodiversity in a fragmented environment. We consider the optimal control problem

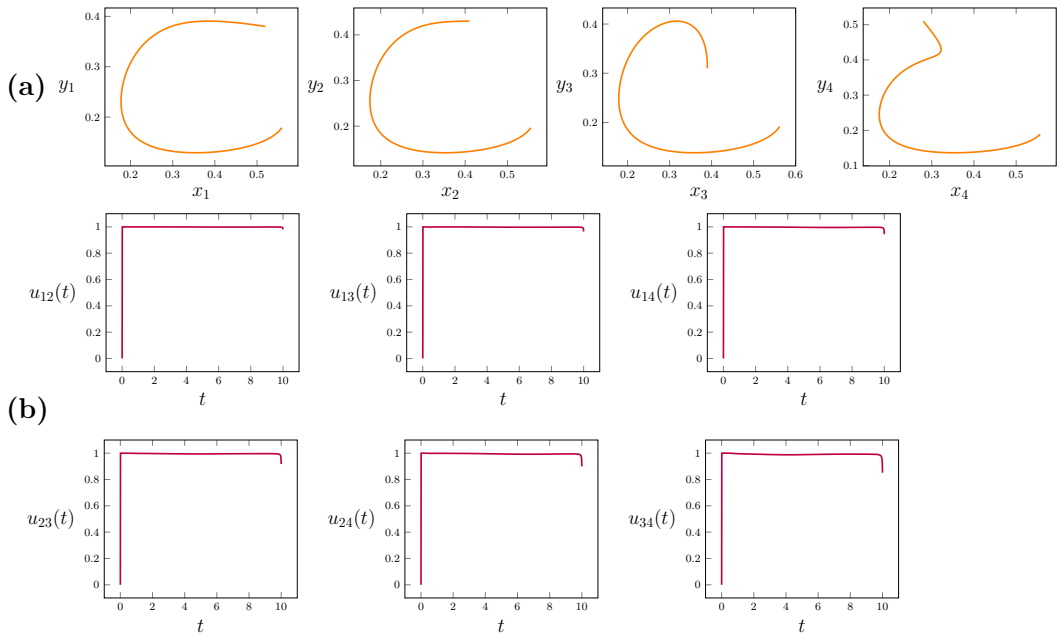


Fig. 13.4: Imposing synchronization in a four nodes network with a control associated with the functionals (13.22), (13.23) can kill oscillations or limit cycles. **(a)** Phase portraits showing the local dynamics (x_i, y_i) on each patch i of the network $(1 \leq i \leq 4)$. **(b)** Time series of the control functions u_{ij} between each pair (i, j) of patches $(1 \leq i, j \leq j,\, i \neq j)$.

(OCP) given by

$$\min_{X,u} \sum_{k=0}^{k^*} \int_{T+k\phi}^{T+(k+1)\phi} \sum_{i=1}^{n} \Big[\left(x_i(t) - \zeta(t)\right)^2 + \left(y_i(t) - \xi(t)\right)^2 \Big]\, dt\,,$$

such that

$$\begin{cases} \dot{x}_i = r_i x_i(1 - x_i) - \dfrac{c_i x_i y_i}{\alpha_i + x_i} - \displaystyle\sum_{j \in \mathcal{N}_i} u_{i,j}(x_i - x_j), \\ \dot{y}_i = -d_i y_i + \dfrac{c_i x_i y_i}{\alpha_i + x_i} - \displaystyle\sum_{j \in \mathcal{N}_i} u_{i,j}(y_i - y_j), \end{cases} \tag{OCP}$$

with $X^*(0) = X_0^* \in (\mathbb{R}^+)^{2n}$ and $u_{i,j}(\cdot) \in \Omega\,.$

The existence of solutions for the (OCP) is ensured by classical sufficient conditions, see e.g. [34] and references cited therein.

Let us now apply the well known first order necessary optimality condition given by the Pontryagin maximum principle [30] to the (OCP) problem. In what follows, we write (x_i, y_i) for $(x_i(t), y_i(t)) \in (\mathbb{R}^+)^{2n}$, $u_{i,j}$ for $u_{i,j}(t) \in \Omega$ and $p = (p_{1,i}, p_{2,i})$ for $(p_{1,i}(t), p_{2,i}(t)) : [0, T] \to (\mathbb{R}^+)^{2n}$,

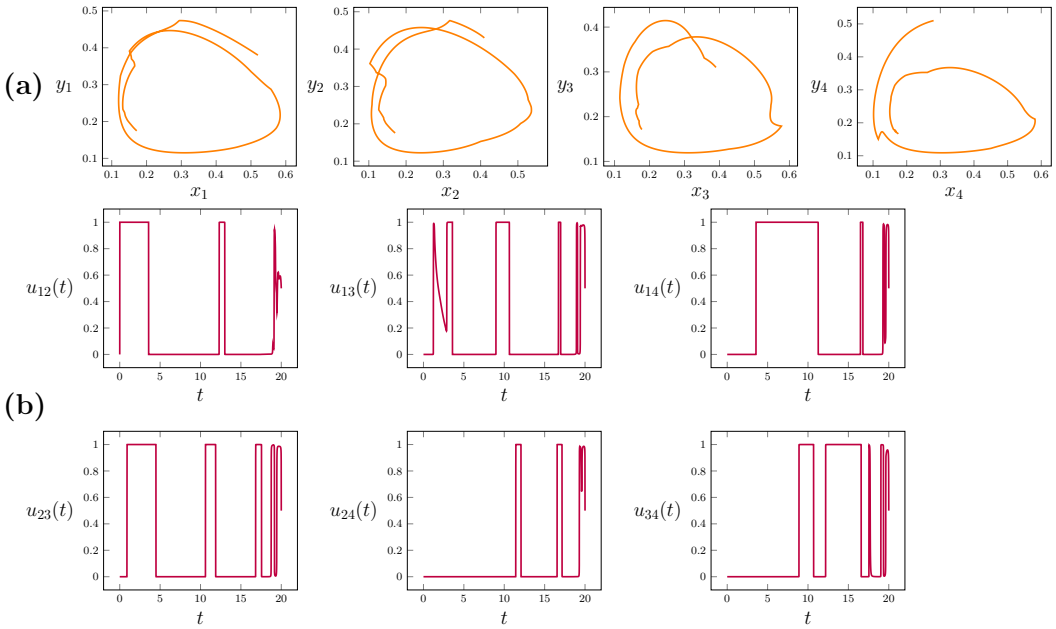


Fig. 13.5: Attempt to reach synchronization of a given limit cycle in a four nodes network. **(a)** Phase portraits showing the local dynamics (x_i, y_i) on each patch i of the network $(1 \leq i \leq 4)$. **(b)** Time series of the control functions u_{ij} between each pair (i, j) of patches $(1 \leq i, j \leq j,\ i \neq j)$.

with $t \in [0, T]$ and $1 \leq i \leq n$. According to the Pontryagin maximum principle, if $u^*_{i,j} \in \Omega$ is optimal for (OCP), then there exists a nontrivial absolutely continuous mapping p^*, the adjoint vector, such that

$$\dot{x}^*_i = \frac{\partial H}{\partial p^*_{1,i}}\,, \quad \dot{y}^*_i = \frac{\partial H}{\partial p^*_{2,i}}\,,$$

and

$$\dot{p^*}_{1,i} = -\frac{\partial H}{\partial x^*_i} \quad \text{and} \quad \dot{p^*}_{2,i} = -\frac{\partial H}{\partial y^*_i} \text{ for } 1 \leq i \leq n\,,$$

where the normalized *Hamiltonian* H is given by

$$\begin{aligned} H\left(x_i, y_i, p_{1,i}, p_{2,i}, u_{i,j}\right) = &-\sum_{i=1}^{n}\Big[\left(x_i(t)-\zeta(t)\right)^2 + \left(y_i(t)-\xi(t)\right)^2\Big] \\ &+\sum_{i=1}^{n} p_{1,i}(t)\left(f_1\left(x_i, y_i, \lambda_i\right) + g_1\left(x_i, X, u_{i,j}\right)\right) \\ &+\sum_{i=1}^{n} p_{2,i}(t)\left(f_2\left(x_i, y_i, \lambda_i\right) + g_2\left(y_i, X, u_{i,j}\right)\right)\,, \end{aligned} \tag{13.25}$$

for $1 \leq i \leq n$ and $j \in \mathcal{N}_i$. The minimization condition is given by

$$H\left(x_i^*, y_i^*, p_{1,i}^*, p_{2,i}^*, u_{i,j}^*\right) = \min_{u_{i,j}\in\Omega} H\left(x_i^*, y_i^*, p_{1,i}^*, p_{2,i}^*, u_{i,j}\right) ,$$

holds almost everywhere on $[0,T]$. Moreover, the *transversality conditions* $\left(p_{1,i}^*(T), p_{2,i}^*(T)\right) = (0,0)$, hold, with $1 \le i \le n$.

The minimizing controls $u_{i,j}^*$ are determined by the *switching functions*

$$\phi_{i,j} = \frac{\partial H}{\partial u_{i,j}} \,, \text{for } 1 \le i \le n \text{ and } j \in \mathscr{N}_i$$

and the *control law*

$$u_{i,j}^*(t) = \begin{cases} 0 & \text{if} \quad \phi_{i,j}(t) > 0 \,, \\ u_{max} & \text{if} \quad \phi_{i,j}(t) < 0 \,, \\ \text{singular} & \text{if} \quad \phi_{i,j}(t) = 0 \quad \forall t \in I_s \subset [0,T] \,, \end{cases} \tag{13.26}$$

for $1 \le i \le n$ and $j \in \mathscr{N}_i$.

If the switching functions $\phi_{i,j}$, $1 \le i \le n$, $j \in \mathscr{N}_i$, do not vanish on any subinterval I of $[0,T]$, then the extremal controls $u_{i,j}^*$ are *bang-bang* on I. The zeros of $\phi_{i,j}$ on I (possibly in infinite number), $0 < \tau_1^* < \ldots < \tau_s^* \ldots$, are called the *switching times*.

Moroever, the *strict bang-bang Legendre condition*, can be applied to the (OCP) problem, that is,

$$\dot{\phi}_{i,j}(\tau_l^*) = \frac{d}{dt}\phi(t)|_{t=\tau_l^*} \neq 0 \,, \tag{13.27}$$

for every switching time.

Next, we consider an optimal control synchronization problem of the type (OCP) with 4 nodes.

13.4.2 Example: optimal control synchronization with 4 nodes

Let γ be limit cycle, that is, $\{\gamma(t) = (\zeta(t), \xi(t))\}_{0\le t\le\phi}$, with period ϕ.

Consider the control system with 4 nodes, again associated with the complete graph shown in Figure 13.1. The equations of the controlled network are given by

$$\begin{cases} \dot{x}_1 = r_1 x_1(1-x_1) - \frac{c_1 x_1 y_1}{\alpha_1 + x_1} - \sum_{j\neq 1} u_{1,j}(x_1 - x_j) \\ \dot{y}_1 = -d_1 y_1 + \frac{c_1 x_1 y_1}{\alpha_1 + x_1} - \sum_{j\neq 1} u_{2,j}(y_1 - y_j) \\ \dot{x}_2 = r_2 x_2(1-x_2) - \frac{c_2 x_2 y_2}{\alpha_2 + x_2} - \sum_{j\neq 2} u_{2,j}(x_2 - x_j) \\ \dot{y}_2 = -d_2 y_2 + \frac{c_2 x_2 y_2}{\alpha_2 + x_2} - \sum_{j\neq 2} u_{1,j}(y_2 - y_j) \\ \dot{x}_3 = r_3 x_3(1-x_3) - \frac{c_3 x_3 y_3}{\alpha_3 + x_3} - \sum_{j\neq 3} u_{3,j}(x_3 - x_j) \\ \dot{y}_3 = -d_3 y_3 + \frac{c_3 x_3 y_3}{\alpha_3 + x_3} - \sum_{j\neq 3} u_{3,j}(y_3 - y_j) \\ \dot{x}_4 = r_4 x_4(1-x_4) - \frac{c_4 x_4 y_4}{\alpha_4 + x_4} - \sum_{j\neq 4} u_{4,j}(x_4 - x_j) \\ \dot{y}_4 = -d_4 y_4 + \frac{c_4 x_4 y_4}{\alpha_4 + x_4} - \sum_{j\neq 4} u_{4,j}(y_4 - y_j) \end{cases} \tag{13.28}$$

and the cost functional

$$J\left(x_i, y_i\right) = \sum_{k=0}^{k^*} \int_{T+k\phi}^{T+(k+1)\phi} \sum_{i=1}^{4} \left[\left(x_i(t) - \zeta(t)\right)^2 + \left(y_i(t) - \xi(t)\right)^2 \right] dt, \tag{13.29}$$

subject to the initial conditions $X(0) = (x_1(0), \ldots, x_4(0)) \in \mathscr{N}(\gamma)$, and control constraints

$$u_{min} \leq u_{i,j}(t) \leq u_{max}\,, \quad t \in [T, T + m\phi],$$

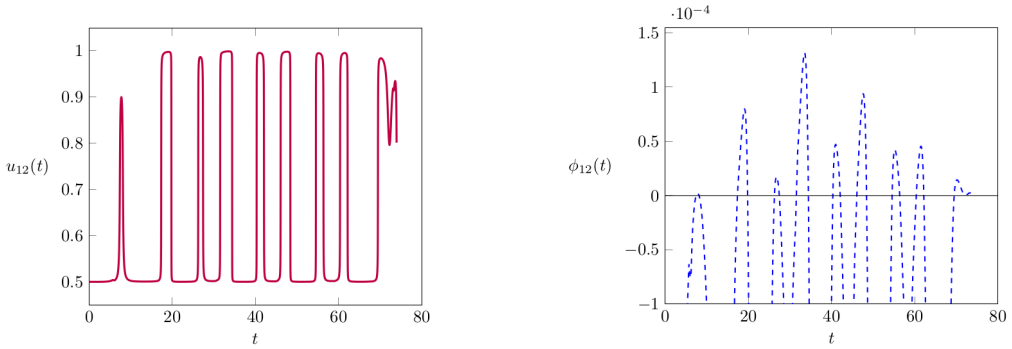
where m is the number of periods and u_{min} satisfies $u_{min} \geq K$, with K a positive threshold, that guarantees the near synchronization in the case of a constant coupling strength σ. The normalized Hamiltonian is given by

