



**HAL**  
open science

# Control and Optimization of Physical Systems: Quantum Dynamics and Magnetic Confinement in Stellarators

Rémi Robin

► **To cite this version:**

Rémi Robin. Control and Optimization of Physical Systems: Quantum Dynamics and Magnetic Confinement in Stellarators. Optimization and Control [math.OC]. Sorbone Université, 2022. English. NNT: . tel-03779871

**HAL Id: tel-03779871**

**<https://inria.hal.science/tel-03779871v1>**

Submitted on 19 Sep 2022

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

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

**SORBONNE UNIVERSITÉ**  
**LJLL**

École doctorale **Sciences Mathématiques de Paris Centre**  
Unité de recherche **Laboratoire Jacques-Louis Lions, Équipe Inria CAGE**

Thèse présentée par **Rémi ROBIN**

Soutenue le **16 septembre 2022**

En vue de l'obtention du grade de docteur de Sorbonne Université

Discipline **Mathématiques appliquées**

Spécialité **Contrôle et optimisation**

**Contrôle et optimisation de systèmes  
physiques : applications à la mécanique  
quantique et au confinement  
magnétique dans les stellarators**

**Thèse dirigée par** Mario SIGALOTTI directeur  
Ugo BOSCAIN co-directeur

**Composition du jury**

<i>Rapporteurs</i>	Pierre ROUCHON Enrique ZUAZUA	professeur aux Mines de Paris professeur à l'Université de Erlan- gen-Nuremberg	
<i>Examineurs</i>	Bruno DESPRES Karine BEAUCHARD Yannick PRIVAT	professeur à Sorbonne Université professeure à l'ENS de Rennes professeur à l'Université de Stras- bourg	président du jury
<i>Directeurs de thèse</i>	Mario SIGALOTTI Ugo BOSCAIN	directeur de recherche à l'INRIA directeur de recherche au CNRS	



**SORBONNE UNIVERSITÉ**  
**LJLL**

Doctoral School **Sciences Mathématiques de Paris Centre**  
University Department **Laboratoire Jacques-Louis Lions, Équipe Inria CAGE**

Thesis defended by **Rémi ROBIN**

Defended on **16<sup>th</sup> September, 2022**

In order to become Doctor from Sorbonne Université

Academic Field **Applied mathematics**

Speciality **Control and optimization**

# Control and Optimization of Physical Systems: Quantum Dynamics and Magnetic Confinement in Stellarators

**Thesis supervised by** Mario SIGALOTTI Supervisor  
Ugo BOSCAIN Co-Supervisor

## Committee members

<i>Referees</i>	Pierre ROUCHON	Professor at Mines de Paris	
	Enrique ZUAZUA	Professor at Université de Erlan- gen–Nuremberg	
<i>Examiners</i>	Bruno DESPRES	Professor at Sorbonne Université	Committee President
	Karine BEAUCHARD	Professor at ENS de Rennes	
	Yannick PRIVAT	Professor at Université de Stras- bourg	
<i>Supervisors</i>	Mario SIGALOTTI	Senior Researcher at INRIA	
	Ugo BOSCAIN	Senior Researcher at CNRS	





Cette thèse a été préparée au

**Laboratoire Jacques-Louis Lions, Équipe Inria  
CAGE**

Sorbonne Université  
Campus Pierre et Marie Curie  
4 place Jussieu  
75005 Paris  
France

☎ +33 1 44 27 42 98  
Site <https://ljl11.math.upmc.fr/>





---

**CONTRÔLE ET OPTIMISATION DE SYSTÈMES PHYSIQUES : APPLICATIONS À LA MÉCANIQUE QUANTIQUE ET AU CONFINEMENT MAGNÉTIQUE DANS LES STELLARATORS**
**Résumé**

Cette thèse porte sur l'optimisation et le contrôle de plusieurs systèmes physiques : elle est composée de trois parties.

La première partie est consacrée aux stellarators. Ce type de réacteur à fusion nucléaire pose de nombreux défis liés à l'optimisation. Nous nous sommes concentrés sur un problème inverse bien connu des physiciens, modélisant la conception optimale de bobines supraconductrices générant un champ magnétique donné. Nous avons conduit une étude théorique et numérique d'une extension de ce problème, portant sur une optimisation de forme. Nous avons ensuite développé une nouvelle méthode afin de prouver l'existence de formes optimales dans le cas de problèmes d'optimisation d'hypersurfaces. Nous avons enfin effectué l'étude et l'optimisation des forces de Laplace s'exerçant sur une densité surfacique de courant.

La deuxième partie porte ensuite sur l'étude du contrôle de systèmes quantiques de dimension finie. Nous avons étudié rigoureusement la combinaison de l'approximation de l'onde tournante avec l'approximation adiabatique. Dans un premier temps, nous avons obtenu la robustesse des méthodes de transfert de population sur les qubits. Cette dernière permet alors d'étendre des résultats de Li et Khaneja sur le contrôle d'ensemble des qubits en se restreignant à l'utilisation d'un seul contrôle. Nous présentons également une seconde contribution, consacrée à l'analyse d'un phénomène de *chattering* pour un problème de contrôle optimal d'un système quantique.

Enfin, la troisième partie est dédiée à la preuve d'un résultat de contrôlabilité à zéro en temps petit pour des équations de Burgers généralisées grâce à l'utilisation d'une couche limite.

**Mots clés :** contrôle optimal, contrôle quantique, contrôlabilité d'ensemble, qubit, optimisation de forme, physique des plasmas, stellarator, équation de Burgers, couche limite

---

**CONTROL AND OPTIMIZATION OF PHYSICAL SYSTEMS: QUANTUM DYNAMICS AND MAGNETIC CONFINEMENT IN STELLARATORS**
**Abstract**

This PhD manuscript deals with the optimization and control of several physical systems. It is divided into three parts.

The first part is devoted to stellarators. This type of nuclear fusion reactor poses many challenges related to optimization. We focus on an inverse problem well known to physicists, modeling the optimal design of superconducting coils generating a given magnetic field. We conduct both a theoretical and a numerical study of an extension of this problem, involving shape optimization. Then, we develop a new method to prove the existence of optimal shapes in the case of hypersurface optimization problems. Finally, we study and optimize the Laplace forces acting on a current surface density.

The second part of this manuscript deals with the control of finite dimensional quantum systems. We rigorously study the combination of the rotating wave approximation with the adiabatic approximation. First, we obtain the robustness of a population transfer method on qubits. The latter then allows to extend results of Li and Khaneja on the ensemble control of qubits by restricting to the use of a single control. We also present a second contribution, devoted to the analysis of a chattering phenomenon for an optimal control problem of a quantum system.

Finally, the third part is dedicated to the proof of a small-time global null controllability result for generalized Burgers' equations using a boundary layer.

**Keywords:** optimal control, quantum control, ensemble controllability, qubit, shape optimization, plasma physics, stellarator, Burgers equation, boundary layer

---

**Laboratoire Jacques-Louis Lions, Équipe Inria CAGE**

Sorbonne Université – Campus Pierre et Marie Curie – 4 place Jussieu – 75005 Paris – France



# Remerciements

Je voudrais tout d'abord remercier mes deux directeurs de thèse, Ugo Boscain et Mario Sigalotti, pour ces trois années. Vous avez tous les deux été des directeurs exceptionnels : disponibles, encourageants, exigeants tout en me laissant une grande liberté que j'ai particulièrement appréciée. Ma rencontre avec Ugo à l'X fût déterminante ; je pensais alors devenir informaticien ou physicien. Avec Mario, vous m'avez alors fait découvrir des mathématiques non seulement d'une grande beauté mais aussi dotées de nombreuses applications en physique et en ingénierie. Mario, tes conseils, tes questions et ta très grande disponibilité furent clairement des éléments décisifs dans la réussite de cette thèse. Merci infiniment à vous deux.

J'aimerais à présent remercier grandement Pierre Rouchon et Enrique Zuazua pour avoir accepté de rapporter ma thèse et pour leurs encouragements. Je suis également fortement reconnaissant aux autres membres de mon jury : Karine Beauchard, Bruno Despres et Yannick Privat ; c'est un privilège de pouvoir présenter ma thèse devant vous.

Une grande partie de cette thèse fut l'objet de collaborations ; outre mes directeurs de thèse, je souhaite rendre hommage à Francesco Volpe pour m'avoir fait découvrir les *stellarators* et présenté de nombreux problèmes extrêmement intéressants. Un grand merci également à Yannick Privat qui m'a introduit à l'optimisation de forme ; travailler avec toi est une super expérience. Dans l'univers quantique, je remercie Dominique Sugny pour ses idées sur le *chattering* et bien sûr Nicolas Augier pour m'avoir aidé à percer un petit peu les mystères de la *rotating wave* et de l'adiabatique. Enfin, j'ai eu le plaisir de travailler un peu et d'apprendre beaucoup avec Jean-Michel Coron. Merci pour ton temps, tes conseils et tes encouragements.

J'ai particulièrement apprécié travailler au sein de laboratoire Jacques-Louis Lions. Je souhaite ainsi remercier le personnel administratif pour leur disponibilité ; ainsi que Kashayar dont l'aide informatique est toujours d'une efficacité redoutable. Je souhaite aussi remercier les permanents de la salle café, toujours disponibles pour parler de montagne, de Sobolev ou d'autres choses.

Un grand merci à tous les doctorants du LJLL toutes cuvées confondues, qui malgré la pandémie ont réussi à entretenir une ambiance sereine et agréable. En particulier, mes collègues de bureau dont les qualités en *ultimate* sont indéniables : Christophe, Amaury, Alex, Gabriela, Élise, Noémi et Yipeng. Une petite pensée pour la dynastie quantique/fusion : Robin et Eugénio, je pense à vous. Merci également aux nombreux doctorants qui font vivre le laboratoire à travers l'organisation du GTT, GT-EDP, respo bureau ou thé du labo : Ramon, Jesus, Augustin, Lucas, Pierre, Charles, Guillaume, Nicolai... Un grand merci aux représentants doctorants, Antoine, Emma, Juliette, Thomas et bien sûr Giorgia et Sylvain qui m'ont accompagné. Merci à Jules pour ses combats incessants pour défendre les doctorants, l'ambiance et l'environnement (avec une pensée pour Janco en prime). Merci aussi aux cyclistes Gontran et Valentin et aux anciens Anouk, Nicolas, Fatima, Allen, Ludo... Enfin, je souhaite remercier grassement les montagnards qui ont accepté de me tirer (et parfois même de se faire tirer) en trail, en ski, en chaussons ou en grosses : merci Lucas, merci Matthieu (l'ordre dans lequel ils descendent le Taillefer).

Pas au LJLL, mais jamais très loin, c'est toujours un plaisir voir Lev-Arcady, Louis, Gaspard, Benoit et tant d'autres en conférences.

Dans la montagne à nouveau, merci Barbo pour tes cabanes originales et Pierre-Christophe pour tes plans de dernières minutes. Sur des versants plus alcoolisés (toujours une bonne manière de présenter ses amis), je n'oublie pas Camille, Cyril, Naïla, Noémie et Thomas. Pour les portraits et les confinements heureux, je remercie ma coloc d'Arcueil : Alfred, Mimou, Gab, Gab, Marion, Mendès, Ludo, Anaëlle et Marilou. Pour le confinement grim pant, un grand merci à Gaëtan, Mimou (encore !) et Marc. Pour les raquettes et beaucoup d'autres choses, je remercie Vincent.

Lise, merci pour ces belles années.

Enfin, je remercie mes petites sœurs, incroyables sur (sous ?) l'eau, en pâtisserie ou encore à *Small-world* ; ainsi que mes parents pour tous ces moments heureux.

*à mes parents,  
Cécile et Thierry,  
mes sœurs,  
Lisa, Lucie et Marilou,  
ma pagaie, mes peaux et mes cordes*





# Contents

Résumé	vii
Remerciements	ix
Contents	xiii
<b>I Introduction et résumé des contributions</b>	<b>1</b>
I.1 Publications scientifiques . . . . .	1
I.2 Introduction générale . . . . .	1
I.3 À l'interface de la fusion nucléaire et de l'optimisation : les stellarators . .	2
I.3.1 Brève introduction à la fusion nucléaire contrôlée . . . . .	2
I.3.2 Confinement magnétique : tokamak et stellarator . . . . .	3
I.3.3 Un problème inverse sur les bobines de stellarator . . . . .	4
I.4 Contributions à l'optimisation des stellarators . . . . .	8
I.4.1 Optimisation de la <i>Coil Winding Surface</i> . . . . .	8
I.4.2 Quelques théorèmes d'existence en optimisation de forme . . . . .	12
I.4.3 Forces de Laplace s'exerçant sur une surface de courant . . . . .	16
I.4.4 Quelques problèmes ouverts et perspectives . . . . .	18
I.5 Contrôle de systèmes quantiques : motivations et outils . . . . .	19
I.5.1 Motivations physiques générales . . . . .	19
I.5.2 Formalisation mathématique . . . . .	20
I.5.3 Contrôlabilité d'ensemble de qubits . . . . .	23
I.6 Contributions au contrôle de systèmes quantiques . . . . .	27
I.6.1 Compatibilité entre l'approximation de l'onde tournante et l'approximation adiabatique . . . . .	27
I.6.2 Un exemple de <i>chattering</i> en contrôle quantique . . . . .	30
I.6.3 Quelques problèmes ouverts et perspectives . . . . .	32
I.7 Contrôlabilité globale à zéro en temps petit des équations de Burgers général- isées . . . . .	33
I.7.1 Description du système étudié . . . . .	33
I.7.2 Motivations et résultats existants . . . . .	33
I.7.3 Contribution et idées de preuves . . . . .	34
I.7.4 Perspectives et conclusion . . . . .	35
<b>First Part : Optimizations for Stellarators</b>	<b>37</b>

<b>II</b>	<b>Optimal shape of stellarators for magnetic confinement fusion</b>	<b>39</b>
II.1	Introduction . . . . .	39
II.1.1	Motivations: towards a shape optimization problem . . . . .	39
II.1.2	State of the art and main contributions of this chapter . . . . .	41
II.1.3	Notations . . . . .	43
II.1.4	Modeling: towards a shape optimization problem . . . . .	44
II.2	Existence issues for Problem ( $\mathcal{P}_{\text{shape}}$ ) . . . . .	46
II.2.1	Existence of an optimal current for a given shape . . . . .	46
II.2.2	Existence of an optimal shape . . . . .	47
II.3	Shape differentiation for Problem ( $\mathcal{P}_{\text{shape}}$ ) . . . . .	54
II.3.1	Shape derivative of the cost functional $C$ . . . . .	54
II.3.2	Proof of Theorem II.11 . . . . .	55
II.4	Numerical implementation . . . . .	60
II.4.1	Parametrization issues . . . . .	60
II.4.2	Implementation . . . . .	64
II.4.3	Numerical results . . . . .	64
II.A	Appendix . . . . .	65
II.A.1	Some differential geometry . . . . .	65
II.A.2	Reach constraint and sets of positive reach . . . . .	69
II.A.3	Jacobian determinant and changes of variables on manifolds . . . . .	70
<b>III</b>	<b>Existence of surfaces optimizing geometric and PDE shape functionals under reach constraint</b>	<b>71</b>
III.1	Framework and main results . . . . .	71
III.1.1	Introduction . . . . .	71
III.1.2	Notations . . . . .	72
III.1.3	Preliminaries on sets of uniformly positive reach . . . . .	73
III.1.4	Main results . . . . .	74
III.2	Proofs . . . . .	77
III.2.1	The extruded surface approach . . . . .	77
III.2.2	Proof of Lemma III.4 . . . . .	80
III.2.3	Proof of Theorem III.6 . . . . .	81
III.2.4	Proof of Theorem III.10 . . . . .	84
III.2.5	Main steps in the proof of Theorem III.11 . . . . .	87
III.3	Conclusion . . . . .	89
III.A	Appendix . . . . .	89
III.A.1	Curvatures of a submanifold . . . . .	89
III.A.2	$R$ -convergence: proof of Theorem III.3 . . . . .	90
III.A.3	The Laplace–Beltrami equation on a manifold: proof of Theorem III.9 . . . . .	90
<b>IV</b>	<b>Minimization of magnetic forces on Stellarator coils</b>	<b>93</b>
IV.1	Introduction . . . . .	93
IV.2	Laplace force on a surface . . . . .	94
IV.2.1	Notations . . . . .	94
IV.2.2	Limit definition of Laplace force exerted by a current-sheet on itself . . . . .	95
IV.2.3	Computing the Laplace force exerted by one current-sheet on another . . . . .	97
IV.2.4	Justification from a 3D current modelisation . . . . .	99
IV.3	Examples of cost functions . . . . .	99
IV.4	Numerical simulations . . . . .	101

IV.4.1	Setup . . . . .	101
IV.4.2	Adding force minimization and improving regularization in REGCOIL . . . . .	103
IV.4.3	Numerical results . . . . .	104
IV.5	Summary, conclusions and future work . . . . .	105
IV.A	Appendix . . . . .	107
IV.A.1	First tangential term . . . . .	107
IV.A.2	Second tangential term . . . . .	109
IV.A.3	First normal term . . . . .	109
IV.A.4	Second normal term . . . . .	111
IV.B	Proof of existence of minimisers . . . . .	112
 <b>Second Part: Quantum control</b>		<b>117</b>
<b>V</b>	<b>Ensemble qubit controllability with a single control via AA and RWA</b>	<b>119</b>
V.1	Introduction . . . . .	119
V.1.1	Rotating wave approximation . . . . .	120
V.1.2	Adiabatic approximation . . . . .	122
V.1.3	Combination of RWA and AA and statement of the population inversion result . . . . .	123
V.2	Application to the ensemble control problem . . . . .	126
V.3	Proof of Theorem V.3 . . . . .	129
V.3.1	A first change of variables . . . . .	129
V.3.2	Idea of the proof . . . . .	130
V.3.3	The rotating wave approximation . . . . .	131
V.3.4	Two scales adiabatic approximation . . . . .	136
V.4	Numerical simulations . . . . .	140
<b>VI</b>	<b>Chattering Phenomenon in Quantum Optimal Control</b>	<b>145</b>
VI.1	Contribution . . . . .	145
VI.1.1	Introduction . . . . .	145
VI.1.2	Model . . . . .	147
VI.1.3	Description of the optimal control . . . . .	147
VI.1.4	Numerical simulations . . . . .	149
VI.1.5	Conclusion . . . . .	152
VI.2	Technical results . . . . .	152
VI.2.1	The model system . . . . .	153
VI.2.2	The Fuller model . . . . .	154
VI.2.3	Properties of the switching function for the three-level quantum system . . . . .	155
VI.2.4	A sufficient condition for chattering . . . . .	158
VI.2.5	Numerical optimization procedure . . . . .	160
 <b>Third Part: Controllability of one dimensional fluid dynamics equations</b>		<b>163</b>

<b>VII Small-time global null controllability of generalized Burgers' equations</b>	<b>165</b>
VII.1 Introduction . . . . .	165
VII.1.1 Description of the system . . . . .	165
VII.1.2 Statement of our main result . . . . .	166
VII.1.3 Preliminaries . . . . .	167
VII.2 Hyperbolic stage, first part: toward a very stable steady state . . . . .	169
VII.2.1 The control strategy . . . . .	169
VII.2.2 Lower bound . . . . .	170
VII.2.3 Upper bound . . . . .	172
VII.3 Hyperbolic stage, second part: toward a neighborhood of zero up to a boundary layer . . . . .	174
VII.4 Passive stage: dissipation of the boundary residue . . . . .	176
VII.5 Parabolic stage: local null exact controllability . . . . .	178
VII.6 Open problems . . . . .	180
VII.A Parabolic regularity estimates for the heat equation . . . . .	181
<b>List of Figures</b>	<b>183</b>
<b>Bibliography</b>	<b>187</b>