$$\begin{aligned}
H = &- \sum_{i=1}^{4} \left[\left(x_i(t) - \zeta(t)\right)^2 + \left(y_i(t) - \xi(t)\right)^2 \right] \\
&+ p_{1,1} \left(r_1 x_1 (1 - x_1) - \frac{c_1 x_1 y_1}{\alpha_1 + x_1} - u_{1,2}(x_1 - x_2) - u_{1,3}(x_1 - x_3) - u_{1,4}(x_1 - x_4) \right) \\
&+ p_{2,1} \left(-d_1 y_1 + \frac{c_1 x_1 y_1}{\alpha_1 + x_1} - u_{1,2}(y_1 - y_2) - u_{1,3}(y_1 - y_3) - u_{1,4}(y_1 - y_4) \right) \\
&+ p_{1,2} \left(r_2 x_2 (1 - x_2) - \frac{c_2 x_2 y_2}{\alpha_2 + x_2} - u_{1,2}(x_2 - x_1) - u_{2,3}(x_2 - x_3) - u_{2,4}(x_2 - x_4) \right) \\
&+ p_{2,2} \left(-d_2 y_2 + \frac{c_2 x_2 y_2}{\alpha_2 + x_2} - u_{1,2}(y_2 - y_1) - u_{2,3}(y_2 - y_3) - u_{2,4}(y_2 - y_4) \right) \\
&+ p_{1,3} \left(r_3 x_3 (1 - x_3) - \frac{c_3 x_3 y_3}{\alpha_3 + x_3} - u_{1,3}(x_3 - x_1) - u_{2,3}(x_3 - x_2) - u_{3,4}(x_3 - x_4) \right) \\
&+ p_{2,3} \left(-d_3 y_3 + \frac{c_3 x_3 y_3}{\alpha_3 + x_3} - u_{1,3}(y_3 - y_1) - u_{2,3}(y_3 - y_2) - u_{3,4}(y_3 - y_4) \right) \\
&+ p_{1,4} \left(r_4 x_4 (1 - x_4) - \frac{c_4 x_4 y_4}{\alpha_4 + x_4} - u_{1,4}(x_4 - x_1) - u_{2,4}(x_4 - x_2) - u_{3,4}(x_4 - x_3) \right) \\
&+ p_{2,4} \left(-d_4 y_4 + \frac{c_4 x_4 y_4}{\alpha_4 + x_4} - u_{1,4}(y_4 - y_1) - u_{2,4}(y_4 - x_2) - u_{3,4}(y_4 - y_3) \right) ,
\end{aligned}$$

and the switching functions are given by

$$\begin{aligned}
\phi_{1,2} &= \frac{\partial H}{\partial u_{12}} = -p_{1,1}(x_1 - x_2) - p_{2,1}(y_1 - y_2) - p_{1,2}(x_2 - x_1) - p_{2,2}(y_2 - y_1)\,, \\
\phi_{1,3} &= \frac{\partial H}{\partial u_{13}} = -p_{1,1}(x_1 - x_3) - p_{2,1}(y_1 - y_3) - p_{1,3}(x_3 - x_1) - p_{2,3}(y_3 - y_1)\,, \\
\phi_{1,4} &= \frac{\partial H}{\partial u_{14}} = -p_{1,1}(x_1 - x_4) - p_{2,1}(y_1 - y_4) - p_{1,4}(x_4 - x_1) - p_{2,4}(y_4 - y_1)\,, \\
\phi_{2,3} &= \frac{\partial H}{\partial u_{23}} = -p_{1,2}(x_2 - x_3) - p_{2,2}(y_2 - y_3) - p_{1,3}(x_3 - x_2) - p_{2,3}(y_3 - y_2)\,, \\
\phi_{2,4} &= \frac{\partial H}{\partial u_{24}} = -p_{1,2}(x_2 - x_4) - p_{2,2}(y_2 - y_4) - p_{1,4}(x_4 - x_2) - p_{2,4}(y_4 - x_2)\,, \\
\phi_{3,4} &= \frac{\partial H}{\partial u_{34}} = -p_{1,3}(x_3 - x_4) - p_{2,3}(y_3 - y_4) - p_{1,4}(x_4 - x_3) - p_{2,4}(y_4 - y_3)\,.
\end{aligned}$$

In Figure 13.6 we observe that the control $u_{1,2}$ and the corresponding switching function $\phi_{1,2}$ satisfy the control law (13.26) and the strict bang-bang Legendre condition (13.27). Analogously, the other controls also satisfy these optimality conditions, but for simplicity we do not provide a figure for the others five controls.



(a) Control $u_{1,2}$.

(b) Switching function $\phi_{1,2}$.

Fig. 13.6: The control $u_{1,2}$ and the switching function $\phi_{1,2}$ satisfy (13.26) and (13.27).

Next, we show in Figure 13.7 the dynamics of the solution to the controlled four nodes network (13.28)-(13.29). We have computed the solution $\Big((x_i, y_i)_{1\leq i\leq 4}, (u_{i,j})_{1\leq i\neq j\leq 4}\Big)$ until the final time $T + m\phi$ with $T = 6.5$, $m = 5$ and $\phi = 13.5$. It is interesting to note that the numbers of switching times of the controls are distinct. Namely, u_{12} reaches 9 times its maximum value, whereas u_{23} does only 5 times. We mainly observe that oscillations are maintained under the action of bang-bang controls. Overall, our goal to synchronize the local dynamics while preserving oscillations is reached.

13.5 Conclusion and Future Work

In this chapter, we considered a controlled complex network of Lotka-Volterra systems, where the strength of the migrations of biological individuals between the patches is replaced by control functions, reproducing the implementation of ecological corridors in a fragmented environment. We assumed that the ecological dynamics are non-identical within the fragmented environment and proved near-synchronization sufficient conditions for the solution of the controlled complex network.

After, we study optimal control problems where the main goal is the minimization of the default of synchronization in the complex network. We consider different cost functionals taking into account that the dynamics of the controlled complex network ensure the conservation of both species, namely, our goal is to impose synchronization or synchronization of limit cycles. Therefore, the solutions of the optimal control problems lead to a restoration of the biodiversity of life species in a heterogeneous habitat by reaching at least a global coexistence equilibrium, or in a better scenario, a global limit cycle which would guarantee biological oscillations, which means rich life dynamics.

In a future work, we aim to enlarge our study of controlled synchronization or near-synchronization in complex networks of nonlinear dynamical systems. First, it is natural to ask

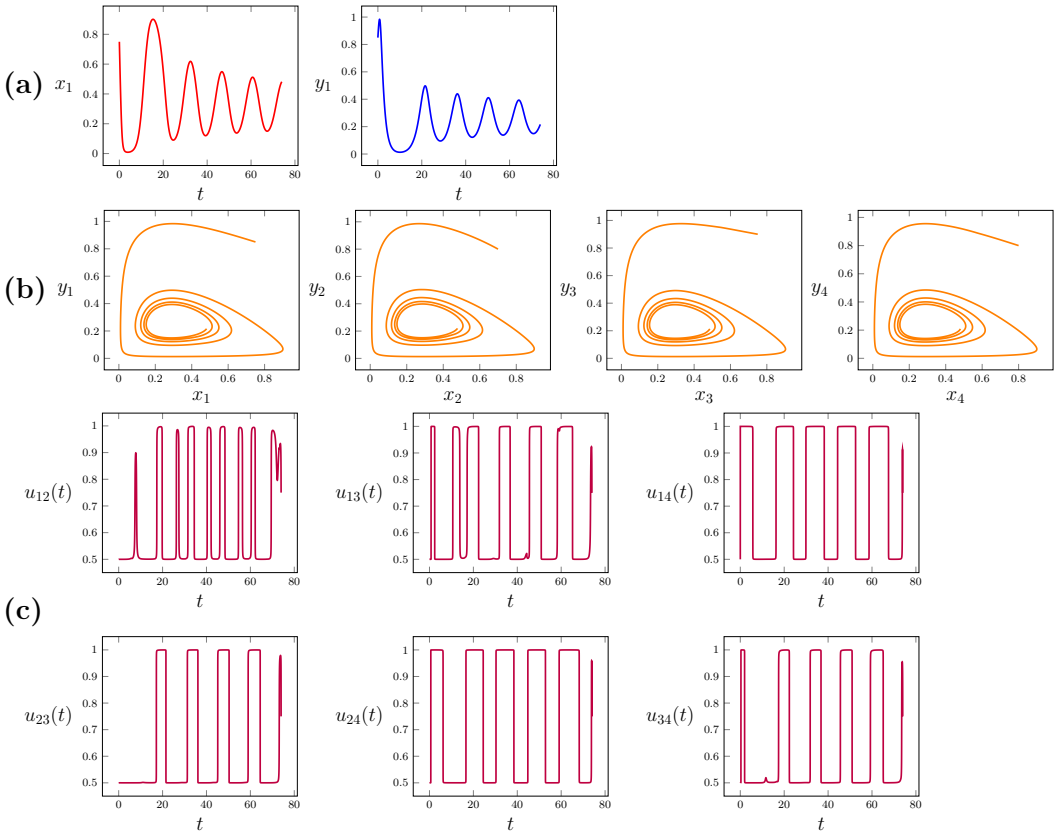


Fig. 13.7: Synchronization towards oscillations of the controlled four nodes network (13.28)–(13.29). **(a)** Time series showing the evolution of x_1, y_1 on the first patch. **(b)** Phase portraits showing the local dynamics (x_i, y_i) on each patch i of the network $(1 \leq i \leq 4)$. **(c)** Time series of the control functions u_{ij} between each pair (i, j) of patches $(1 \leq i, j \leq j,\ i \neq j)$.

if the possibility to near-synchronize oscillations in finite-dimensional systems can be generalized to infinite dimensional systems, such as reaction-diffusion systems, which are likely to admit bifurcations of periodic solutions (see for instance [26] for a study of oscillatory solutions in a spatial Holling-Tanner reaction-diffusion system). Next, another exciting perspective would be to investigate the optimal control of synchronization of chaotic systems, since it is known that such systems can be synchronized by constant couplings (see notably [1]). Hence, we believe that optimal control of synchronization in complex networks of nonlinear dynamical systems will produce original results in a near future.

Acknowledgments

This work is partially supported by Portuguese funds through CIDMA, The Center for Research and Development in Mathematics and Applications of University of Aveiro, and the Portuguese Foundation for Science and Technology (FCT – Fundação para a Ciência e a Tecnologia), within project UIDB/04106/2020 (https://doi.org/10.54499/UIDB/04106/2020) and by the project "Mathematical Modelling of Multi-scale Control Systems: applications to human diseases (CoSysM3)", 2022.03091.PTDC, financially supported by national funds (OE), through FCT/MCTES.

References

1. Aziz-Alaoui, M.A.: Synchronization of Chaos. Encyclopedia of Mathematical Physics, Elsevier, Vol. 5, pp : 213-226, (2006).
2. Aziz-Alaoui, M.A. and Cantin, Guillaume and Thorel, Alexandre Synchronization of Turing patterns in complex networks of reaction-diffusion systems set in distinct domains. To be published in Nonlinearity (2024).
3. Arenas, A., Díaz-Guilera, A., Kurths, J., Moreno, Y. and Zhou, C.: Synchronization in complex networks. Physics reports, 469 (3), 93–153 (2008)
4. Apreutesei, N. and Dimitriu, G.: Optimal control for Lotka–Volterra systems with a hunter population. In: International Conference on Large-Scale Scientific Computing, 277–284 (2008)
5. Barahona, M. and Pecora, L. M: Synchronization in small-world systems. Physical review letters, 89 (5), 054101 (2002)
6. Bayen, T., Mairet, F., Martinon, P. and Sebbah, M.: Analysis of a periodic optimal control problem connected to microalgae anaerobic digestion. Optimal Control Applications and Methods, 36 (6), 750-773 (2015)
7. Bazykin, A. D.: Nonlinear dynamics of interacting populations. World Scientific, (1998)
8. Belykh, I., Hasler, M., Lauret, M. and Nijmeijer, H.: Synchronization and graph topology. International Journal of Bifurcation and Chaos, 15 (11), 3423–3433 (2005)
9. Bernard, S., César, T., Nuiro, S. P. and Piétrus, A.: Unexistence of limit cycle in an optimal control problem of a population of diabetics. Revista de Matemática: Teoría y Aplicaciones, 25 (2), 239–259 (2018)
10. Biscani, F. and Izzo, D. Heyoka. Optimal Control of the Lotka-Volterra equations. `https://bluescarni.github.io/heyoka.py/notebooks/Optimal%20Control%20of%20the%20Lotka-Volterra%20equations.html`, Accessed in 2023.
11. Bonnard, B. and Rouot, J.: Optimal Control of the Controlled Lotka-Volterra Equations with Applications - The Permanent Case, SIAM Journal on Applied Dynamical Systems, 2023, 22 (4), 2761–2791.
12. Bonnard, B., Rouot, J. and Silva, C. J.: Geometric optimal control of the Generalized Lotka-Volterra model of the intestinal microbiome. Accepted in OCAM (2023)
13. Cantin, G., Verdière, N. and Lanza, V.: Synchronization under control in complex networks for a panic model. In: International Conference on Computational Science, 262–275 (2019)
14. Cantin, G. and Verdière, N.: Mathematical modeling and optimal control of complex epidemiological networks. In: Complex Systems, Smart Territories and Mobility, 169–186, Springer (2021)
15. Cantin, G. and Aziz-Alaoui, M. A.: Dimension estimate of attractors for complex networks of reaction-diffusion systems applied to an ecological model. Communications on Pure & Applied Analysis, 20 (2), 623 (2021)
16. Cantin, G. and Silva, C. J.: Complex network near-synchronization for non-identical predator-prey systems, AIMS Mathematics, 7 (11), 19975–19997 (2022)