# Chapitre I

## Introduction et résumé des contributions

### I.1 Publications scientifiques

La production scientifique de cette thèse est l'objet des publications suivantes :

1. Y. PRIVAT, R. ROBIN et M. SIGALOTTI. "Optimal shape of stellarators for magnetic confinement fusion". In : *Journal de Mathématiques Pures et Appliquées* 163 (2022), p. 231-264,
2. R. ROBIN et F. A. VOLPE. "Minimization of magnetic forces on stellarator coils". In : *Nuclear Fusion* 62.8 (2022), p. 086041,
3. R. ROBIN, N. AUGIER, U. BOSCAIN et M. SIGALOTTI. "Ensemble qubit controllability with a single control via adiabatic and rotating wave approximations". In : *Journal of Differential Equations* 318 (2022), p. 414-442,

ainsi que des articles en relecture suivants :

1. Y. PRIVAT, R. ROBIN et M. SIGALOTTI. *Existence of surfaces optimizing geometric and PDE shape functionals under reach constraint*. 2022. arXiv : 2206.04357 [math],
2. R. ROBIN, U. BOSCAIN, M. SIGALOTTI et D. SUGNY. *Chattering phenomenon in quantum optimal control*. 2022. arXiv : 2206.13868 [quant-ph],
3. R. ROBIN. *Small-time global null controllability of generalized Burgers' equations*. 2022. arXiv : 2206.05931 [math].

### I.2 Introduction générale

Ce manuscrit est composé de trois parties.

Dans une première partie, nous considérons des problèmes liés au confinement magnétique d'un plasma, dans le cadre de la réalisation d'un réacteur à fusion nucléaire de type stellarator.

Ces travaux ont été effectués dans le cadre de l'action exploratoire Inria StellaCage<sup>1</sup>, regroupant l'équipe Inria CAGE, Yannick Privat<sup>2</sup> et la start-up Renaissance Fusion<sup>3</sup>. Cette dernière ambitionne de réaliser un stellarator. Sous son impulsion, nous avons étudié des problèmes d'optimisation appliqués au confinement magnétique, et plus précisément à l'optimisation de la forme des bobines du stellarator. Nous présenterons tout d'abord un premier travail d'optimisation de forme dans le Chapitre II. Le Chapitre III portera sur des problèmes d'existence en optimisation de forme. Enfin, le Chapitre IV aura pour objet l'étude des forces de Laplace s'exerçant sur une surface de courant.

La deuxième partie de ce manuscrit est dédiée au contrôle de quelques systèmes quantiques de dimension finie. Dans le Chapitre V, nous commencerons par étudier la compatibilité entre l'approximation de l'onde tournante et l'approximation adiabatique. Nous en déduirons des résultats de contrôlabilité d'ensemble sur des qubits. Dans un second temps, nous analyserons dans le Chapitre VI un problème de contrôle optimal présentant un phénomène de *chattering*.

Enfin, une troisième partie correspondant au Chapitre VII, est consacrée à l'étude de la contrôlabilité globale à zéro en temps petit d'une famille d'équations provenant de la mécanique des fluides.

## I.3 À l'interface de la fusion nucléaire et de l'optimisation : les stellarators

### I.3.1 Brève introduction à la fusion nucléaire contrôlée

Commençons par quelques rappels de physique nucléaire. La fusion nucléaire est une réaction mettant en jeu deux atomes légers qui fusionnent pour former un atome plus lourd. Un tel mécanisme produit des quantités phénoménales d'énergie dues à une perte de masse des réactifs au cours de la réaction. À titre d'exemple, la fusion nucléaire est responsable de l'activité thermique du Soleil ainsi que de la majorité des étoiles.

La découverte expérimentale de la fusion nucléaire remonte à E. Rutherford qui réalise la fusion d'atomes de deutérium en 1934. Les conditions permettant une réaction de fusion nucléaire étant extrêmes, la maîtrise de ce phénomène reste aujourd'hui encore un défi technologique majeur. En effet, afin de fusionner les noyaux des deux atomes chargés positivement, ces derniers doivent outrepasser la barrière de potentiel coulombien. Malgré l'aide de l'effet tunnel, il est toutefois nécessaire d'amener les atomes à des énergies extrêmement élevées.

La première application de la fusion nucléaire fût militaire : en 1952, la première bombe H, c'est-à-dire régie par une réaction de fusion nucléaire, est testée par les États-Unis. Son mécanisme consiste à faire exploser une bombe nucléaire à fission afin d'enclencher la réaction de fusion nucléaire, produisant alors beaucoup plus d'énergie que la bombe à fission.

Dans le cas où l'énergie de la fusion nucléaire vise une utilisation industrielle (production d'électricité, propulsion ...) on parle de fusion nucléaire contrôlée. Cette dernière est le sujet de recherches particulièrement actives depuis la fin de la seconde guerre mondiale. Pour la production d'électricité, les promesses d'un tel réacteur sont attrayantes : pas de déchets radioactifs, pas de risque d'emballement de la réaction, combustibles très abondants<sup>4</sup>, pas d'émissions directes de gaz à effet de serre... Dans le contexte géopolitique et climatique actuel, la fusion nucléaire est un candidat sérieux, à moyen/long terme, pour résoudre partiellement la crise énergétique.

1. <https://www.ljll.math.upmc.fr/sigalotti/cage/stellacage.html>

2. Institut de Recherche Mathématique Avancée, Université de Strasbourg

3. <https://stellarator.energy/>

4. La réaction deutérium-tritium est privilégiée dans la plupart des dispositifs. Le deutérium est abondant sur Terre. Le tritium est actuellement un résidu de fission mais pourrait à l'avenir être produit par le réacteur

Deux grandes familles technologiques de solutions sont considérées pour répondre à ce défi : les technologies de fusion par confinement inertiel, et celles par confinement magnétique. Les premières se fondent sur l'utilisation de lasers à haute puissance pour piéger la matière et amorcer la réaction. Le laser Mégajoule du Commissariat à l'Énergie Atomique (CEA) est un exemple de confinement inertiel. Cependant, dans le cadre de cette thèse, nous nous focaliserons uniquement sur le confinement magnétique.

Afin de réaliser une réaction de fusion nucléaire contrôlée, les réacteurs à confinement magnétique utilisent un plasma extrêmement chaud. L'ordre de grandeur pour la température est de 150 millions de Kelvin. La température du Soleil, à titre de comparaison, est de 6000 Kelvin en surface et 15 millions de Kelvin au centre. Grâce à la pression gravitationnelle colossale qui y règne, les réactions de fusion s'y déroulent sans problème. Dans les plasmas terrestres, il est nécessaire de compenser la faible densité par des températures sensiblement plus élevées. De ce fait, le critère de qualité générale d'un réacteur est le célèbre *triple product*  $nT\tau_e$ , où  $n$  est la densité du plasma,  $T$  sa température et  $\tau_e$  la durée pendant laquelle le confinement est maintenu. Le facteur dit de gain d'énergie de fusion, noté  $Q$ , est un autre indicateur important.  $Q$  est le ratio entre l'énergie produite et l'énergie injectée dans le système. Un objectif à court terme est d'atteindre le *break-even* ( $Q = 1$ ). Le record actuel,  $Q = 0.7$ , a été établi en 2021 par le *National Ignition Facility* avec un confinement inertiel<sup>5</sup>. Le réacteur *International Thermonuclear Experimental Reactor* (ITER), devrait atteindre la valeur de  $Q = 10$ . Enfin, mentionnons que l'objectif d'un réacteur commercialisable serait d'atteindre l'*ignition*, c'est-à-dire d'obtenir une réaction autoentretenu, i.e. lorsque  $Q = \infty$ .

### I.3.2 Confinement magnétique : tokamak et stellarator

Il existe deux types principaux de réacteurs de fusion nucléaires à confinement magnétique : les tokamaks et les stellarators. Leur principe de fonctionnement est le suivant : une fois que le combustible est sous forme de plasma, il devient sensible au champ magnétique externe. Rappelons qu'en présence d'un champ magnétique uniforme, une particule chargée a un mouvement hélicoïdal le long des lignes de champ. Ainsi, dans un champ magnétique uniforme intense, les particules ne se dispersent pas dans les directions normales au champ magnétique. On souhaite alors "refermer" les lignes de champ afin de confiner les particules. La géométrie la plus simple pour ce faire est celle du tore.

Cependant, une difficulté importante apparaît. Le champ magnétique créé par un assemblage axisymétrique de bobines autour de l'axe  $Oz$  est proportionnel au champ  $\frac{e_\theta}{R}$  à l'intérieur des bobines.  $e_\theta$  désignant le vecteur unitaire dans la direction toroïdale, et  $R$  la distance à l'axe  $Oz$ . Le champ magnétique a donc une intensité inhomogène. Le mouvement d'une particule dans ce champ n'est alors plus hélicoïdal le long d'une ligne de champ : la particule subit une déviation verticale appelée *dérive* (ou *drift*) dépendant de la charge. La simulation de la Figure I.1 illustre ce phénomène. Le rayon de Larmor est en effet plus court dans les régions à fort champ magnétique et plus faible dans celles à faible intensité. Nous renvoyons par exemple à [IPW19, chapitre 5] pour une analyse du mouvement d'une particule chargée dans le cas d'un champ non homogène. Afin de résoudre ce problème, deux solutions principales sont disponibles. Les tokamaks, inventés par les physiciens soviétiques I. Tamm, A. Sakharov et O. Lavrentiev, utilisent un champ magnétique poloïdal afin de déplacer l'axe de rotation de Larmor et ainsi moyennent les effets de la dérive (voir Figure I.2). Les tokamaks représentent la technologie la plus mature actuellement ; le projet ITER étant le prochain réacteur majeur en construction. Afin de générer un champ magnétique poloïdal tout en préservant l'axisymétrie, un courant électrique induit doit

5. Dans le cas du confinement inertiel, seule la quantité d'énergie injectée dans le système via les lasers est comptabilisée comme énergie fournie, et non l'énergie dépensée pour faire fonctionner les lasers



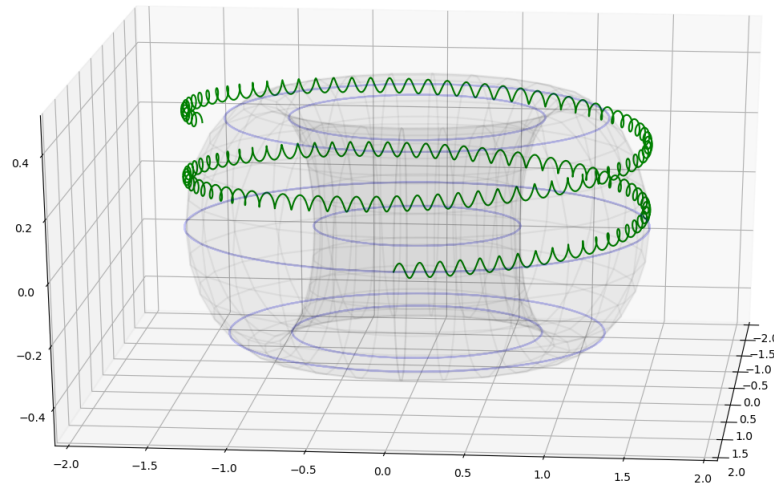


FIGURE I.1 – Simulation de la trajectoire d’une particule (en vert) dans un champ magnétique axisymétrique (en bleu). Courtoisie de Robin Roussel.

circuler dans le plasma. L’induction et la stabilisation de ce courant représentent une difficulté majeure, intrinsèque à l’utilisation des tokamaks.

La seconde solution technologique est celle du stellarator. Inventés en 1951 par l’américain L. Spitzer, et contrairement aux tokamaks, les stellarators ne nécessitent pas de courants électriques importants au sein du plasma. L’élimination de la dérive est induite par l’utilisation d’un champ magnétique complexe et non axisymétrique. En particulier, les bobines, qui ne sont pas coplanaires, sont extrêmement chères et difficiles à produire. En contrepartie, un stellarator présente beaucoup moins d’instabilités relatives au courant circulant dans le plasma. Les défenseurs de cette technologie ont coutume d’affirmer qu’un stellarator est plus complexe à construire au bénéfice d’une plus grande stabilité.

La conception d’un stellarator est un processus éminemment complexe, à l’interface de l’ingénierie de pointe et de la physique des plasmas.

Une des premières questions à se poser porte sur la forme du plasma et le choix du champ magnétique. Il faut pour cela optimiser de nombreux paramètres physiques afin de maximiser le temps de confinement. Par exemple, la stabilité de l’équilibre magnéto-hydrodynamique d’un plasma soumis à un tel champ magnétique. Une fois qu’un champ magnétique présentant de bonnes propriétés a été choisi, il est nécessaire de trouver un agencement de bobines permettant de le réaliser. C’est sur cette étape que nous nous concentrons dans le cadre de cette thèse.

### I.3.3 Un problème inverse sur les bobines de stellarator

Considérons le problème suivant : la forme du plasma ainsi que le champ magnétique à générer à l’intérieur sont donnés. On cherche le meilleur arrangement possible des bobines pour générer ce champ magnétique tout en garantissant que les bobines soient réalisables. Par exemple, il est nécessaire de laisser un espace suffisant entre les bobines et le plasma afin de réaliser l’enceinte à vide. Il faut également veiller à ce que les bobines ne soient ni trop complexes, ni le support d’un courant avec une intensité trop élevée. En résumé, il faut générer une approximation du champ magnétique à l’aide de bobines soumises à des contraintes d’ingénierie.

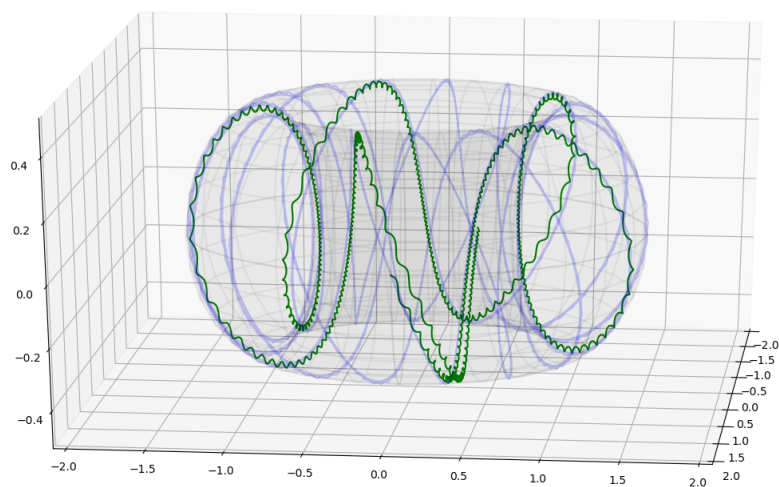


FIGURE I.2 – Simulation de la trajectoire (en vert) d'une particule dans un champ magnétique (en bleu) avec une composante poloïdale. Courtoisie de Robin Roussel.

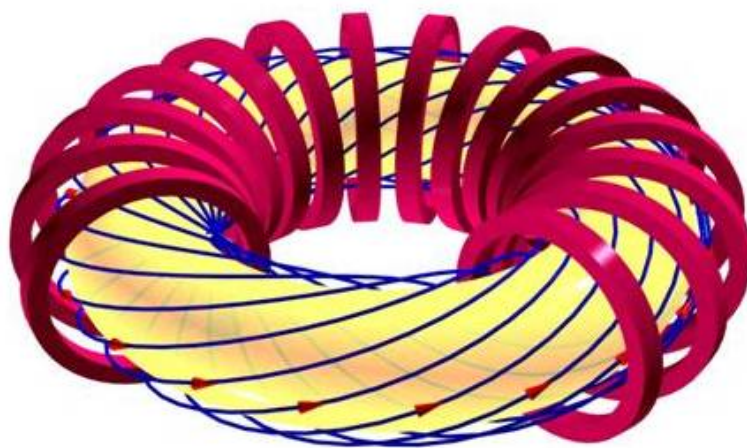


FIGURE I.3 – Schéma d'un tokamak, la torsion des lignes de champ magnétique est produite grâce à un courant électrique circulant à l'intérieur le plasma.

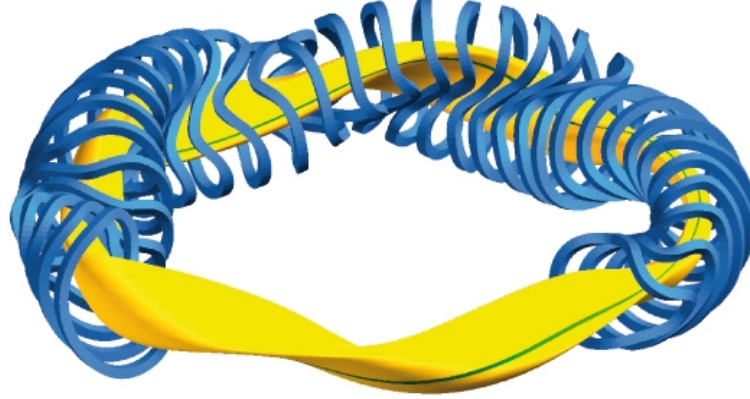


FIGURE I.4 – Schéma de Wendelstein 7-X, stellarator du Max-Planck Institut für Plasmaphysik achevé en 2015. Les bobines sont en bleue, le plasma en jaune et une ligne de champ magnétique est représentée en vert. L'image provient du Max-Planck Institut für Plasmaphysik

Pour toutes ces raisons, le problème est généralement relaxé en considérant une surface imaginaire sur laquelle seront disposés les bobines. Cette surface est dénommée *Coil Winding Surface* (CWS). Les bobines discrètes sont remplacées par un courant surfacique supporté sur la CWS. Avec la donnée de ce courant, on reconstruit ensuite les bobines ayant pour support la CWS.

Modélisons à présent la recherche de la densité surfacique optimale pour une CWS donnée. Introduisons pour cela les notations suivantes :

- $P$  est un ouvert de  $\mathbb{R}^3$  représentant le domaine du plasma.  $P$  est en pratique de forme toroïdale.
- $S$  est la CWS, cette surface enveloppe  $P$ . Notons également  $d\mu_S$  l'élément de surface associé.
- $\mathfrak{X}(S)$  correspond à l'ensemble des champs de vecteurs tangents lisses sur  $S$ .
- $\mathcal{F}_S$  est la complétion de  $\mathfrak{X}(S)$  pour le produit scalaire

$$\langle X_1, X_2 \rangle_{\mathcal{F}_S} = \int_S \langle X_1, X_2 \rangle d\mu_S, \quad (X_1, X_2) \in \mathfrak{X}(S)^2.$$

- $\mathcal{F}_S^0$  est la fermeture pour la norme  $\|\cdot\|_{\mathcal{F}_S}$  des champs à divergence (tangentielle) nulle de  $\mathfrak{X}(S)$ .

Rappelons également que le champ magnétique généré par un courant surfacique  $j \in \mathcal{F}_S^0$  au point  $y \notin S$  est donné par la loi de Biot et Savart :

$$\text{BS}(j)(y) = \int_S \frac{x - y}{|x - y|^3} \times j(x) d\mu_S(x). \quad (\text{I.1})$$

Le problème d'optimisation consistant à trouver la densité de courant optimale sur  $S$  afin de générer  $B_T \in L^2(P, \mathbb{R}^3)$  se formule donc comme suit :

$$\inf_{j \in \mathcal{F}_S^0} \|\text{BS}(j) - B_T\|_{L^2(P)}^2. \quad (\text{I.2})$$

Le problème inverse (I.2) a été introduit par Merkel dans [Mer86 ; Mer87]. Une fois l'équation discrétisée, on est ramené à un problème de moindres carrés que l'on peut résoudre explicitement. L'implémentation numérique a été effectuée dans le code de référence en langage Fortran dénommé NESCOIL en utilisant une base de type Fourier tronquée pour  $\mathcal{F}_S^0$ .