17. Crespo, L. and Sun, J.: Optimal control of populations of competing species. Nonlinear Dyn., 27, 197–210 (2002)
18. El-Gohary, A. and Yassen, M. T.: Optimal control and synchronization of Lotka–Volterra model. Chaos, Solitons & Fractals, 12 (11), 2087–2093, (2001)
19. Faria, J. R.: Limit cycles in an optimal control problem of diabetes. Applied Mathematics Letters, 16 (1), 127–130 (2003)
20. Fribourg, L. and Soulat, R.: Limit cycles of controlled switched systems: Existence, stability, sensitivity. Journal of Physics: Conference Series, 464, 012007 (2013)
21. Holling, C. S.: The functional response of predators to prey density and its role in mimicry and population regulation, The Memoirs of the Entomological Society of Canada, 97 (S45), 5–60, Cambridge University Press (1965)
22. Ibañez, A. and Zuazua, E.: Optimal control and turnpike properties of the Lotka–Volterra model. Master Thesis, Universidad del País Vasco/Euskal Herriko Unibertsitatea (2013)
23. Ibañez, A.: Optimal control of the Lotka–Volterra system: turnpike property and numerical simulations, Journal of Biological Dynamics, 11 (1), 25–41 (2017)
24. Kuznetsov, Y. A., Kuznetsov, I. A. and Kuznetsov, Y.: Elements of applied bifurcation theory, 112, Springer (1998)
25. Ledzewicz, U. and Schättler, H.: On the role of the objective in the optimization of compartmental models for biomedical therapies. Journal of Optimization Theory and Applications, 187, 305–335 (2020)
26. Li, X. and Jiang, W. and Shi, J.: Hopf bifurcation and Turing instability in the reaction–diffusion Holling–Tanner predator–prey model. The IMA Journal of Applied Mathematics, 78(2), 287-306 (2013).
27. Markus, L.: Optimal control of limit cycles or what control theory can do to cure a heart attack or to cause one. In: Symposium on Ordinary Differential Equations, Springer, 108–134 (1973)
28. May, R. M.: Stability and complexity in model ecosystems, 1, Princeton University Press (2019)
29. Miranville, A., Cantin, G. and Aziz-Alaoui, M. A.: Bifurcations and Synchronization in Networks of Unstable Reaction–Diffusion Systems, Journal of Nonlinear Science, 31 (2), 1–34 (2021)
30. Pontryagin, L., Boltyanskii, V., Gramkrelidze, R. and Mischenko, E.: The Mathematical Theory of Optimal Processes, Wiley Interscience (1962)
31. Real, L. A.: The kinetics of functional response. The American Naturalist, 111 (978), 289–300 (1977)
32. Sager, S., Bock, H. G., Diehl, M., Reinelt, G. and Schloder, J. P.: Numerical Methods for Optimal Control with Binary Control Functions Applied to a Lotka-Volterra Type Fishing Problem. In: Seeger, A. (eds) Recent Advances in Optimization. Lecture Notes in Economics and Mathematical Systems, 563, Springer, Berlin, Heidelberg, 269–289 (2006)
33. Skalski, G. T. and Gilliam, J. F.: Functional responses with predator interference: viable alternatives to the Holling type II model, Ecology, 82 (11), 3083–3092 (2001)
34. Trélat, E.: Contôle optimal, théorie & applications, Mathématiques Concrètes, Vuibert, Paris, 2005.
35. Trélat, E. and Zuazua, E.: The turnpike property in finite-dimensional nonlinear optimal control, J. Differential Equations 258, 81–114 (2015).
36. Trélat, E., Zhang, C. and Zuazua, E.: Steady-state and periodic exponential turnpike property for optimal control problems in Hilbert spaces, SIAM J. Control Optim. 56 (2), 1222–1252 (2018).
37. Trélat, E.: Linear turnpike theorem, Math. Control Signals Systems, 35 (3), 685–739 (2023)
38. Wirl, F.: Cyclical strategies in two-dimensional optimal control models: necessary conditions and existence, Annals of Operations Research, 37 (1), 345–356 (1992)
39. Yosida, S.: An optimal control problem of the prey–predator system, Funck. Ekvacioj, 25, 283–293 (1982)
40. Dynamic Optimization. Lotka Volterra Fishing Optimization.`https://apmonitor.com/do/index.php/Main/LotkaVolterra`, Accessed in 2023.

14

On Quantitative Approaches to Model and Control Biomedical Systems

Sean T. McQuade[1], Christopher Denaro[1], and Benedetto Piccoli[1 2]

[1] Center for Computational and Integrative Biology, Rutgers–Camden email
[2] Department of Mathematical Sciences, Rutgers–Camden email

Summary. The growing availability of large omics datasets has provided a great opportunity to address difficulties in drug development for many different disease areas. Several quantitative approaches to analyze these datasets rely on encoding biological networks as mathematical graphs, thus pointing to an increased need to develop and categorize mathematical methods specifically geared towards networks describing biological systems. In this review, we categorize network-based approaches useful to answer questions posed in the drug development field, with an emphasis on the mathematical background of each of them. This categorization includes compartmental systems, Laplacian dynamics, zero-deficiency theory, cooperative and monotone systems, flux balance analysis, linear-in-flux-expressions and others. Despite such extensive history of the application of graph theory to network biology, the mathematical tools dealing with hypergraphs are not well-developed for specific application to biological networks. After motivating the necessity of dealing with hypergraphs to correctly represent complex bio-chemical reactions and drug effects, some results for the linear-in-flux-expressions method and general challenges for a network flow theory for hypergraphs are provided. These approaches may support drug development for various diseases, such as Tuberculosis and Parkinson's.

14.1 Introduction

Mathematical models and computational approaches have been used in biology and medical sciences for a long time. However, only in the recent past did the official role of such methods start to be fully recognized. For instance, the US Food and Drug Administration is now accepting computer simulations, based on mathematical models, as a valid tool for decision-making on medical devices. Figure 14.1, reproduced from [1] with permission from the authors, illustrates the four main methods for scientific evidence in such decisions: animal models, clinical trials, bench experiments, and computer simulations. The latter gives enormous opportunities for quantitative scientists, such as mathematicians, physicists, and engineers, to contribute to biomedical sciences. On the other side, there are many challenges creating obstacles to the applicability of such methods to real systems. It is beyond the scope to discuss all such challenges, especially those related to data paucity and other forms of uncertainty, but we rather focus on the specific problem of designing models and methods that are feasible to tackle the naturally high dimension of such systems.

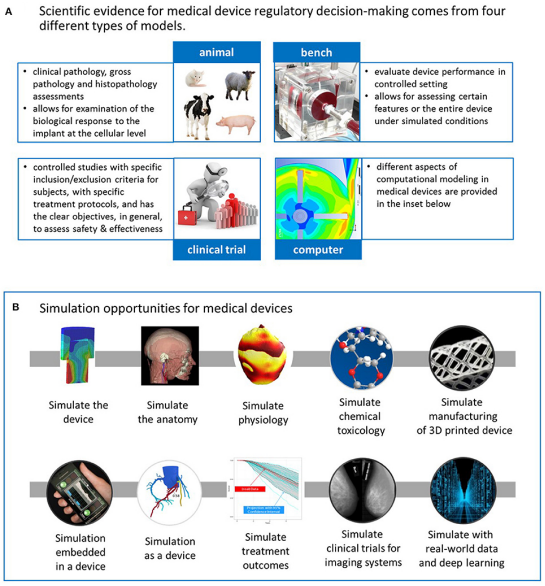


Fig. 14.1: A. Regulatory decisions about medical devices are made with evidence collected from four different models: animal, bench, computational, and human (i.e., clinical trials). B. Simulation opportunities for QSP models. The upper row indicates applications that are used to design and evaluate a medical device, the lower indicates other processes, such as within a device or simulation itself (figure (with permission) from Morrison, Tina M., et al.). "Advancing regulatory science with computational modeling for medical devices at the FDA's office of science and engineering laboratories." Frontiers in medicine 5 (2018): 241.

14.1.1 How big is a real system

As mentioned, one of the biggest challenges to the use of mathematical approaches is the large size of the systems to be modeled. Technological advancements gave rise to large-scale data from the sequencing of genomes and other "-omes" datasets, such as proteome, transcriptome, metabolome, epigenome, etc. System-level data started to appear as soon as half a century ago, with many attempts to create consortia, data repositories, and maps to exhaustively model cells, tissues, and organs, all the way to the whole human body. In general, it is difficult to tackle generalistic projects, but few attempts have been made in the last twenty years. The multiscale nature of biological systems calls for such efforts, but the enormous complexity often compromises the potential success of these projects. Examples of large-scale attempts include the BioModels, Physiome Project [2], and the Drug Disease Modeling Resources Consortium [3, 4], with the last two apparently not fully active. A number of projects focused on developing maps of biological networks, such as the Cell Atlas Project [5] and the Virtual Metabolic Human [6]. There are also efforts in mapping diseases, for example, the Parkinson's Disease (PD) Map developed by the University of Luxembourg and the Systems Biology Institute of Tokyo, which provides a network of genes and molecules that describes the pathophysiology of PD [7]. Just to mention some figures, the Human BioMolecular Atlas Program (HubMAP) Portal, which is part of the Cell Atlas Project, mentions 57 reference organs with 1,588 anatomical structures, while the Virtual Metabolic Human main map includes

more than 5,000 metabolites and close to 20,000 reactions. The PD Map includes roughly 2000 nodes corresponding to molecules, complexes, and ions over roughly 400 different compartments. At risk of pushing such considerations a bit too far, we may say that while the big unifying projects in physics hit the boundary of impossible experiments, the big unifying projects in these areas of biology hit the boundary of unmanageable amounts of experiments and data.

14.1.2 Quantitative systems pharmacology

A wealth of mathematical methods and approaches have been employed using techniques from different areas, such as graph theory, differential equations, boolean networks, agent-based models, and many others. One of our main focuses will be on Quantitative Systems Pharmacology (briefly QSP), a research area where there is an increasing interest in the use of such methods [8, 9]. The term Quantitative Systems Pharmacology gained substantial recognition in a 2011 National Institutes of Health (NIH) white paper, which was a collaborative effort between academia, industry, and government. The white paper describes the combination of experimental and quantitative approaches [10]. The field of QSP was born by recognizing the need for a field that uses both pharmacology and systems biology for application to drug discovery and development. More precisely, QSP was described as "... an approach to translational medicine that combines computational and experimental methods to elucidate, validate and apply new pharmacological concepts to the development and use of small molecule and biologic drugs" [10]. QSP provides an alternative to the traditional approach of "one drug, one target, one pathway" by offering network-focused approaches to facilitate the understanding of drug mechanism of action on cellular pathways and the impact on disease pathophysiology.

Mechanistic modeling of biological processes utilizes techniques from both mathematical and biological sciences [8, 9, 11, 12]. Mechanistic models aim to reproduce biological processes with enough accuracy to be relevant to an applied problem. As large biological data becomes more available, the field of mechanistic modeling becomes more feasible [13, 14, 15, 16, 17]. One of the main draws to mechanistic modeling in the pharmaceutical industry is its cheap predictive ability during translational and pre-clinical steps. The use of mechanistic models to aid in the development of pharmaceutical therapies is referred to as Model Informed Drug Development (MIDD). As mentioned previously, MIDD is an increasing field, with approval of the use of mechanistic models looking promising [18]. MIDD has a plethora of uses, especially for diseases that lack robust biomarkers for target-drug engagement. Mechanistic models can make it possible to identify potential biomarkers for a therapy-alleviating the issue of therapies failing clinical trials due to lack of target engagement. For example, there are no known disease-modifying therapies for Parkinson's Disease, and there have been trial therapies that have failed clinical trials due to the lack of biomarkers [19]. Mechanistic models can also help predict which treatments may be best for additional translational and pre-clinical testing. For example, treatment of Tuberculosis can involve treatment with four antibiotics over the course of four months [20, 21, 22]. Testing all possible four-drug combinations (out of a pool of 20 antibiotics) is unfeasible in a pre-clinical setting; a mechanistic model may provide input into which treatment combinations may work the best, quickly trimming the necessary number of pre-clinical tests. These example applications of mechanistic modeling highlight the ability of MIDD to de-risk clinical trials by alleviating issues faced in translational and pre-clinical trials.

14.1.3 An example of Quantitative Systems Pharmacology focused on tuberculosis treatment

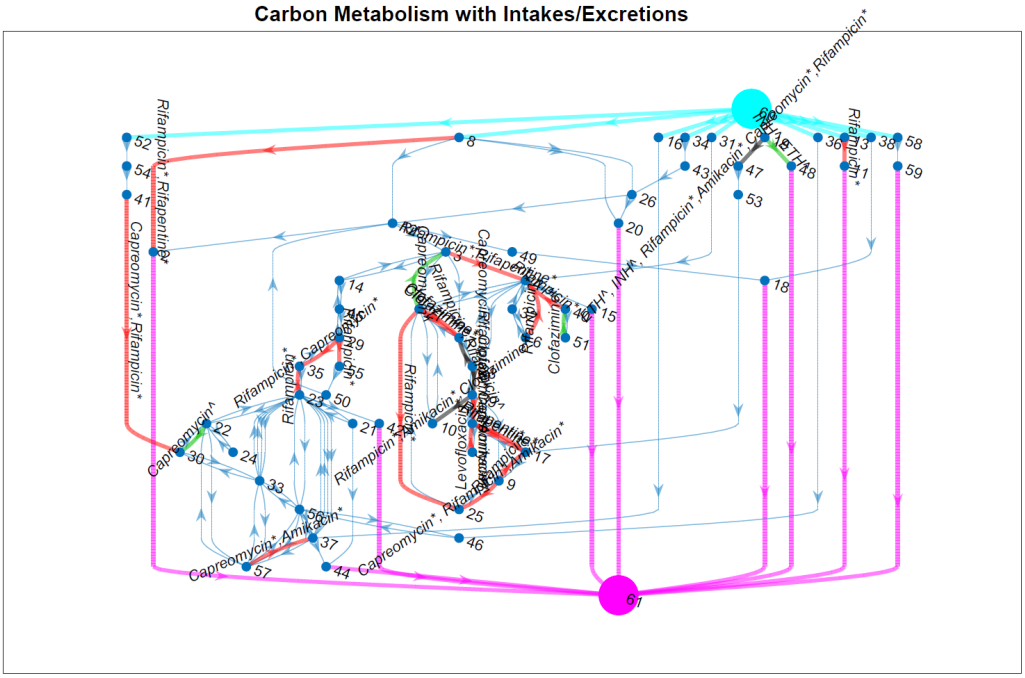


Fig. 14.2: Simulation of the effect of a drug treatment on the Carbon Metabolism of Mycobacterium Tuberculosis. Edges labeled by antibiotics that significantly affect them, with red color for downregulation and green color for upregulation. Cyan edges represent intakes, and magenta edges excretions [23], which were generated synthetically to complete the network.

Beyond the common application of graph theory, one can build metabolic networks using generalized graphs that capture relationships between reactions involving more than two metabolites, i.e. *hypergraphs*, or the enhancer inhibitor effects of drugs on given reactions, i.e. *ubergraphs*. With these tools, metabolic networks can be encoded in the following way: nodes correspond to metabolites while chemical reactions between metabolites correspond to edges - note that by using hyperedges, we are able to represent chemical reactions with multiple substrates or products faithfully. Additionally, the inclusion of uberedges, which are connections from nodes to other edges, allows for a systematic labeling of edges (reactions) that are affected when treated with a drug. Figure 14.2 demonstrates this application to the Central Carbon Metabolism network in Mycobacterium Tuberculosis; nodes correspond to metabolites while edges correspond to chemical reactions between metabolites. The underlying metabolic network information is sourced from the Kyoto Encyclopedia of Genes and Genomes (KEGG), while the drug-gene expression data is sourced from the Gene Expression Omnibus (GEO) [24, 25, 26, 27]. Note that the edges are colored depending on whether or not the enzyme facilitating the corresponding reaction was up- or down-regulated after treatment

with an antibiotic; red edges indicate down-regulation, green edges indicate up-regulation, and blue edges indicate no differential regulation. Furthermore, it is possible to simulate the flow of mass through this metabolic network under the effects of drug treatment according to the differential equation: $\dot{x} = S(x) \cdot f$, where S is defined as in equation (14.7). An important step in enabling the non-trivial flow of mass through the metabolic network is network "completion." In order to "complete" the network, a "virtual" source node is added, leading into all nodes with no incoming edges. It is also necessary to add a "virtual" sink node to all nodes without outgoing edges. This step ensures that there is no trivial dissipation/accumulation of mass in the system. Figure 14.2 includes this step of completing the network, where the virtual source and sink nodes are represented in cyan and magenta, respectively.

14.1.4 The role of artificial intelligence

There has also been an increase in big data in the systems biology field which makes establishing model priors more readily available. Note that mechanistic models contrast artificial intelligence (AI) models as they are typically built with reference to known biological systems - contrasting the "black box" behavior of AI models. In the past, there was not enough curated data to train such models; deep learning models are becoming an option as big data becomes available to researchers.

New Machine learning and AI methods have been developed that can assist in building QSP models and simulations. Given very large data sets, it can be difficult to select which data should inform a model. AI may be used to find literature and mine databases to inform the QSP models and combine it with Clinical studies. For some models, AI may be leveraged to construct those parts of the model for which the underlying mechanism is not well known, such as when there is a lack of clinical data regarding particular metabolic pathways, or incomplete data from a population of interest. It has been proposed that AI models can help extrapolate data sets or aggregate data from many studies to fill in gaps.

AI can also specify the qualitative characteristics of a metabolic network. These qualities are valuable to guide researchers in constructing QSP models. In [28], the authors show a proof of concept that deep learning models can determine the existence of an equilibrium on a metabolic network, and the metabolite levels at equilibrium. These and other machine learning techniques are expected to become integrated with mechanistic models.

14.2 Systems Biology Models for Metabolic Networks.

In this section, we illustrate some of the most popular approaches to the modeling of biological networks, with special focus on metabolic ones, based on systems biology ideas.
To fix basic notation, we define a network to be a directed graph $G = (V, E)$, with vertices $V = \{v_1, \ldots, v_N\}$ that correspond to state variables (metabolites) $\{x_{v_1}, \ldots, x_{v_n}\} = X(t)$. The latter changes over time at a fast time scale. Biological networks are usually not isolated, thus we will represent inflows by adding a *virtual vertex* v_0, which will act as a source, and outflows by adding a second virtual vertex v_{n+1}, which acts as a sink. We set $\bar{V} = V \cup \{v_0, v_{n+1}\}$, i.e. the set of nodes including the virtual ones. The edges $E \subset \bar{V} \times \bar{V}$ are directed and have weights given by numbers $f_e = f_{(v_i,v_j)} \in \mathbb{R}_+$, representing the fluxes of the respective edge, which in turn model a biochemical reaction. Most of the time the fluxes evolve at a much slower scale w.r.t. metabolites; thus the numbers f_e are assumed to be constant.

Given a vertex v, an edge $e = (v, w)$ represents a dynamics (in the linear case) of the type $\dot{x}_v = -f_e x$, $\dot{x}_w = f_e x$. Therefore, the dynamics of the whole system can be written as a system of ordinary differential equations:

$$\dot{x} = S(f) \cdot x \tag{14.1}$$

with x the vector of metabolites, f the vector of fluxes, and S a matrix depending on fluxes. Many approaches were proposed to deal with such systems, oftentimes exploiting the connection between (14.1) and the underlying graph G. Methods focusing on the formulation (14.1) include compartmental systems based on control of networked systems [29, 30, 31, 32], Laplacian dynamics using the weighted Laplacian of the graph [33, 34, 35, 36], Zero Deficiency Theory linking graph structural properties to the existence and stability of equilibria [37, 38], and cooperative and monotone systems [39, 40]. On the other side, the well-established Flux Balance Analysis (briefly FBA, see [41, 42]) is based on writing the dynamics as:

$$\dot{x} = S \cdot f, \tag{14.2}$$

where $S = \{S_{ve}\}_{v \in V, e \in E}$ is the stoichiometric matrix, $f = f_e$ is the vector of fluxes, thus equilibria are given by kernel of S, see [42]. In this paper, we will discuss these approaches and others, including network flows, Markov chains focusing on stochastic processes on a directed graph [43], and network motifs [44]. Before doing that, we introduce some more notation. G has a *terminal component*, means there exists a set of vertices $T \subset V$ such that there is no directed path from $v_i \in T$ to v_{n+1}. A weakly connected component of G is a set of vertices $V_1 \subseteq V$ such that every vertex $v_i \in V_1$ has an undirected path to every other vertex $v_j \in V_1$ i.e. not necessarily moving along each edge in the direction the edge indicates.

14.2.1 Compartmental systems

A compartmental system is defined as a collection of compartments C_i, $i = 1, \ldots, n$, each with a given substance, whose total amount x_i varies in time. It is assumed that there is mass exchange between the compartments and the outside world. More precisely a_{ij} is the flow from compartment C_i to C_j, u_i is the input from outside the systems, and w_i the output to outside the system. Then one can write the dynamics as:

$$\dot{x}_i = \sum_j a_{ij} + u_i - \sum_j a_{ji} - w_i \tag{14.3}$$

where $a_{ij} = a_{ij}(x_j)$ and $w_i = w_i(x_i)$ are all assumed smooth (continuously differentiable) and positive. In simple terms, the flows are positive and those from a compartment C_i depend only on the amount x_i. It is easy to see that, starting from initial data with positive components, solutions will remain positive if $a_{i,j}(0) = 0$ and $w_i(0) = 0$. If the coefficients are monotonic, these systems happen to be nonoscillatory, i.e. they don't exhibit periodic solutions (see also Section 14.2.4). Moreover, the existence and stability of equilibria are linked to the topological property of the underlying network. More precisely, one defines a network adding edges for nonvanishing coefficients, then the existence of equilibria is guaranteed if every compartment connected upstream to the outside world is also connected downstream to the outside world. We omit details since results will be stated in section 14.3. This was one of the first developed approaches, see [32], linking the network topology to properties of the dynamics.

14.2.2 Laplacian Dynamics

Laplacian dynamics were introduced to the mathematical biology literature by [36], where the authors study the equilibria and convergence of chemical reaction networks. Laplacian dynamics were studied on a graph as defined above G excepting no self-loops. The edge labels f_e are the kinetic reaction rates corresponding to the edges of the graph. These systems have the following structure:

$$\dot{X} = \mathscr{L}(G) \cdot X$$

$\mathscr{L}(G)$ is the Laplacian matrix of G, and the state variables $X = (x_1, \ldots, x_n)^T$ is a column vector of concentrations, with x_i being the concentration of compound represented by $v_i \in G$.

$$\mathscr{L}(G)_{ij} = \begin{cases} e_{ij} & \text{if } i \neq j, \\ -\sum_{v \neq j} e_{vj} & \text{if } i = j. \end{cases}$$

This is a framework built for analyzing networks, and many central results of molecular biology can be derived from it, such as mechanisms in gene regulatory networks. The inherent linear framework replaces nonlinear dynamics and simple rate constants for linear dynamics with more complex labels. The purpose here is to facilitate the calculation of steady states. We recount a proposition from [36].

Proposition 1. *For any graph G call the corresponding laplacian matrix L. If λ is an eigenvalue of L then either $\lambda = 0$ or $Re(\lambda) \leq 0$*

14.2.3 Zero-deficiency theory

In this section, we summarize the zero-deficiency theory and recall a main result. Zero-deficiency theory is found in the literature on chemical reaction networks, with an influential paper being [38]. The article [45] describes a chemical reaction network with m reactions and n metabolites from a zero deficiency theory standpoint. It is a directed graph $G = (V, E)$ where every vertex is adjacent to an edge, and there exists $e = (v_j, v_k) \in E : v_i = v_j$ or $v_i = v_k$ for $i = 1, \ldots, n$. Each vertex is called a *complex* and comprises one or more of ℓ different chemical species. The reactant map gives the mapping of species to complexes $\rho : \ell \to V$, the product map gives the mapping of reactions to complexes $\pi : E \to V$. In zero deficiency theory, a weakly connected component of G is called a *linkage class*. For a chemical system with n complexes, ℓ linkage classes, and m dimensions of the flux space, the deficiency is defined by the formula:

$$\delta = n - l - s.$$

A chemical network is classified by its deficiency δ which is always non-negative. A large number of chemical networks have deficiency $\delta = 0$, and zero-deficiency theory gives particular attention to this class of networks.

Proposition 2 (Zero-deficiency Theorem). *Consider a free closed chemical reaction network with mass-action kinetics. If its deficiency is zero, then: there exists an equilibrium with strictly positive entries if and only if the system is weakly reversible (i.e. each weakly connected component is also strongly connected).*

Moreover, this strictly positive equilibrium is unique in each stoichiometric class (i.e. each weakly connected component has a space of equilibria that has dimension one, so that its equilibrium is unique up to a multiplicative constant representing the total mass in the component), and it is locally asymptotically stable.

14.2.4 Cooperative and monotone systems

Consider a general system of differential equations

$$\dot{x} = f(x) \tag{14.4}$$

and let $t \mapsto x(t, x_0)$ be the solution taking initial data $x(0) = x_0$. We say that (14.4) is *monotone* if there exists a partial order $\preceq$ such that the following holds. If $x_0 \preceq y_0$ then for all $t \geq 0$ we have $x(t, x_0) \preceq x(t, y_0)$.
A particular class of monotone systems is given by cooperative systems. More precisely, (14.4) is called *cooperative* if the Jacobian $J(f)$ has non-negative off-diagonal entries:

$$\frac{\partial f_i}{\partial x_j}(x) \geq 0$$

for every $i, j \in \{1, \ldots, n\}$, $i \neq j$ and every x. The following proposition contains the main results for monotone systems.

Proposition 3. *The following holds:*
(i) If (14.4) is cooperative then it is monotone with respect to the partial order $\preceq_+$.
(ii) If (14.4) is cooperative then there exist no periodic solutions (except constants).
(iii) If (14.4) is cooperative and $J(f)(x)$ is irreducible (i.e. corresponding signed graph is connected) for every x, where $J(f)$ is the Jacobian matrix of f, then almost (in the sense of Lebesgue measure) every bounded solution approaches the set of equilibria.