Une fois la détermination d'une densité de courant effectuée, l'obtention des bobines consiste à suivre les courbes de niveau d'un *potentiel de courant généralisé*. Il est ensuite possible d'optimiser les positions des bobines ainsi obtenues en initialisant avec la configuration donnée sur la CWS. Cette étape n'est pas étudiée dans cette thèse.

Revenons à l'Équation (I.2) : dans la pratique  $P$  et  $S$  ne s'intersectent pas. Par conséquent, on obtient facilement que  $BS : \mathcal{F}_S^0 \rightarrow L^2(P, \mathbb{R}^3)$  est un opérateur compact. Rappelons qu'un opérateur compact ne peut avoir d'inverse continue en dimension infinie. Ainsi, **le problème inverse (I.2) est mal posé**. D'un point de vue numérique, cela implique que lorsque l'on augmente la taille de la famille utilisée pour générer  $\mathcal{F}_S^0$ , la suite des normes des minimiseurs explose.

Afin de résoudre ce problème, M. Landreman a introduit dans [Lan17] une régularisation de Tychonoff au problème via un paramètre  $\lambda > 0$ . Pour une CWS  $S$  donnée, on cherche alors à résoudre :

$$C(S) := \inf_{j \in \mathcal{F}_S^0} \|BS(j) - B_T\|_{L^2(P)}^2 + \lambda \|j\|_{\mathcal{F}_S^0}^2. \quad (\text{I.3})$$

Ce problème est strictement convexe. Il est bien posé et admet un unique minimiseur dans l'espace de dimension infini  $\mathcal{F}_S^0$ . De plus, l'expression du minimiseur  $j_S$  est explicite (cf. Lemme II.3) :

$$j_S = \left( BS^\dagger BS + \lambda \text{Id} \right)^{-1} BS^\dagger B_T, \quad (\text{I.4})$$

où  $BS^\dagger$  est l'adjoint de  $BS$  dans  $\mathcal{F}_S^0$ . On prouve facilement (voir Lemme II.3) que  $BS^\dagger BS + \lambda \text{Id}$  est inversible et que cet inverse est continue pour  $\lambda > 0$ . Par ailleurs,  $\lambda$  pénalise l'utilisation de courants électriques trop élevés.

En résumé, cette approche utilisant une CWS permet de linéariser le problème de placement des bobines et rend sa résolution numérique simple et rapide. Cependant, un des défauts de cette approche est que la CWS est fixée *a priori*, alors même que sa forme influence fortement les performances.

Il est alors pertinent de se demander comment optimiser le coût (I.3) sur un ensemble de surfaces candidates, c'est-à-dire optimiser  $C(S)$  pour  $S$  dans un ensemble de formes admissibles. Un premier travail d'optimisation de la forme de la CWS dans ce cadre a été effectué par Paul et al. dans [Pau+18]. L'approche utilisée consiste à commencer par discrétiser ; la surface est alors représentée par des coefficients de Fourier. Puis dans un second temps, d'optimiser ces coefficients afin de minimiser une combinaison linéaire de  $C(S)$  et d'autres coûts relatifs à la surface. Ces coûts représentent des caractéristiques physiques de la surface : distance au plasma, aire, volume, mais aussi pénalisations ( $H^k$ ) des coefficients de Fourier élevés.

## I.4 Contributions à l'optimisation des stellarators

### I.4.1 Optimisation de la *Coil Winding Surface*

Ce travail a été effectué en collaboration avec Yannick Privat<sup>2</sup> et Mario Sigalotti<sup>6</sup>. Il est présenté en détails dans le chapitre II.

#### I.4.1.1 Motivations et présentation du problème traité

La première contribution que nous allons présenter attaque l'optimisation de la CWS sous un autre angle. Nous regardons le problème d'optimisation posé sur un espace de formes abstrait que nous appellerons les formes admissibles, et nous déduisons à la fois l'existence et l'expression d'un gradient de forme. Cette étape est réalisée sans discrétisation ni paramétrisation. C'est seulement dans un second temps que nous discrétiserons l'espace des formes admissibles afin de résoudre numériquement le problème.

Cette approche présente plusieurs avantages. Tout d'abord, toutes les quantités considérées sont indépendantes de la paramétrisation choisie puisqu'elles sont intrinsèques. Ensuite, optimiser puis discrétiser est généralement plus robuste que l'approche discrétisation puis optimisation. Enfin, notre méthode est compatible avec une très grande variété de choix de surfaces admissibles et de paramétrisations.

Commençons par définir les critères définissant les surfaces admissibles.

- On souhaite que  $S$  soit une surface suffisamment lisse et homéomorphe à un tore.
- Le plasma doit être enveloppé par ce tore et se situer à une distance supérieure à  $d_{min} > 0$ .
- On impose aussi que l'aire de la surface soit inférieure à une constante et que la courbure de la surface soit suffisamment faible. Nous reviendrons sur ce second point dans la Section I.4.1.5.

Notons  $\mathcal{O}_{ad}$  cet ensemble de forme. Nous cherchons donc à résoudre le problème suivant

$$\inf_{S \in \mathcal{O}_{ad}} \inf_{j \in \mathcal{F}_S^0} \|BS(j) - B_T\|_{L^2(P)}^2 + \lambda \|j\|_{\mathcal{F}_S^0}^2. \quad (\text{I.5})$$

#### I.4.1.2 Le gradient de forme

Afin de pouvoir optimiser la surface, nous allons utiliser la notion de dérivée de forme au sens d'Hadamard. Plus précisément, nous utiliserons l'approche de F. Murat et J. Simon présentée dans [MS76a; MS76b]. Pour cela nous considérons des perturbations de l'identité de la forme

$$\text{Id} + \tau \text{ avec } \|\tau\|_{W^{2,\infty}} < 1. \quad (\text{I.6})$$

Dans ce cas,  $(\text{Id} + \tau)S$  est toujours une surface avec régularité  $\mathcal{C}^{1,1}$ . On peut alors calculer la dérivée directionnelle selon la perturbation  $\tau$ ,

$$\frac{d}{dt} C((\text{Id} + t\tau)S)|_{t=0}. \quad (\text{I.7})$$

On dira que  $C$  est dérivable au sens des formes si l'on peut écrire

$$C((\text{Id} + \tau)S) = C(S) + \langle dC(S), \tau \rangle + o(\|\tau\|_{W^{2,\infty}(\mathbb{R}^3, \mathbb{R}^3)}), \quad (\text{I.8})$$

---

6. Laboratoire Jacques-Louis Lions, Inria Paris

et l'on nommera dérivée de forme l'application linéaire  $dC(S)$ . Il est pratique et souvent possible grâce à des théorèmes de structure (cf. [HP18, Section 5.9]) d'obtenir une expression de la forme

$$\langle dC(S), \tau \rangle = \int_S \langle X, \tau \rangle dS. \quad (\text{I.9})$$

Une fois l'expression de  $X$  obtenue, effectuer l'algorithme de la plus profonde descente revient simplement à choisir comme direction de perturbation le champ de vecteurs  $X$ .

Dans le Théorème II.11, nous prouvons la différentiabilité au sens des formes du coût  $C$ . Nous donnons également l'expression de  $X$  dans l'équation (I.9) ci-dessus.

### I.4.1.3 Originalité, difficultés et idées de preuve

Commençons par justifier en quoi notre problème (I.3) se différencie de ceux habituellement traités en optimisation de forme pour lesquels un certain nombre de méthodes classiques ont été développées.

- Le coût  $C(S)$  dépend d'un problème de minimisation posé sur la surface et non d'une équation aux dérivées partielles (EDP) posée sur le domaine entouré par la surface.
- Il n'est *a priori* pas évident de pouvoir étendre un champ vecteur à divergence nulle sur  $S$  vers une surface  $(\text{Id} + \tau)S$  proche de  $S$ .

La première étape consiste alors à construire explicitement un opérateur de  $\mathcal{F}_S^0$  vers  $\mathcal{F}_{(\text{Id} + \tau)S}^0$ . Une fois cet opérateur étudié, on peut ramener le problème de minimisation posé sur  $\mathcal{F}_{(\text{Id} + \tau)S}^0$  vers  $\mathcal{F}_S^0$ .  $\mathcal{F}_S^0$  étant un espace vectoriel, on peut alors utiliser le calcul différentiel. On déduit ainsi la variation du minimiseur  $j_S$  lorsque l'on fait varier  $S$  en utilisant l'équation (I.4) et l'opérateur reliant  $\mathcal{F}_S^0$  vers  $\mathcal{F}_{(\text{Id} + \tau)S}^0$ . Enfin, on peut faire un calcul de variation du coût en injectant (I.4) dans (I.3).

### I.4.1.4 L'implémentation numérique

Grâce à l'expression du gradient de forme, nous avons effectué une implémentation numérique de l'optimisation de forme. Les détails concernant la discrétisation des objets (formes et champs de vecteurs) sont présentés dans la Section II.4. Mentionnons brièvement qu'il est possible de représenter les champs à divergence nulle sur un tore à l'aide de deux champs dits harmoniques et d'un potentiel scalaire que l'on développe alors en série de Fourier tronquée. De plus, au lieu d'approcher le champ magnétique dans le plasma au sens  $L^2(P, \mathbb{R}^3)$ , on se restreint à la composante normale sur la surface du plasma (donc un élément de  $L^2(\partial P, \mathbb{R})$ ). Les justifications de ces choix, largement utilisés dans la littérature, sont fondés sur la décomposition de Hodge. Nous prouvons des résultats les justifiant dans l'appendice II.A.1.

Le logiciel que nous avons développé s'appelle **Stellacode**. L'ensemble des outils nécessaires à son utilisation ainsi que la documentation sont disponibles sous licence libre <sup>7</sup>. Quelques tutoriels permettent de prendre en main la bibliothèque.