While cooperative systems are arguably the most frequent case of monotone systems, other cases are of interest. An orthant is defined by: $\Omega^r = \{(x_1, \ldots, x_n) : \epsilon_i \cdot x_i \geq 0\}$, where $\epsilon_i \in \{-1, +1\}$ for $i = 1, \ldots, n$. Given an orthant Ω^r we can define a partial order $\preceq^r_\Omega$ by setting $x \preceq^r_\Omega y$ if and only if $y - x \in \Omega^r$. The differential equation (14.4) is said *monotone with respect to the orthant Ω^r* if it is monotone with respect to $\preceq^r_\Omega$. To link monotone systems w.r.t. an orthant to the topology of the underlying network, we need to introduce a piece of notation. Given a signed graph (G, ϕ), i.e. $\phi : E \to \{-1, +1\}$, $P = i_1\ i_2\ \ldots\ i_\ell\ i_{\ell+1}$, with $i_{\ell+1} = i_1$, is an *indirected closed path* if for every $j = 1, \ldots, \ell$ either $(i_j, i_{j+1}) \in E$ or $(i_{j+1}, i_j) \in E$. Moreover, the *parity* of P is given by:

$$\mathrm{Par}(P) = \mathrm{sign}(i_1; i_2)\ \mathrm{sign}(i_2; i_3) \cdots \mathrm{sign}(i_\ell; i_1), \tag{14.5}$$

where $(i_1; i_2) = (i_1, i_2)$ or $(i_1; i_2) = (i_2, i_1)$. We have the following:

Proposition 4. *The differential equation (14.4) is monotone with respect to an orthant Ω^r if and only if all indirected closed path of the associated signed graph (G, ϕ) have positive parity (i.e. equal to 1).*

14.2.5 Flux Balance Analysis

The Flux Balance Analysis approach (briefly FBA) was successfully applied to many biological networks, see [46, 47, 42]. This approach models the reactions of the network as a matrix of stoichiometric coefficients S multiplied by a vector f representing all chemical reactions. $f_e \in f$ is the flux corresponding to edge e. An FBA system models the metabolites in graph G by

$$\dot{X} = S \cdot f$$

The method identifies an allowable solution in the flux space using constraints imposed by the stoichiometric matrix, S, and capacity constraints imposed by the fluxes of the network. These constraints intersect with the positive orthant of the flux space to guarantee positive (and thus, biologically relevant) fluxes. The edges of these flux cones correspond to extreme pathways (as shown in 14.3), which are special paths connecting exchange fluxes of the network (source v_0 to sink v_{n+1}). *Extreme pathways* are a sequence of adjacent edges through a biochemical network. From [48] they have the following properties:

1. they are contiguous sets of fluxes (a flux map) that each satisfy the mass balance of the system and reaction irreversibility constraints; and 4) they can have multiple inputs or outputs;
2. they characterize time-invariant properties of biochemical networks;
3. they are a unique set of convex basis vectors that circumscribe all possible steady-state flux distributions through the network.

Non-negative combinations of extreme pathways form a positive cone in the flux space and can be used to generate all feasible flows through the network. The extreme pathways correspond to the edges of the cone see Fig 14.3 right.

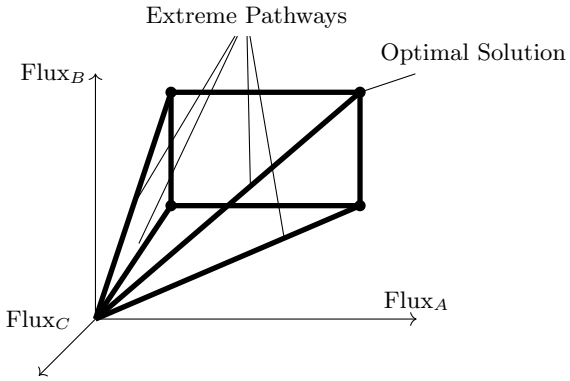


Fig. 14.3: The fluxes intersect the positive orthant to indicate a biologically feasible cone. The edges of this cone correspond to the extreme pathways in the network. The Optimal solution will lie on an edge of the feasible cone.

14.2.6 Network flows

The study of network flows concerns algorithms and methods for finding positive fluxes of a network $G = (V, E)$ (with n vertices and m edges that admit an equilibrium. Directed edges of G represent fluxes between the vertices. The maximum flow on an edge is called the "capacity" of the edge and is denoted $c(v_i, v_j)$. A flow is a mapping $f : E \to \mathbb{R}^m_+$ that satisfies $0 \leq f_{(v_i,v_j)} \leq c(v_i, v_j)$ and

$$\sum_{(v_i,v_j)\in E} f_{(v_i,v_j)} - \sum_{(v_j,v_k)\in E} f_{(v_j,v_k)} = 0$$

Network flow problems usually assume that the system is in equilibrium. This respects equation (14.2.6), which enforces *Kirchoff's first law*: the total flow entering a vertex equals the flow exiting the vertex. Network flows are less directed toward modeling the network dynamics, and instead, they are designed to describe the properties of a network. A common network flow problem is to find the maximum flow on the network, i.e. the largest amount of flow from a source to a sink.

In [49] the authors consider a network that contains one source and sink (in our case these are the virtual vertices v_0 and v_{n+1} respectively). A well-known result is given, known as the max-flow min-cut theorem, which indicates that the maximum flow through a network is equal to the minimum cut capacity. The *cut set* is a set of edges whose removal disconnects the source from the sink and the cut capacity is the sum of capacities of the *cut set*).

Proposition 5 (Max-Flow Min-Cut Theorem). *The maximum flow value obtainable in a network is the minimum capacity of all cut sets that disconnect v_0 and v_{n+1}.*

14.2.7 Markov chains

A system defined as a Markov chain on a labeled graph G interprets edge weights differently. The flow value $f_e = f_{(v_i,v_j)} \in \mathbb{R}_+$ is the probability that a random variable X which at time step t_h, with $h \in \mathbb{R}_+$, is in state v_i i.e. $X(t_h) = v_i$, transitions to a state v_j at next time step t_{h+1}, i.e. $X(t_{h+1}) = v_j$. More precisely, a homogeneous discrete-time Markov chain over a finite set of states $V = \{v_1, \dots, v_n\}$ is a stochastic process $X(t)$ taking values in V and satisfying the condition: for all times $t_0 \le t_1 \le \cdots \le t_h \le t_{h+1}$, one has

$$\Pr(X(t_{h+1}) = v_i \mid X(t_0) = v_{i_0}, X(t_1) = v_{i_1}, \dots, X(t_h) = v_{i_h}) = \\ \Pr(X(t_{h+1}) = v_i \mid X(t_h) = v_{i_h}),$$

and which is homogeneous in time:

$$\Pr(X(t_{h+1}) = v_i \mid X(t_h) = v_{i_h}) = \Pr(X(t_{h+1} - t_h) = v_{i_{h+1}} \mid X(t_0) = v_{i_h}) = \\ P_{i_h, i_{h+1}}(t_{h+1} - t_h)\,.$$

In simple words, the transition from one state to another is not dependent on the past and can be encoded in a transition matrix P. The (i, j)-th entry of P is the probability of transitioning from v_i to v_j.
The continuous-time version of a Markov chain is given by having a transition matrix $P(t)$ defined for all times t and right-differentiable. In this case, one defines the generator matrix as $Q = \dot{P}(0)$. The vector $\pi(t)$ such that $\pi_i(t) = \Pr(X(t) = i)$, is a solution of $\dot{\pi}^T = \pi^T Q$ with the initial condition $\pi(0)$.

14.2.8 Network motifs

A well-known constructive approach to complex networks is that of *network motifs*, mainly due to Uri Alon [50]. The main idea is that biological networks are composed of small units, with a few nodes, which have a precise biological function. The work is justified by noticing how some specific small graphs are recurrent in nature, from bacteria as *Escherichia coli* to fungi as *Saccharomyces cerevisiae* and all the way to mammals. The justification of this recurrence is often time performed

by comparing with random networks, such as Erdős-Renyi ones [51]. This approach was particularly successful among systems biologists but somehow did not expand into a well-defined mathematical theory. One of the reasons is biological networks may not be easily composed, as pointed out by the modularity theory of Del Vecchio [52].

14.3 Linear-In-Flux-Expressions (LIFE) Approach

Linear-in-Flux-Expression (LIFE) method. Generalizing the FBA dynamics, the following class of systems were recently introduced:

$$\frac{dx}{dt} = S(X) \cdot f, \tag{14.6}$$

where the stoichiometric matrix S is as in FBA models, but depending on X, and f is the column vector of fluxes. $S_{ve}(X)$ indicates the matrix entry corresponding to node v and edge e. This allows one to leverage the natural linearity w.r.t. fluxes of biological systems. In this setting, the network equilibria, for a fixed flux vector f, correspond to the kernel (null space) of $S(X)$ and thus depend both on the fluxes and metabolite levels, and a detailed analysis is provided in [53]. The biologically significant equilibria correspond to positive fluxes, thus to the cone given by the intersection of the positive orthant with the kernel of $S(x)$. This cone can be generated using only positive coefficients from a positive basis, whose cardinality of such basis may exceed the dimension of the ambient space. The elements of such bases are linked to the so-called *extreme pathways*, and algorithms to compute such bases for (14.6) were provided in [53]. Linear systems without intakes nor excretions are related to *continuous-time Markov chains* [43], while linear systems with intakes and excretions are known as *compartmental systems* [29, 30]. Nonlinear systems could be treated by using the results of [32], enabling the determination of existence, uniqueness, and stability of equilibria.

To include the inflow and outflow of the network, which are fundamental to guarantee the existence of equilibria for metabolic networks, we consider fluxes corresponding to edges having only terminal vertices (intakes), which have a virtual vertex v_0 acting as a source, and fluxes corresponding to edges having only initial vertices (excretions) which have a virtual vertex acting as a sink.

We use the reverse cholesterol transport network to demonstrate a simple system defined by linear in flux expressions. The network is shown in Fig. 14.4. The stoichiometric matrix $S(x)$ for Fig. 14.4 is a six by ten matrix where each column corresponds to an edge, and each row corresponds to a vertex. The first three columns of $S(x)$ correspond to exchange fluxes into the network.

$$(A) \qquad S_{ve}(X) = \begin{cases} = \text{-}F_e(x_v) & e = (v, w),\ v \in V, w \in V \cup \{v_{n+1}\} \\ = F_e(x_w) & e = (w, v),\ w \in V \\ = 1 & e = (v_0, v)\ v \in V \\ = 0 & \text{otherwise,} \end{cases} \tag{14.7}$$

where $F_e : \mathbb{R} \to \mathbb{R}_+$ is differentiable, strictly increasing, and $F_e(0) = 0$. The stoichiometric matrix of the system depicted by Fig. 14.4 is shown in (14.8); we've dropped the subscript indicating the edges since this is given by the column of the entry.

$$S_{ve}(X) =$$

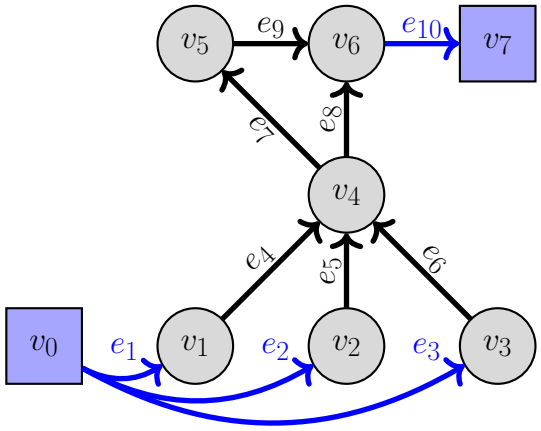


Fig. 14.4: A graph representing the reverse cholesterol transport. Each node represents a metabolite involved in human cholesterol metabolism, edges represent reactions, and two virtual vertices v_0, v_7, are defined to complete the network.

$$\begin{pmatrix} 1 & 0 & 0 & -F(x_{v_1}) & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -F(x_{v_2}) & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -F(x_{v_3}) & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & F(x_{v_1}) & F(x_{v_2}) & F(x_{v_3}) & -F(x_{v_4}) & -F(x_{v_4}) & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & F(x_{v_4}) & 0 & -F(x_{v_5}) & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & F(x_{v_4}) & F(x_{v_5}) & -F(x_{v_6}) \end{pmatrix}. \tag{14.8}$$

The important subspace of this matrix is the nullspace which comprises the space of fluxes f such that $S_{ve}(X) \cdot f = 0$ and is denoted $\mathcal{N}(S(X))$. Later in this section, we will see how to compute $\mathcal{N}(S(x)) \cap (\mathbb{R}_+)^m$ yielding the biologically relevant fluxes in this space.

14.3.1 Relationships between Markov chains and LIFE approach

The generator matrix Q is a Metzler matrix, i.e. rows sum to 0 thus the equation $\dot{\pi}^T = \pi^T Q$ corresponds to the linear LIFE system $\dot{x} = J(f)\dot{x}$, simply by taking $Q = J(f)^T$. The main difference is that in Markov chains $\sum_i \pi_i = 1$ since it represents a probability, while in the LIFE approach we may have any positive value. Notice that the vector π_i restricted to a terminal component G_i is a stationary distribution, thus an equilibrium for the LIFE dynamics. The uniqueness of the equilibrium is linked to strong connectedness, and, in particular, graphs G with a unique terminal component have a unique equilibrium of a given total mass. Otherwise stated, the kernel of the matrix $J(f)$ has dimension one.

14.4 Hyperedges in Biological Networks

In this section, we discuss the importance of including generalized classes of graphs to model biological systems. Most biochemical reactions involve multiple reactants and multiple products. Moreover,

often times enzymes are necessary for the reaction to occur. Finally, proteins and other molecules may act as enhancers or inhibitors, and the same is true for drugs or their cascade effects. Such complexity is difficult to capture by representing reactions as simple edges, as in standard graphs. Figure 14.5 provides an illustration of how complex reactions, respectively enhancer/inhibitor actions, can be represented by hyperedges, respectively uberedges. A directed hyperedge is, in simple words, an edge with multiple initial nodes and terminal ones. Figure 14.5 (left) includes weights for every node, given by α_i, $i = 1, \ldots, 5$, and for the hyperedge itself, given by h. The former decodes the stoichiometry of the given reaction, while the latter is the number of reactions occurring simultaneously, i.e. the flow through the hyperedge. Figure 14.5 (right) includes a sign ($\pm$) to distinguish the type of action (enhancer/inhibitor) and a weight u_i, $i = 1, 2$, for each uberedge representing the level at which the action occurs.

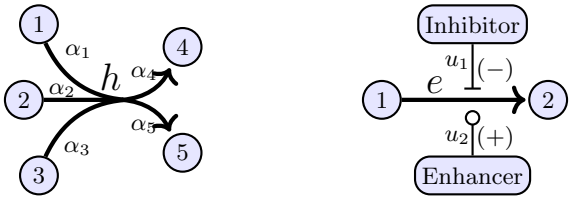


Fig. 14.5: Left: A hyperedge h connecting three reactants to two products. Each reactant and product has a weight corresponding to the stoichiometry of the reaction. Right: An enhancer(inhibitor) molecule promotes(inhibits) a chemical reaction. The edges u_1, u_2 are called "uberedges" and connect an enzyme or drug to an edge e.

14.4.1 A key example: central carbon metabolism

Figure 14.6 represents a section of the central carbon metabolism. The main function of the network is carbon metabolism from lipid and sugar catabolism: glucose is a product of catabolism and the cell uses this energy to synthesize enzymes for the pentose phosphate pathway and provide ribose 5-P for nucleotide synthesis. Glycolysis yields metabolites phosphoenolpyruvate (PEP), pyruvate, and acetyl-CoA; see the KEGG database [26] for details.
This is an example of a metabolic pathway well conserved among different species. Simple directed edges model most reactions; however, the complete dynamics involves several hyperedges. The symbol $h_{1,2}$ indicates two directed hyperedges h_1 and h_2: each connecting three metabolites. (the first having two reactants and the second having two products). h_1 represents the production of isocitrate from oxaloacetate and acetyl CoA, and h_2 production of acetyl CoA and oxaloacetate from isocitrate. The hyperedge h_3 models acetyl CoA reacting with glyoxylate to produce malate. To illustrate the importance of representing reactions as hyperedges, we focused on the two hyperedges h_1 and h_2, and the four metabolites Pyruvate, Acetyl CoA, Isocitrate, and Oxaloacetate. We assume the other metabolites are constant, thus the other reactions act as sources or sinks. In Figure 14.7, we compare the evolution of the system obtained by representing the hyperedges as a collection of simple edges with that of the system using true hyperedges.

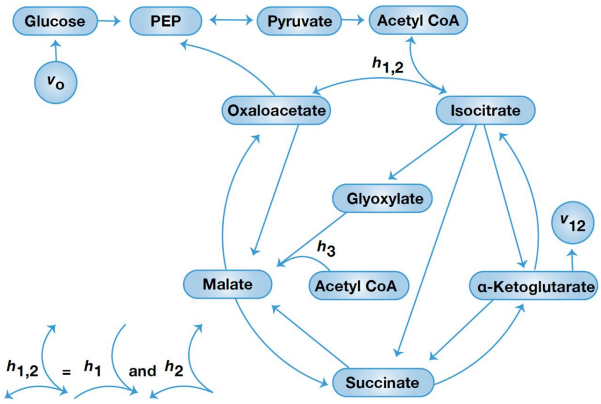


Fig. 14.6: The central carbon metabolism network. The edges labeled $h_{1,2}$ represent two different hyperedges, and h_3 is a third hyperedge connecting three metabolites.

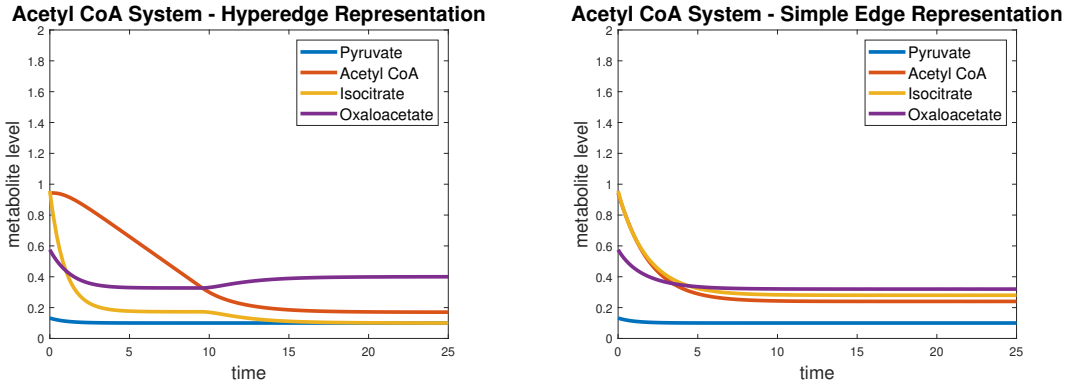


Fig. 14.7: Simulations for the hyperedge system (left) and the simple edge system (right). The metabolites pyruvate, isocitrate, and oxaloacetate exhibit an "intermediate equilibrium" before reaching a true equilibrium in the hypergraph system.

14.4.2 Hyperedges in LIFE dynamics

Directed graphs are commonly used to model metabolic networks, with nodes representing metabolites and edges representing the biochemical reactions. A hypergraph is a more general structure where an edge can connect more than two nodes thus more faithfully depicting these chemical reactions.

For the following definitions, let $V = \{v_1, \ldots, v_n\}$ be a set of nodes and let $\mathscr{P}(V)$ be the power set of V. A *hyperedge* h is a subset of nodes, i.e. $h \in \mathscr{P}(V) \setminus \{\emptyset\}$, representing a biochemical reaction involving multiple reactants and products. A *directed hyperedge* is an ordered pair $h = (Y, Z)$ with $Y \in \mathscr{P}(V) \cup \{v_0\}$, $Z \in \mathscr{P}(V \cup \{v_{n+1}\})$. The set Y represents incoming nodes, while Z outgoing ones, thus elements of Y (Z) are called initial vertices (terminal vertices). for the hyperedge h. Moreover, using $|\cdot|$ to indicate the cardinality of a set, we define $d_{\text{in}}(h) = |Y|$, i.e. the incoming

degree, and $d_{\text{out}}(h) = |Z|$, i.e. the outgoing degree. The set of directed hyperedges is denoted $\mathscr{H}$. Each hyperedge needs to be supplemented by node weights representing the stoichiometry of the reaction. Thus an hyperedge $h = (Y, Z)$ is weighted defining the map $\Psi_h : h \mapsto (\Psi_h^{\text{out}}, \Psi_h^{\text{in}})$ where $\Psi_h^{\text{out}} : Y \mapsto \mathbb{R}_+$ and $\Psi_h^{\text{in}} : Z \mapsto \mathbb{R}_+$.
To represent drug actions, we need to add a definition of uberedges, which connect vertices to directed hyperedges. More precisely, An *e/i-uberedge* is a couple $u = (v, h)$ with $v \in V, h \in \mathscr{H}$ and the set of ubereges is indicated by $\mathscr{U}$.

We are now ready to state a general assumptions for LIFE dynamic on ubergraphs:

(B) For every hyperedge $h = (Y, Z)$ and vertex $v \in X$, we have:

$$S_{vh}(X) = \begin{cases} -\alpha_v \mathbf{F}_h(X)\mathbf{K}_h(X) & v \in X \\ \alpha_v \mathbf{F}_h(X)\mathbf{K}_h(X) & v \in Y \\ 1 & X = \{v_0\}, v \in Y, \\ 0 & \text{otherwise,} \end{cases} \tag{14.9}$$

with $\alpha_w = \Psi_h^{\text{in}}(w)$ if $w \in Y$ and $\alpha_w = \Psi_h^{\text{out}}(w)$ if $w \in Z$, $\mathbf{F}_h : \mathbb{R}^{d_{in}(h)} \to \mathbb{R}_+$ is given by

$$\mathbf{F}_h(X) = \min_{w \in Y} \left\{ F_{w,h}(x_w) \frac{1}{\alpha_w} \right\}, \tag{14.10}$$

$F_{w,h} : \mathbb{R}_+ \to \mathbb{R}_+$ (flow of metabolite x_w due to reaction h), and

$$\mathbf{K}_h = \prod_{w \in U_h} K_{(w,h)}(x_w), \tag{14.11}$$

where $K_{(w,h)} : \mathbb{R}_+ \to \mathbb{R}_+$ and U_h the set of vertices w such that there exists e/i-uberedge $(w, h) \in \mathscr{U}$, with the convention that $K_h = 1$ if $U_h = \emptyset$.

Similarly to (A), under assumption (B) each function $F_{w,h}$ depends only on the metabolite x_w, but there is the additional factor $\mathbf{K}$ which corresponds to the action of one or more e/i-uberedges. This gives a nonlocal dependence, with respect to network topology, because the vertex(vertices) corresponding to an enhancer(s) or inhibitor(s) may be anywhere in the network not necessarily close to the edge it is affecting.
For a LIFE system to be in equilibrium, the fluxes must be in the nullspace of the stoichiometric matrix, $\mathscr{N}(S(X))$. We also noted that finding a positive basis for the flux space was necessary for application as it represents the biologically relevant part of the space. This space, $\mathscr{N}(S(x)) \cap (\mathbb{R}_+)^m$, is represented by a cone in the positive orthant. Finding a basis for this cone is more difficult, and in general, there are more basis vectors than there are dimensions in the space. Here we summarize a method to build extreme pathways described by [54]. The full description of the method (Algorithm 1) can be found in [53] which includes an example (Example 3.1). To construct the positive basis, start by considering the equation $V_0 \cdot S(x)^T = C_0$, and the solution given by $V_0 = I_m$, the $m \times m$ identity matrix, $S(x)$ the stoichiometric matrix and $C_0 = S(X)^T$. At each iteration new matrices $V_i \in M_{n_i \times m}$, $C_i \in M_{n_i \times n}$ are defined which satisfy $V_i \cdot S(X)^T = C_i$.

Proposition 6. *The extreme pathways of $S(X)$ form a positive basis of $\mathscr{N}(S(X)) \cap (\mathbb{R}_+)^m$.*

The proof is given in [53].

14.5 Flows on Weighted Hypergraphs

Definition 1. *A flow on a hypergraph with stoichiometric coefficients $G = (V, \mathscr{H}, \Psi_{\mathscr{H}})$ is a function $g : \mathscr{H} \to \mathbb{R}_+$ such that g satisfies Kirchhoff's law for metabolic graphs, i.e. for every node v*

$$\sum_{h \in \Gamma^{out}(v)} \psi_h^{out}(v) \cdot g(h) = \sum_{k \in \Gamma^{in}(v)} \psi_k^{in}(v) \cdot g(k) \tag{14.12}$$

where $\Gamma^{out}(v) = \{(X, Y) \in \mathscr{H} : v \in X\}$ and $\Gamma^{in}(v) = \{(W, Z) \in \mathscr{H} : v \in Z\}$.

With (14.12), we can now examine the characteristics of potential flows on hypergraphs with stoichiometric coefficients. This section will demonstrate some interesting examples of how hyperedges within hypergraphs can impose non-local constraints on the values of the flow. We hope that this work will elucidate the nontrivial difficulties when attempting to apply well-established results in graph theory, such as the Maximum-Flow Minimum-Cut Theorem 5, to hypergraphs.

We will begin by examining a small hypergraph with stoichiometric coefficients that is given in Figure 14.8 such that there are no flows that satisfy Kirchhoff's law for hypergraphs due to the choice of stoichiometric coefficients. Examining in more detail, we see that Kirchhoff's law for nodes v_1, v_2 implies different values for the flows assigned to h_1, depending on which node is considered, leading to no valid flows existing.

Although Figure 14.8 demonstrates a weighted hypergraph with no valid flows, the choice of stoichiometric coefficients was decided without respect to an underlying chemical reaction. In short, the choice of stoichiometric coefficients in Figure 14.8 may not conserve the mass of the chemical species across the reaction - this example may allude that additional restrictions on the choice of stoichiometric coefficients/branch weights within the hypergraph need to be considered for the Max-Flow Min-Cut Theorem to hold.

This problem of flows over hypergraphs, and subsequently the Max-Flow Min-Cut Theorem, has been studied in various settings. A method for converting undirected hypergraphs with a single weight on each edge to corresponding "flow networks" has been studied in [55]. Lawler provides a method of converting un-directed, weighted hypergraphs to directed, weighted graphs in a manner that preserves a one-to-one correspondence of cutsets between the two networks; importantly, no definition or correspondence of flow is provided in the hypergraph setting. Cambini provides additional work in describing flows over hypergraphs in [56]. In a setting with specific types of hyperedges, notably with hypertrees, Cambini provides a definition for a flow that includes a Kirchhoff-like law that incorporates a "demand" vector on each node - alleviating the occurrence of an over-determined system. Work has also been done in generalizing the Max-Flow Min-Cut Theorem by Hoffman in [57]. Hoffman provides a method of loosening the definition of a cut to a weight requirement instead; however, no example is provided for a hypergraph. Despite the success in defining a flow for weighted hypertrees, a concise, generalizable definition for a flow on a hypergraph is elusive.

A more general description of networks that do not admit feasible flows can be derived from the matrix form of Kirchhoff's law for each node. As an illustrative example, we provide the matrix encoding Kirchhoff's law for each node using notation for the stoichiometric edge weights for Figure 14.9.