Afin de lancer une optimisation avec **Stellacode**, il est nécessaire de fournir au programme un fichier de configuration comportant les informations suivantes :

1. Les coefficients de Fourier permettant de représenter la surface du plasma ainsi que la CWS qui servira d'initialisation.
2. Les coefficients représentant le champ magnétique cible sur le plasma ainsi que deux nombres encodant le courant total poloïdal et toroïdal (voir appendice II.A.1.3). Le paramètre de régularisation de Tychonoff dans (I.3) doit également être donné.

---

7. <https://rrobin.pages.math.cnrs.fr/stellacode/>

3. Les différents paramètres de discrétisation : taille des mailles pour la surface du plasma et la CWS, nombre d'harmoniques dans chaque direction pour le potentiel scalaire. . .
4. Certains choix d'implémentation, qui peuvent être modifiés (taille des chunks, utilisation de cluster, utilisation de cartes graphiques. . .).
5. Les contraintes et leur pénalisation relative.

La gestion des contraintes (courbure, aire et distance au plasma) est implémentée de manière lisse. Concrètement, on fixe un seuil doux, un seuil dur et un paramètre de proportionnalité. Dans le cas de la distance au plasma par exemple (pénalisation des faibles valeurs) on ajoute au coût  $C(S)$  l'intégrale suivante

$$\int_S f(d(x, P)) dS(x), \quad (\text{I.10})$$

où  $d(x, P)$  est la distance au plasma et  $f$  est une fonction  $\mathcal{C}^1(\mathbb{R})$  valant 0 au-dessus du seuil doux et  $+\infty$  en dessous du seuil dur.

Ces coûts liés aux contraintes sont ajoutés au coût du problème inverse régularisé (I.3). On calcule alors le gradient de forme lié au problème inverse à travers le calcul du champ  $X$  dans l'équation (I.9) et on y ajoute le gradient de forme des coûts représentant les contraintes.

Ce gradient de forme se représente alors numériquement comme un gradient sur l'espace des coefficients de Fourier de la CWS. On effectue un algorithme d'optimisation en boîte noire (comme l'algorithme BFGS par exemple) en fournissant une fonction de coût et son gradient.

Le calcul du gradient de forme est optimisé numériquement : le code est parallélisé et utilise des bibliothèques de calcul scientifique fortement optimisées et adaptées au calcul haute performance. Les simulations ont été réalisées sur un cluster de calcul du Laboratoire Jacques-Louis Lions.

En présence de pénalisation de la courbure, on observe des résultats très satisfaisants. Les Figures I.5 et I.6 représentent respectivement la CWS qui a été utilisée pour la conception de NCSX ([Zar+01]) et celle optimisée par notre algorithme. Dans ce cas, nous avons obtenu une réduction d'un facteur quatre de l'erreur d'approximation du champ magnétique cible et une réduction d'un tiers sur la norme  $L^2$  du courant électrique.

#### I.4.1.5 Courbure et retour sur la question de l'existence

Revenons à présent sur ce que nous entendons par courbure et sur la question de l'existence d'un minimiseur au problème d'optimisation de forme (I.5). Rappelons qu'il est très facile en optimisation de forme de poser des problèmes qui n'admettent pas de minimiseur. Par exemple, essayer de maximiser le périmètre d'une courbe fermée en dimension deux tout en gardant l'aire délimitée constante (un problème isopérimétrique inverse) : en prenant une surface toujours plus irrégulière, on peut construire une suite de courbes ayant un périmètre arbitrairement grand tout en délimitant un volume fini.

Les simulations numériques du problème d'optimisation (I.5) avec une distance minimale au plasma ainsi qu'une aire maximale imposée sur les formes admissibles ont donné lieu à des formes très exotiques. Dans la Figure I.7 on observe des piques très fines et très longues. Si de telles déformations semblent bien réduire le coût du problème inverse, elles posent un double problème pratique et théorique.

- D'un point de vue pratique pour un numéricien, notre discrétisation du problème commencent à ne plus être pertinente. Il faudrait fortement réduire la taille du maillage au voisinage de ces points.
- D'un point de vue pratique pour un ingénieur, cette surface est inconstructible.

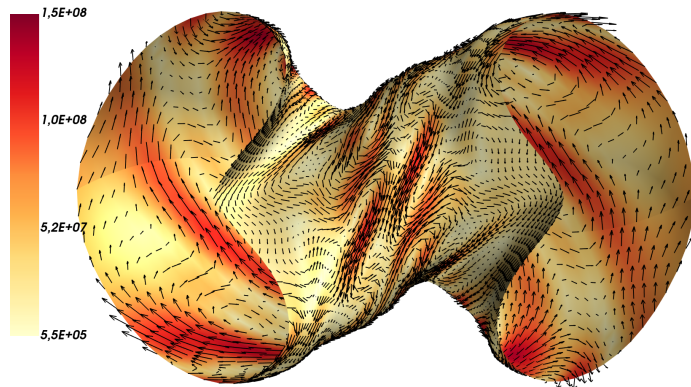


FIGURE I.5 – Courant surfacique optimal pour la CWS utilisée pour la conception de NCSX. La CWS présente une symétrie discrète sous la rotation d'angle  $\frac{2\pi}{3}$  selon l'axe  $Oz$ .

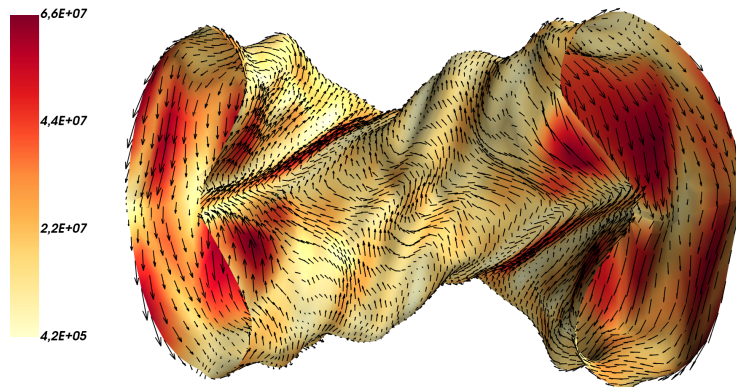


FIGURE I.6 – Courant surfacique optimal sur une surface optimisée par **Stellacode**. L'échelle montre une réduction d'un facteur trois de l'intensité maximale du courant avec la situation initiale.





FIGURE I.7 – Une CWS obtenue numériquement dans une optimisation ne pénalisant pas la courbure.

- D'un point de vue théorique, effectuer une descente de gradient sur un problème qui n'admet pas de minimiseur est une idée plutôt curieuse.

Une solution pour obtenir l'existence d'un minimiseur est de réduire la classe des formes admissibles en imposant une contrainte supplémentaire. C'est ce que nous avons fait en imposant à nos surfaces une condition dite de *reach*, ou de boule uniforme. Cette condition impose notamment une borne inférieure sur les rayons de courbure de la surface.

La méthode que nous avons développée pour résoudre ce problème d'existence se généralise assez facilement à d'autres types de problèmes d'optimisation de forme comprenant un coût intégral sur une hypersurface. Ces questions d'existences sont l'objet du chapitre III.

## I.4.2 Quelques théorèmes d'existence en optimisation de forme

Ce travail a été effectué en collaboration avec Yannick Privat<sup>2</sup> et Mario Sigalotti<sup>6</sup>. Il est présenté en détails dans la Section II.2.2 et dans le Chapitre III.

Dans cette partie nous allons nous intéresser à la question d'existence de minimiseurs pour le problème (I.5) ainsi que pour des problèmes d'optimisation sur des hypersurfaces suffisamment régulières de la forme

$$\inf_{\Omega \in \mathcal{E}_{\text{ad}}} \int_{\partial\Omega} j(x, \nu_{\partial\Omega}(x), u_{\partial\Omega}(x)) d\mu_{\partial\Omega}(x), \quad (\text{I.11})$$

où  $\nu_{\partial\Omega}(x)$  est la normale extérieure à  $\partial\Omega$  au point  $x$  et  $u_{\partial\Omega}$  peut être une quantité purement

géométrique, comme la courbure moyenne au point  $x$  ou la solution d'une EDP dépendant de  $\Omega$ .

### I.4.2.1 Méthode directe du calcul des variations en optimisation de forme

La méthode classique pour prouver l'existence d'un minimiseur à un problème d'optimisation est la méthode directe du calcul des variations. Résumons-la brièvement.

Fixons un ensemble admissible  $A$  et d'une fonctionnelle  $f : A \rightarrow \mathbb{R}$  bornée inférieure. On cherche à prouver qu'il existe  $x^* \in A$  tel que

$$f(x^*) = \inf_{x \in A} f(x). \quad (\text{I.12})$$

Prenons alors une suite minimisante  $(x_n)_{n \in \mathbb{N}} \in A$  telle que  $f(x_n) \rightarrow \inf_{x \in A} f(x)$ . Si l'on dispose d'une topologie sur  $A$  telle que

1.  $(x_n)_n$  est séquentiellement compacte,
2.  $f$  est semi-continue inférieurement,

alors, à une sous-suite près, on a la convergence  $x_n \rightarrow x^*$  et  $f(x^*) = \inf_{x \in A} f(x)$ . L'égalité (I.12) est donc prouvée. Toute la difficulté est de trouver une topologie satisfaisant à la fois les Points 1 et 2 qui sont des propriétés antagonistes. Puisque, plus une topologie est fine, plus il y a de fonctions continues mais moins il y a de compacts.

Par ailleurs, le choix de l'ensemble  $A$  est une difficulté à part entière dans le cadre de l'optimisation de forme.