The matrix that encodes Kirchhoff's law is generated as follows: for each node within the network, excluding source and sink, generate a row of the matrix. For each edge within the hypergraph, generate a column. Each column should include the stoichiometric coefficients in the corresponding row. Similarly, each row should contain the stoichiometric coefficients of the node in the corresponding hyperedge. The matrix in equation 14.13 encodes Kirchhoff's law for any flow on the weighted

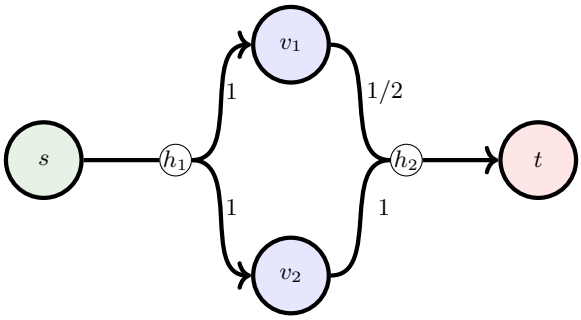


Fig. 14.8: An example of a hypergraph with stoichiometric coefficients such that no possible flow function $g : \mathscr{H} \to \mathbb{R}_+$ exists due to the choice of stoichiometric coefficients and the topology of the network.

hypergraph in Figure 14.9.
In general, the matrix that encodes Kirchhoff's law for flows on a weighted hypergraph is, $M \in \mathscr{M}_{v\times e}(\mathbb{R})$, where v, e are the number of nodes and edges, respectively. For weighted hypergraphs that do not admit feasible flows, the matrix equation will have only a trivial nullspace.

$$\left(\begin{array}{c|ccccc} [l|ccccc] & h_0 & h_1 & h_2 & h_3 & h_4 \\ \hline v_0 & \psi^{in}_{h_0}(v_0) & -\psi^{out}_{h_1}(v_0) & 0 & 0 & 0 \\ v_1 & 0 & \psi^{in}_{h_1}(v_1) & -\psi^{out}_{h_2}(v_1) & 0 & 0 \\ v_2 & 0 & \psi^{in}_{h_1}(v_2) & -\psi^{out}_{h_2}(v_2) & -\psi^{out}_{h_3}(v_2) & 0 \\ v_3 & 0 & 0 & \psi^{in}_{h_2}(v_3) & -\psi^{out}_{h_3}(v_3) & 0 \\ v_4 & 0 & 0 & 0 & \psi^{in}_{h_3}(v_4) & -\psi^{out}_{h_4}(v_4) \end{array}\right) \begin{pmatrix} g(h_0) \\ g(h_1) \\ g(h_2) \\ g(h_3) \\ g(h_4) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad (14.13)$$

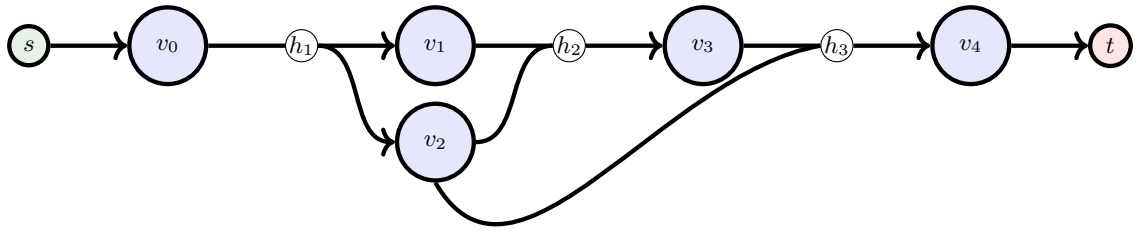


Fig. 14.9: An example of a weighted hypergraph (weights excluded) with five nodes and five edges. Any flow on this hypergraph must satisfy Kirchhoff's law for hypergraphs which can be encoded as in equation 14.13.

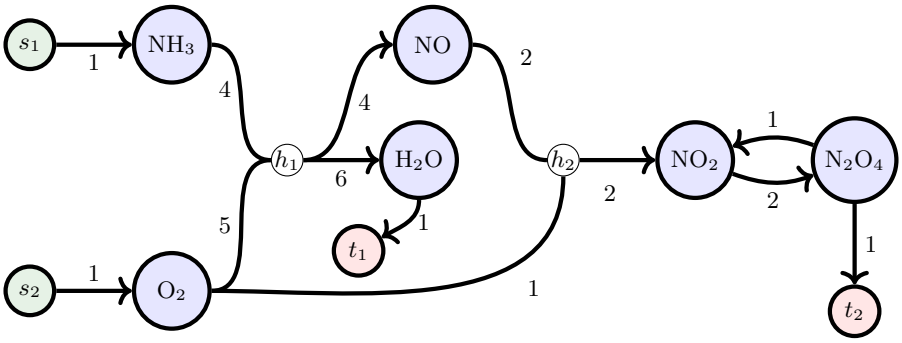


Fig. 14.10: Example of reaction without equilibrium flow: Nitric oxide is oxidized, and then nitrogen dioxide is dimerized into dinitrogen tetroxide. h_1 and h_2 are two hyperedges representing reactions.

14.5.1 Hypergraph Motifs

As introduced previously, hypergraphs provide an avenue to represent many real-world systems, from complex chemical networks to vast communication/interaction networks. Classifying local and non-local characteristics of hypergraphs in these settings can provide insight into the underlying system.

In [58], the authors present a method to classify hyperedges called motifs. These are groups of three hyperedges in a graph that are given a class based on which of the three edges have an empty intersection. This method can be expanded to include more than three hyperedges and will be sufficient for graphs with a large number of vertices. This serves as the "building blocks" for hyperedges and are used to construct a profile for large hypergraphs. The profile describes the local structures in the hypergraph.

The authors indicate some desirable properties of this method of classifying hyperedges:

- Exhaustive: the hyperedge motifs include all possible types of intersections among three hyperedges.
- Unique: the intersections of hyperedges correspond to only one motif.
- Size independent: the motif classification is independent of the number of vertices in the hyperedge.

In, [58], hyperedges are unoriented sets of vertices. They notate a hypergraph $\mathscr{H} = (V, \mathscr{E})$ as having V vertices, and a set of hyperedges $\mathscr{E} = \{e_1, e_2, \ldots, e_{|\mathscr{E}|}\}$ where $|\mathscr{E}|$ is the number of hyperedges in $\mathscr{H}$. We propose a similar system of classifying hyperedges, but with the addition of weights to the specific class of hyperedges based on their intersection.

14.6 Conclusion

In this paper, we presented some general considerations on the use of mathematical models for biomedical systems. Thanks to recent recognition from the FDA, such models can be used in real medical applications as an alternative to classical clinical trials and animal models. This opportunity faces multiple challenges ranging from the high dimensionality of the considered systems to the

paucity of data for specific diseases. Despite these difficulties, an increasing interest manifested itself, especially in the area of Quantitative System Pharmacology, which combines systems biology tools with more classical pharmacodynamics. Moreover, as -omics data become available an increasing role for Artificial Intelligence is on the way.

After revising multiple mathematical approaches, we focused on a recent method called Linear-In-Flux Expression (briefly LIFE), which allows a systematic representation of complex metabolic networks, including biochemical reactions with multiple compounds and drug effects. Those features require the use of generalized graphs, called, respectively, hypergraphs and ubergraphs. These new mathematical frameworks call for the development of new results stemming from classical graph theory. We provide some initial examples of difficulties in establishing a network flow theory for hypergraphs.

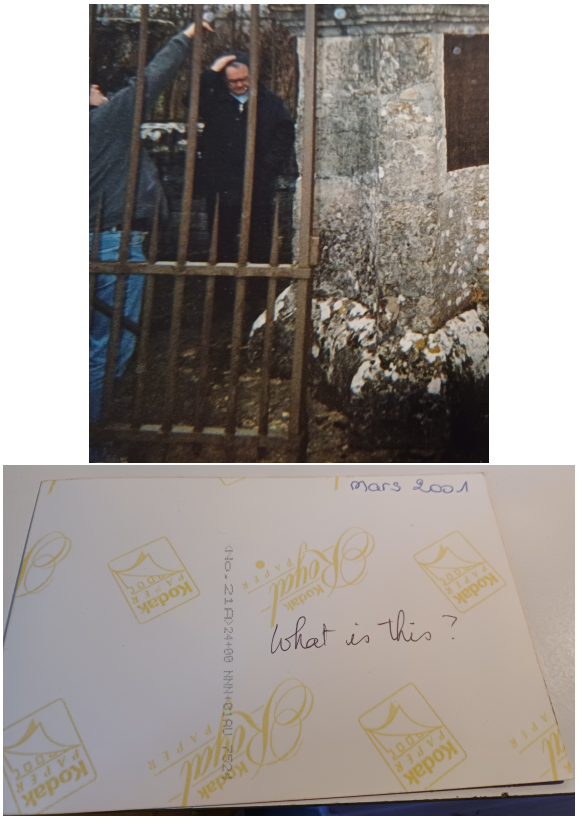
References

1. Morrison, T., Pathmanathan, P., Adwan, M., Margerrison, E.: Advancing regulatory science with computational modeling for medical devices at the FDA's office of science and engineering laboratories. Frontiers in Medicine, Frontiers Media, Vol. 5, pp : 241, (2018)
2. Hunter, P.J., Borg, T.K.: Integration from proteins to organs: the Physiome Project. Nature reviews molecular cell biology, Nature Reviews, Vol. 4, pp : 237-243, (2003)
3. Le Novere, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., Snoep, J., Hucka, M.: BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic Acids Research, Oxford University Press, Vol. 34, pp : D689-D691, (2006)
4. Sarwar, D.M., Kalbasi, R., Gennari, J.H., Carlson, B.E., Neal, M.L., de Bono, B., Atalag, K., Hunters, P.J., Nickerson, D.P.: Model annotation and discovery with the Physiome Model Repository. BMC Bioinformatics, BioMed Central, Vol. 20, pp : , (2019)
5. Regev, Aviv and Teichmann, Sarah A. and Lander, Eric S. and Amt, Ido and Benoist, Christophe and Birney, Ewen and Bodenmiller, Bernd and Campbell, Peter and Carninci, Piero and Clatworthy, Menna and Clevers, Hans and Deplancke, Bart and Dunham, Ian and Eberwine, James and Elis, Roland and Enard, Wolfgang and Farmer, Andrew and Fugger, Lars and Gottgens, Berthold and Hacohen, Nir and Haniffa, Muzlifah and Hemberg, Martin and Kim, Seung and Klenerman, Paul and Kriegstein, Arnold and Lein, E. D. and Linnarsson, Sten and Lundberg, Emma and Lundeberg, Jaokim and Majumder, Partha and Marioni, John C. and Merad, Miriam and Mhlanga, Musa and Nawijin, Martijn and Netea, Mihai and Nolan, Garry and Pe'er, Dana and Phillipakis, Anthony and Ponting, Chris P. and Quake, Stephen and Reik, Wolf and Rozenblatt-Rosen, Orit and Sanes, Joshua and Satija, Rahul and Schumacher, Ton N. and Shalek, Alex and Shapiro, Ehud and Sharma, Padmanee and Shin, Jay W. and Stegle, Oliver and Stratton, Michael and Stubbington, Michel J. T. and Theis, Fabian J. and Uhlen, Matthias and Van Oudenaarden, Alexander and Wagner, Allon and Watt, Fiona and Weissman, Jonathan and Wold, Barbara and Xavier, Ramnik and Yosef, Nir and Human cell atlas meeting: The Human Cell Atlas. ELIFE, eLife Sciences, Vol. 6, pp : , (2017)
6. Brunk, Elizabeth and Sahoo, Swagatika and Zielinski, Daniel C. and Altunkaya, Ali and Drager, Andreas and Mih, Nathan and Gatto, Francesco and Nilsson, Avlant and Gonzalez, German Andres Preciat and Aurich, Maike Kathrin and Prlic, Andreas and Sastry, Anand and Danielsdottir, Anna D. and Heinken, Almut and Noronha, Alberto and Rose, Peter W. and Burley, Stephen K. and Fleming, Ronan M. T. and Nielsen, Jens and Thiele, Ines and Palsson, Bernhard O.: Recon3D enables a three-dimensional view of gene variation in human metabolism. Nature Biotechnology, Nature Portfolio, Vol. 36, pp : 272 - , (2018)

7. Fujita, Kazuhiro A. and Ostaszewski, Marek and Matsuoka, Yukiko and Ghosh, Samik and Glaab, Enrico and Trefois, Christophe and Crespo, Isaac and Perumal, Thanneer M. and Jurkowski, Wiktor and Antony, Paul M. A. and Diederich, Nico and Buttini, Manuel and Kodama, Akihiko and Satagopam, Venkata P. and Eifes, Serge and del Sol, Antonio and Schneider, Reinhard and Kitano, Hiroaki and Balling, Rudi: Integrating Pathways of Parkinson's Disease in a Molecular Interaction Map. Molecular Neurobiology, Springer, Vol. 49, pp : 88-102, (2014)
8. Azer, Karim and Kaddi, Chanchala D and Barrett, Jeffrey S and Bai, Jane P F and McQuade, Sean T and Merrill, Nathaniel J and Piccoli, Benedetto and Neves-Zaph, Susana and Marchetti, Luca and Lombardo, Rosario and Parolo, Silvia and Immanuel, Selva Rupa Christinal and Baliga, Nitin S: History and Future Perspectives on the Discipline of Quantitative Systems Pharmacology Modeling and Its Applications. Frontiers in physiology, Frontiers Media SA, Vol. 12, pp : 637999, (2021)
9. Bloomingdale, Peter and Karelina, Tatiana and Cirit, Murat and Muldoon, Sarah F. and Baker, Justin and McCarty, William J. and Geerts, Hugo and Macha, Sreeraj: Quantitative systems pharmacology in neuroscience: Novel methodologies and technologies. CPT Pharmacometrics Syst. Pharmacol., CPT, Vol. 10, pp : 412-419, (2021)
10. Srinivas, Vivek and Arrieta-Ortiz, Mario L. and Kaur, Amardeep and Peterson, Eliza J. R. and Baliga, Nitin S.: PerSort Facilitates Characterization and Elimination of Persister Subpopulation in Mycobacteria. Msystems, American Society for Microbiology, Vol. 5, pp : , (2011)
11. Lacroix, Clemence and Soeiro, Thomas and Le Marois, Marguerite and Guilhaumou, Romain and Casse-Perrot, Catherine and Jouve, Elisabeth and Rohl, Claas and Belzeaux, Raoul and Micallef, Joelle and Blin, Olivier: Innovative approaches in CNS clinical drug development: Quantitative systems pharmacology: Therapie, Elsevier, Vol. 76, pp : 111-119, (2021)
12. Sorger, PK and Allerheiligen, SRB and Abernethy, DR and Altman, RB and Brouwer, KLR and Califano, A and D'Argenio, DZ and Iyengar, R and Jusko, WJ and Lalonde, R and others: Quantitative and systems pharmacology in the post-genomic era: new approaches to discovering drugs and understanding therapeutic mechanisms. An NIH white paper by the QSP workshop group, NIH Bethesda, Vol. , pp : 1-48, (2011)
13. Bradshaw, EL and Spilker, ME and Zang, R and Bansal, L and He, H and Jones, RDO and Le, K and Penney, M and Schuck, E and Topp, B and Tsai, A and Xu, C and Nijsen, MJMA and Chan, JR: Applications of quantitative systems pharmacology in model-informed drug discovery: perspective on impact and opportunities. CPT Pharmacometrics Syst. Pharmacol., CPT, Vol. 8, pp : 777-791, (2019)
14. Balbas-Martinez, Violeta and Ruiz-Cerdá, Leire and Irurzun-Arana, Itziar and González-García, Ignacio and Vermeulen, An AND Gómez-Mantilla, José David and Trocóniz, Iñaki F.: A systems pharmacology model for inflammatory bowel disease. PloS one, Public Library of Science, Vol. 13, pp : 1-19, (2018)
15. Kaddi, C. D. and Niesner, B. and Baek, R. and Jasper, P. and Pappas, J. and Tolsma, J. and et al.: Quantitative systems pharmacology modeling of acid Sphingomyelinase deficiency and the enzyme replacement therapy Olipudase Alfa is an innovative tool for linking pathophysiology and pharmacology. CPT Pharmacometrics Syst. Pharmacol., CPT, Vol. 7, pp : 442-452, (2018)
16. Coletti, Roberta and Leonardelli, Lorena and Parolo, Silvia and Marchetti, Luca: A QSP model of prostate cancer immunotherapy to identify effective combination therapies. Scientific Reports, Nature Portfolio, Vol. , pp : 9063, (2020)
17. Jefferey E. Ming and Ruth Abrams and Derek W. Barlett and Tu Nguyen and Katherine Kudrycki and Ananth Kadambi and Christina M. Friedrich and Nassim Djebli and Britta Goebel and Joseph Elassal and Poulabi Banerjee and Michael J. Reed and Jeffrey S. Barrett and Karim Azer: A Quantitative Systems Pharmacology Platform to Investigate the Impact of Alirocumab and Cholesterol-Lowering Therapies on Lipid Profiles and Plaque Characteristics. Gene Regulation and Systems Biology, Libertas Academica, Vol. , pp : , (2017)
18. Peterson, M. C. and Riggs, M. M.: FDA Advisory Meeting Clinical Pharmacology Review Utilizes a Quantitative Systems Pharmacology (QSP) Model: A Watershed Moment?. CPT Pharmacometrics Syst. Pharmacol., CPT, Vol. 4, pp : 189-192, (2015)

19. Kuchimanchi, Mita and Monine, Michael and Kandadi Muralidharan, Kumar and Woodward, Caroline and Penner, Natalia: Phase II Dose Selection for Alpha Synuclein-Targeting Antibody Cinpanemab (BIIB054) Based on Target Protein Binding Levels in the Brain. CPT Pharmacometrics Syst. Pharmacol., CPT, Vol. 9, pp : 515-522, (2020)
20. Nahid, Payam and Dorman, Susan E. and Alipanah, Narges and Barry, Pennan M. and Brozek, Jan L. and Cattamanchi, Adithya and Chaisson, Lelia H. and Chaisson, Richard E. and Daley, Charles L. and Grzemska, Malgosia and Higashi, Julie M. and Ho, Christine S. and Hopewell, Philip C. and Keshavjee, Salmaan A. and Lienhardt, Christian and Menzies, Richard and Merrifield, Cynthia and Narita, Masahiro and O'Brien, Rick and Peloquin, Charles A. and Raftery, Ann and Saukkonen, Jussi and Schaaf, H. Simon and Sotgiu, Giovanni and Starke, Jeffrey R. and Migliori, Giovanni Battista and Vernon, Andrew: Executive Summary: Official American Thoracic Society/Centers for Disease Control and Prevention/Infectious Diseases Society of America Clinical Practice Guidelines: Treatment of Drug-Susceptible Tuberculosis. Clinical infectious diseases, Oxford University Press, Vol. 63, pp : 853-867, (2016)
21. Mdluli, Khisimuzi and Kaneko, Takushi and Upton, Anna: Tuberculosis drug discovery and emerging targets. Antimicrobial therapeutics reviews: infectious diseases of current and emerging concern, Annals of the New York Academy of Sciences, Vol. 1323, pp : 56-75, (2014)
22. Palomino, Juan Carlos and Martin, Anandi: Drug Resistance Mechanisms in Mycobacterium tuberculosis. Antibiotic-basel, MDPI, Vol. 3, pp : 317-340, (2014)
23. Boshoff, Helena IM and Myers, Timothy G and Copp, Brent R and McNeil, Michael R and Wilson, Michael A and Barry, Clifton E: The transcriptional responses of Mycobacterium tuberculosis to inhibitors of metabolism novel insights into drug mechanisms of action. Journal of Biological Chemistry, American Society for Biochemistry and Molecular Biology, Vol. 279, pp : 40174-40184, (2004)
24. Kanehisa, Minoru and Sato, Yoko and Furumichi, Miho and Morishima, Kanae and Tanabe, Mao: New approach for understanding genome variations in KEGG. Nucleic Acids Research, Oxford University Press, Vol. 47, pp : D590-D595, (2018)
25. Kanehisa, Minoru and Furumichi, Miho and Tanabe, Mao and Sato, Yoko and Morishima, Kanae: KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research, Oxford University Press, Vol. 45, pp : D353-D361, (2016)
26. Kanehisa, Minoru and Goto, Susumu: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research, Oxford University Press, Vol. 28, pp : 27-30, (2000)
27. Edgar, Ron and Domrachev, Michael and Lash, Alex E: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research, Oxford University Press, Vol. 30, pp : 207-210, (2002)
28. Charton, François and Hayat, Amaury and McQuade, Sean T and Merrill, Nathaniel J and Piccoli, Benedetto: A deep language model to predict metabolic network equilibria. Preprint, Arxiv, Vol. , pp : , (2021)
29. Bullo, F. and Cortes, J. and Dorfler, F. and Martinez, S.: Lectures on network systems. , , Vol. , pp : , (2016)
30. Jacquez, John A and Simon, Carl P: Qualitative theory of compartmental systems. Siam Review, SIAM, Vol. 35, pp : 43-79, (1993)
31. Klinke, David J and Finley, Stacey D: Timescale analysis of rule-based biochemical reaction networks. Biotechnology progress, Wiley Online Library, Vol. 28, pp : 33-44, (2012)
32. Maeda, Hajime and Kodama, Shinzo and Ohta, Yuzo: Asymptotic behavior of nonlinear compartmental systems: nonoscillation and stability. IEEE Transactions on Circuits and Systems, IEEE, Vol. 25, pp : 372-378, (1978)
33. Biggs, Norman and Biggs, Norman Linstead and Biggs, Emeritus Norman: Algebraic graph theory. Cambridge university press, Cambridge university press, Vol. 67, pp : , (1993)
34. Caughman, John S and Veerman, JJP: Kernels of directed graph Laplacians. Electronic Journal of Combinatorics, EJC, Vol. 13, pp : 39, (2006)

35. Gunawardena, Jeremy: A linear framework for time-scale separation in nonlinear biochemical systems. PloS one, Public Library of Science, Vol. 7, pp : e36321, (2012)
36. Mirzaev, Inomzhon and Gunawardena, Jeremy: Laplacian dynamics on general graphs. Bulletin of mathematical biology, Springer, Vol. 75, pp : 2118-2149, (2013)
37. De Leenheer, Patrick: The zero deficiency theorem. Notes for the Biomath Seminar I–MAP6487, Fall, , Vol. 9, pp : , (2009)
38. Feinberg, Martin and Horn, Friedrich JM: Dynamics of open chemical systems and the algebraic structure of the underlying reaction network. Chemical Engineering Science, Elsevier, Vol. 29, pp : 775-787, (1974)
39. Leenheer, Patrick De and Angeli, David and Sontag, Eduardo D.: Monotone Chemical Reaction Networks. Journal of Mathematical Chemistry, Springer, Vol. 41, pp : 295-314, (2007)
40. Sontag, Eduardo D.: Monotone and near-monotone biochemical networks. Systems and Synthetic Biology, Springer Science, Vol. 1, pp : 59-87, (2007)
41. Covert, Markus W and Schilling, Christophe H and Palsson, Bernhard: Regulation of gene expression in flux balance models of metabolism. Journal of theoretical biology, Elsevier, Vol. 213, pp : 73-88, (2001)
42. Palsson, Bernhard: Systems biology. , Cambridge university press, Vol. , pp : , (2015)
43. Cinlar, Erhan: Introduction to stochastic processes. , Courier Corporation, Vol. , pp : , (2013)
44. Alon, Uri: An Introduction to Systems Biology. , Chapman and Hall/CRC, Vol. , pp : , (2006)
45. Fortun, N and Lao, A and Razon, L and Mendoza, E: A deficiency zero theorem for a class of power-law kinetic systems with non-reactant-determined interactions. MATCH Commun. Math. Comput. Chem, Faculty of Science and University of Kragujevac, Vol. 81, pp : 621-638, (2019)
46. Kauffman, Kenneth J and Prakash, Purusharth and Edwards, Jeremy S: Advances in flux balance analysis. Current opinion in biotechnology, Elsevier, Vol. 14, pp : 491-496, (2003)
47. Orth, Jeffrey D and Thiele, Ines and Palsson, Bernhard O: What is flux balance analysis?. Nature biotechnology, Nature Publishing Group, Vol. 28, pp : 245, (2010)
48. Price, Nathan D and Famili, Iman and Beard, Daniel A and Palsson, Bernhard Ø: Extreme pathways and Kirchhoff's second law. Biophysical journal, Elsevier, Vol. 83, pp : 2879-2882, (2002)
49. Ford, Lester R and Fulkerson, Delbert R: Maximal flow through a network. Canadian journal of Mathematics, Cambridge University Press, Vol. 8, pp : 399-404, (1956)
50. Alon, Uri: Network motifs: theory and experimental approaches. Nature Reviews Genetics, Nature Portfolios, Vol. 8, pp : 450-461, (2007)
51. Erdős, Paul and Rényi, Alfréd and others: On the evolution of random graphs. Publ. math. inst. hung. acad. sci, , Vol. 5, pp : 17-60, (1960)
52. Domitilla Del Vecchio: Modularity, context-dependence, and insulation in engineered biological circuits. Trends in Biotechnology, Elsevier, Vol. 33, pp : 111-119, (2015)
53. Merrill, Nathaniel J. and An, Zheming and McQuade, Sean T. and Garin, Federica and Azer, Karim and Abrams, Ruth and Piccoli, Benedetto: Stability of metabolic networks via Linear-in-Flux-Expressions. Networks & Heterogeneous Media, American Institute of Mathematics, Vol. 14, pp : 101-130, (2019)
54. Schilling, Christophe H and Letscher, David and Palsson, Bernhard O: Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. Journal of theoretical biology, Elsevier, Vol. 203, pp : 229-248, (2000)
55. Lawler, Eugene L.: Cutsets and partitions of hypergraphs. Networks, Wiley Online Library, Vol. 3, pp : 275-285, (1973)
56. Cambini, Riccardo and Gallo, Giorgio and Scutella, Maria Grazia: Flows on Hypergraphs. Mathematical Programming, Springer Science, Vol. 78, pp : 195-217, (1997)
57. Hoffman, A.J: A generalization of Max Flow-Min Cut. Mathematical Programming, Springer Science, Vol. 6, pp : 352-359, (1974)
58. Lee, Geon and Ko, Jihoon and Shin, Kijung: Hypergraph motifs: Concepts, algorithms, and discoveries. Preprint, Arxiv, Vol. , pp : , (2020)



Bernard Bonnard and Ivan Kupka with a note from Prudence Kupka on the back of the picture

Bernard Bonnard, Monique Chyba
David Holcman and Emmanuel Trélat (Eds.)
IVAN KUPKA LEGACY：A Tour Through Controlled Dynamics

This volume is a tribute to Ivan Kupka who passed away at Easter 2023. It contains a collection of articles written by colleagues working on dynamical systems or control theory. Those colleagues interacted with Ivan Kupka directly or indirectly through his articles.

In his earliest work (1963) Ivan proved independently the Kupka-Smale theorem: "in a compact manifold the set of vector fields with the following properties are generic: all closed orbits are hyperbolic and heteroclinic orbits are transversal". At the end of the 70s he made significant contributions to controllability properties of invariant vector fields on semi-simple Lie groups, which are being applied to space mechanics and quantum control decades later. In the mid 80s Ivan proved the ubiquity of Fuller phenomenon which allows to describe the complex behaviors of the extremal trajectories in optimal control and provides an obstruction to the construction of optimal syntheses in the sub-analytic category. In 2001, he co-authored a well-cited book "Deterministic Observation Theory and Applications" which deals with geometric observation, state and parameters estimation. Since the end of the 20th century and for the next two decades he made significant contributions to stochastic systems with life and chemical science applications.

The first chapter of this book is an unpublished handwritten paper from Ivan Kupka which discuss optimality of regular extremals. The remaining articles are original contributions devoted to geometric optimal control with applications to Lotka-Volterra equations, Zermelo navigation problem, Quadcoper dynamics, Quantum control and Control of biological systems. Additional articles deal with Hybrid control and Reversible dynamical systems and some contributions couple geometric control which advanced numerical methods in optimal control. This book is a legacy to Ivan Kupka and his research interests and provides a modern perspective for researchers and graduate students to geometric control and its applications connecting applied mathematics and control engineering.

aimsciences.org