Présentons à présent la convergence au sens des fonctions caractéristiques, qui est une topologie classique dans ce contexte. Nous pouvons représenter un ensemble par sa fonction caractéristique dans  $L^\infty(\mathbb{R}^d)$ . Cette opération implique déjà de quotienter notre espace de formes, puisque deux ensembles différant d'un ensemble de mesure nulle auront la même fonction caractéristique. La convergence de  $\Omega_n \rightarrow \Omega$  au sens des fonctions caractéristiques est définie comme la convergence  $\mathbb{1}_{\Omega_n} \xrightarrow[n \rightarrow \infty]{L^1_{loc}} \mathbb{1}_{\Omega}$ . On peut montrer que si la suite  $(\Omega_n)_n$  est à périmètres<sup>8</sup> uniformément bornés, on peut extraire une sous-suite qui converge au sens des fonctions caractéristique vers  $\Omega_\infty$  (ce qui prouve le Point 1). Cependant, ce cadre n'est pas adapté à nos fonctionnelles, posées sur des hypersurfaces. Par exemple, comme nous venons de le mentionner, ajouter un ensemble de mesure nulle ne change pas la fonction caractéristique dans  $L^\infty$  (alors que nous faisons des intégrales sur des hypersurfaces, donc de mesure nulle). Par ailleurs,  $\partial\Omega_\infty$  n'aura a priori aucune régularité, alors que nous cherchons dans un souci de modélisation des surfaces relativement régulières et réalisables physiquement.

Nous renvoyons à [HP18, Chapitre 2] pour la preuve des propositions énoncées dans le paragraphe précédent, ainsi qu'une présentation des topologies usuelles. D'un point de vue informel, les topologies usuelles sont adaptées à des problèmes que nous qualifierons plutôt de "volumiques". Citons deux cas d'étude classiques : les problèmes isopérimétriques et les minimisations de l'énergie de Dirichlet sous contrainte de volume. Dans les deux cas, les convergences que l'on considère sont alors trop faibles pour s'assurer de définir un coût comme celui de (I.3) sur l'objet limite ou pour espérer la semi-continuité dudit coût.

Notre approche va donc consister à réduire radicalement l'espace des formes admissibles en ne considérant que des formes que nous pouvons informellement qualifier d'uniformément  $\mathcal{C}^{1,1}$ .

---

8. au sens de De Giorgi, ce qui est équivalent à la variation totale de la fonction caractéristique

### I.4.2.2 Condition de *reach* et régularité

La condition de régularité sur les formes que l'on va imposer regroupe deux notions presque équivalentes. La condition de boule uniforme sur l'hypersurface  $\partial\Omega$  qui se formule comme suit

$$\forall x \in \partial\Omega, \exists d_x \in \mathbb{R}^n \mid \|d_x\|_{\mathbb{R}^n} = 1, B_h(x - hd_x) \subset \Omega \text{ et } B_h(x + hd_x) \subset \mathbb{R}^n \setminus \Omega, \quad (\text{I.13})$$

pour un certain  $h > 0$  mesurant l'uniforme régularité de la surface. La condition de *reach* uniforme est fondée sur la notion de distance signée, que l'on note  $b_\Omega$ . Cette fonction est définie via la fonction distance  $d_\Omega$  par les relations suivantes :

$$d_\Omega(x) = \inf_{y \in \Omega} \|x - y\| \quad \text{et} \quad b_\Omega(x) = d_\Omega(x) - d_{\mathbb{R}^d \setminus \Omega}(x). \quad (\text{I.14})$$

Introduisons également le voisinage tubulaire de  $\partial\Omega$  par la formule

$$U_h(\partial\Omega) = \{x \in \mathbb{R}^d \mid |b_\Omega(x)| \leq h\}. \quad (\text{I.15})$$

Le *reach* de  $\partial\Omega$  est donné par

$$\text{Reach}(\partial\Omega) = \sup\{h > 0 \mid b_\Omega \text{ est différentiable sur } U_h(\partial\Omega) \setminus \partial\Omega\}. \quad (\text{I.16})$$

Cette notion a été introduite par H. Federer dans [Fed69]. De nombreux résultats ont été obtenus par M. C. Delfour et J.-P. Zolesio et sont compilés dans [DZ11, Chapitre 7]. On y trouve notamment le fait que sous la condition que la mesure de  $\partial\Omega$  soit nulle et que  $\Omega$  soit compact, la condition de *reach* strictement positive est équivalente à la régularité  $\mathcal{C}^{1,1}$  de l'hypersurface. Par ailleurs la distance signée est reliée directement à la géométrie de la surface via de nombreuses propriétés :

- Pour  $x \in \partial\Omega$ ,  $\nabla b_\Omega(x)$  est la normale extérieure.
- Dans un voisinage de  $\partial\Omega$ , le projecteur orthogonal sur l'hypersurface s'écrit  $p(x) = x - b_\Omega(x)\nabla b_\Omega(x)$ .
- La courbure moyenne est la trace de la hessienne de  $b_\Omega$  :  $\text{Tr } \nabla^2 b_\Omega$ .

### I.4.2.3 Résultats existants

Fixons  $D$  un ensemble compact,  $r_0 > 0$  et considérons l'ensemble admissible

$$\mathcal{O}_{r_0} = \{\Omega \subset D \mid \Omega \text{ est fermé, } \text{Reach}(\partial\Omega) \geq r_0, \partial\Omega \text{ est une } (d-1)\text{-sous-variété}\}. \quad (\text{I.17})$$

La question de l'existence du problème d'optimisation de forme (I.11) avec l'ensemble admissible (I.17) a été considérée par B. Guo et al. dans [GY13] puis généralisée par J. Dalphin dans [Dal18] et [Dal20]. La stratégie utilisée fait appel à la méthode directe du calcul des variations et se fonde sur un usage très technique de cartes locales afin de passer à la limite.

Afin de résoudre le problème (I.5), nous avons utilisé une méthode plus directe utilisant la distance signée et les voisinages tubulaires comme outils permettant d'assurer la semi-continuité inférieure des fonctionnelles, cette idée ayant auparavant été développée dans [Del00]. Nous avons ensuite décidé de mettre au point un cadre permettant d'obtenir des preuves plus directes que celles obtenues dans les papiers susmentionnés.

#### I.4.2.4 Existence d'un minimiseur pour le problème (I.5)

Commençons par introduire une notion de convergence sur  $\mathcal{O}_{r_0}$ , que nous avons nommée la *R-convergence*. Soit  $(\Omega_n)_n \in (\mathcal{O}_{r_0})^{\mathbb{N}}$ , on dira que  $(\Omega_n)_n$  *R-converge* vers  $\Omega_\infty \in \mathcal{O}_{r_0}$  si

$$b_{\Omega_n} \rightarrow b_{\Omega_\infty} \quad \begin{cases} \text{dans } \mathcal{C}(\overline{D}), \\ \text{dans } \mathcal{C}^{1,\alpha}(U_r(\partial\Omega_\infty)), \forall r < r_0, \forall \alpha \in [0, 1), \\ \text{faible-* dans } W^{2,\infty}(U_r(\partial\Omega_\infty)), \forall r < r_0. \end{cases} \quad (\text{I.18})$$

On peut montrer (cf. Chapitre III) que  $\mathcal{O}_{r_0}$  muni de la *R-convergence* est séquentiellement compact.

Par ailleurs, les contraintes de distance au plasma

$$d(\partial\Omega, P) = \inf_{x \in \partial\Omega, y \in P} |x - y| = \inf_{x \in S} d_P(x) \geq \delta, \quad (\text{I.19})$$

d'aire maximale

$$\mathcal{H}^2(\partial\Omega) \leq P_{\max}, \quad (\text{I.20})$$

ou encore d'isotopie à un tore sont continues pour la *R-convergence* (voir Lemme III.4). Ainsi, l'ensemble des formes admissibles formées des éléments de  $\mathcal{O}_{r_0}$  satisfaisant (I.19), (I.20), et une condition d'isotopie forment un ensemble séquentiellement compact.

Il reste donc à prouver que le coût (I.3) est semi-continue inférieurement pour la *R-convergence*.

On prend alors une suite minimisante  $\Omega_n \xrightarrow{R} \Omega_\infty$  et on étend les minimiseurs sur un voisinage tubulaire de  $\partial\Omega_n$ . Cela définit une suite de fonctions sur  $\partial\Omega_\infty$  et on montre qu'à extraction près, cette suite converge vers un champ de vecteurs dans  $\mathcal{F}_{\partial\Omega_\infty}^0$  qui est le minimiseur du problème (I.3) posé sur  $\partial\Omega_\infty$ .

À l'aide des différentes convergences, on en déduit l'existence d'un minimiseur pour (I.5).

#### I.4.2.5 Nouvelles preuves d'existence pour certaines fonctionnelles

Comme nous le mentionnons, la méthode développée dans la Section II.2.2 est facilement généralisable aux cas traités par [GY13; Dal18; Dal20]. Le Chapitre III contient des preuves concises de semi-continuité inférieure de nombreuses fonctionnelles faisant intervenir des hypersurfaces satisfaisant une contrainte de *reach* strictement positif. Ces résultats ne sont donc pas originaux. Cependant, nous pensons que le traitement systématique obtenu par l'utilisation de la distance signée et du voisinage tubulaire a le mérite de concilier clarté et concision. Par ailleurs, afin d'illustrer la puissance des outils développés, nous avons traité le cas original d'une équation elliptique posée sur la surface.

#### I.4.2.6 Cas d'une EDP elliptique sur la frontière

Soit  $f \in \mathcal{C}^0(D)$  et  $\Omega \in \mathcal{O}_{r_0}$ . On définit  $v_{\partial\Omega}$  la solution de

$$\Delta_{\partial\Omega} v_{\partial\Omega}(x) = f(x) \quad \text{dans } \partial\Omega, \quad (\text{I.21})$$

où  $\Delta_{\partial\Omega}$  est le Laplacien Beltrami sur  $\partial\Omega$ . Comme la surface n'est pas  $\mathcal{C}^2$ , on adopte une définition variationnelle comme minimiseur de l'énergie de Dirichlet. L'existence et l'unicité des solutions

s'ensuivent. On s'intéresse alors à la fonctionnelle

$$F(\Omega) = \int_{\partial\Omega} j(x, \nu(x), v_{\partial\Omega}(x), \nabla_{\partial\Omega} v_{\partial\Omega}(x)) d\mu_{\partial\Omega}(x),$$

où  $j : \mathbb{R}^d \times \mathcal{S}^{d-1} \times \mathbb{R} \times \mathbb{R}^d \rightarrow \mathbb{R}$  est continue. Nous prouvons alors la semi-continuité inférieure de  $F$  dans le Théorème III.10 et en déduisons l'existence d'un minimiseur au problème

$$\inf_{\Omega \in \mathcal{O}_{r_0}} F(\Omega).$$

### I.4.3 Forces de Laplace s'exerçant sur une surface de courant

Ce travail a été effectué en collaboration avec Francesco Volpe<sup>9</sup>. Il est présenté en détails dans le chapitre IV. Nous mentionnons que ce travail a été publié dans le journal de physique *Nuclear Fusion*.

#### I.4.3.1 Contexte physique

Si nous retournons au problème inverse (I.3), le choix de pénaliser la norme  $L^2$  de l'intensité électrique de la densité surfacique de courant est assez naturel d'un point de vue mathématique. Cependant, sa pertinence physique peut être remise en question.

Renaissance Fusion souhaite construire un prototype compact de stellarator. Pour maintenir un confinement équivalent, il faut utiliser un champ magnétique plus élevé. Cela nécessite des courants électriques plus importants, et implique que les forces de Laplace augmentent de manière quadratique. Il devient alors important de s'assurer que ces forces ne soient pas trop intenses.

Rappelons brièvement la physique en jeu. Dans un conducteur soumis à un champ magnétique, les électrons sont soumis à la force de Lorentz dont la composante magnétique est  $q\vec{v} \wedge \vec{B}$ . En régime stationnaire, cela induit un champ électrique appelé champ de Hall compensant cette force. Il en résulte une force s'exerçant sur un volume infinitésimal  $dV$  s'exprimant comme suit

$$d\vec{F} = (\vec{j} \wedge \vec{B}) dV. \quad (\text{I.22})$$

Pour un conducteur filiforme, on utilise souvent l'expression

$$d\vec{F} = I d\vec{l} \wedge \vec{B}, \quad (\text{I.23})$$

où  $I$  est l'intensité parcourant le fil. Derrière cette expression se cache déjà une ambiguïté. Qu'est-ce que  $\vec{B}$ ? Rappelons que la norme du champ magnétique produit par un fil explose au voisinage de celui-ci. On admet généralement que la contribution principale est donnée par le champ  $\vec{B}$  extérieur, c'est-à-dire généré par le reste du système.

Qu'en est-il dans le cas d'une surface avec un courant surfacique  $\vec{j}$ ? Le champ magnétique généré en tout point est donné par l'opérateur de Biot et Savart dont l'expression se trouve dans l'équation (I.1). On observe que le champ magnétique ainsi obtenu n'est pas continu sur la surface. D'un point de vue mathématique, cela se lit dans la non-intégrabilité du "noyau"

$$K(x, y) = \frac{x - y}{\|x - y\|^3}. \quad (\text{I.24})$$

---

9. Renaissance Fusion

En effet, pour  $x \in S$ ,  $K(x, \cdot)$  n'est pas dans  $L^1(S)$  puisque  $\frac{1}{\|y\|^2}$  n'est pas intégrable au voisinage de 0 dans  $\mathbb{R}^2$ . Mais alors, comment définir les forces de Laplace s'exerçant sur une surface de courant ? C'est à cette question, puis à celle de la minimisation de cette force, que nous nous sommes attelés dans le Chapitre IV.

### I.4.3.2 Modélisation

Une dérivation physique rigoureuse de la force de Laplace nécessite de retourner à des objets volumiques et passer à la limite d'une épaisseur négligeable. Cette approche est présentée dans la Section IV.2.4. Afin de simplifier les calculs, nous présentons un cas légèrement simplifié. Considérons pour ce faire, pour  $\varepsilon > 0$  donné,

$$\mathbf{L}_\varepsilon(\vec{j}_1, \vec{j}_2)(y) = \frac{1}{2} \vec{j}_1(y) \times [BS(\vec{j}_2)(y + \varepsilon\nu(y)) + BS(\vec{j}_2)(y - \varepsilon\nu(y))], \quad y \in S. \quad (\text{I.25})$$

$\mathbf{L}_\varepsilon$  est donc une application bilinéaire représentant le produit vectoriel entre  $\vec{j}_1$  et la demi-somme du champ magnétique généré par  $\vec{j}_2$  en  $y \pm \varepsilon\nu(y)$ . Ainsi, nous utilisons le champ magnétique légèrement à l'extérieur de la surface. Une définition raisonnable des forces de Laplace (et équivalente à celle découlant de l'approche volumique) exercée par un courant  $\vec{j}$  est alors

$$L(\vec{j}, \vec{j})(y) = \lim_{\varepsilon \rightarrow 0} \mathbf{L}_\varepsilon(\vec{j}, \vec{j})(y). \quad (\text{I.26})$$

Nous avons alors prouvé que  $\mathbf{L}_\varepsilon$  en tant qu'opérateur de  $H^1 \times H^1$  vers  $L^p(S, \mathbb{R}^3)$ ,  $p < \infty$ , convergeait vers un opérateur limite  $L$ . Le théorème IV.2 énonce cette convergence et donne l'expression explicite de la limite. Mentionnons brièvement que dans l'approche volumique, il existe une contribution des forces de Laplace qui tend à élargir la surface, mais qui n'est pas capturée par  $L$ , puisque la moyenne de cette force sur la direction normale est nulle. L'expression de  $L$  est alors obtenue en faisant une intégration par parties sur les composantes tangentielles et en utilisant des inégalités fonctionnelles pour contrôler les contributions qui se moyennent.

### I.4.3.3 Optimisation numérique

Une fois obtenue une expression faisant office de définition pour les forces de Laplace, on cherche à minimiser ces dernières numériquement.

Pour cela, on résout une généralisation de (I.3) sous la forme

$$\inf_{j \in E} \|BS(j) - B_T\|_{L^2(P)}^2 + \lambda \|j\|_E^2 + \gamma f(\|L(j, j)\|), \quad (\text{I.27})$$

$E$  étant un sous espace de  $\mathcal{F}_S^0$  et  $f$  étant soit une pénalisation  $L^p$ , soit une pénalisation similaire à celle utilisé dans (I.10). Nous reviendrons sur les questions d'existences dans la Section I.4.3.4.

Numériquement, nous avons observé que si l'on pénalise les forces de Laplace avec une norme  $L^2$ , le résultat est très comparable à une pénalisation  $L^2$  de la norme de la densité surfacique de courant. Cependant, si on considère des coûts pénalisant les valeurs extrêmes, on peut alors fortement réduire les valeurs maximales. Ceci a d'autant plus de sens d'un point de vue ingénierie que les problèmes sont liés aux valeurs importantes des forces de Laplace s'exerçant sur la structure et non les forces moyennes.

Pour attaquer le problème numériquement, nous avons développé un code reproduisant le comportement du code Fortran **REGCOIL** de M. Landreman [Lan17] résolvant (I.3) et nous avons implémenté l'optimisation de la densité de courant. Pour cela nous avons calculé le gradient

du coût (I.27) et utilisé un algorithme de type quasi-Newton. Nous renvoyons à la Section IV.4.3 pour une présentation détaillée des résultats.

#### I.4.3.4 Retour sur la question d'existence

Notons tout d'abord qu'il est nécessaire de supposer de la régularité sur  $\vec{j}$  afin de pouvoir définir  $L(\vec{j}, \vec{j})$ . Il est donc naturel de régulariser (I.27) avec une pénalisation  $H^1$  sur  $\vec{j}$ . Nous avons observé que l'utilisation de  $H^1$  (et non  $L^2$ ) a peu d'impact numériquement sur les forces et la qualité de la reproduction du champ magnétique. Cependant, sans pénalisation  $H^1$ , on obtient une densité de courant très oscillante (voir Figure IV.7) et donc moins intéressante physiquement.

Par ailleurs, la pénalisation que nous avons implémentée pour les forces de Laplace est un coût intégral similaire à (I.10) dépendant de la norme des forces et consistant en un seuil doux (aucune pénalisation sous une certaine valeur) et un seuil dur (pénalisation infinie au-dessus du seuil dur). Ceci n'est pas justifié théoriquement puisque l'opérateur  $L$  n'est pas à valeur dans  $L^\infty(S, \mathbb{R}^3)$  mais seulement  $L^p(S, \mathbb{R}^3)$ ,  $p < \infty$  et jusqu'à l'espace de Orlicz  $L^\varphi(S, \mathbb{R}^3)$ , avec  $\varphi(t) = e^{t^2} - 1$ . Ce choix est motivé par des aspects pratiques (seuils plus faciles à fixer).

Afin de répondre à cette faiblesse, nous prouvons dans la Section IV.B l'existence d'un minimiseur au problème (I.27) pour  $E = H^1$  et une pénalisation de la norme  $L^p(S, \mathbb{R}^3)$ ,  $p < \infty$ , des forces de Laplace.

### I.4.4 Quelques problèmes ouverts et perspectives

Comme nous l'avons illustré dans cette partie dédiée à la fusion nucléaire, l'optimisation est un élément clef dans la conception des stellarators. Ce thème fait l'objet de nombreuses recherches menées par des physiciens et il nous a semblé qu'une plus grande implication de la communauté des mathématiques appliquées serait très bénéfique pour relever ce grand défi qu'est la fusion nucléaire contrôlée.

Par ailleurs, l'étude des stellarators fait apparaître des aspects subtils et profonds des mathématiques contemporaines. Les formes harmoniques et la théorie de Hodge sont en effet des éléments essentiels pour bien comprendre les solutions des équations de Maxwell dans le vide (divergence et rotationnel nul) sur des domaines non simplement connexes comme le tore. L'étude de la dynamique le long d'une ligne de champ magnétique est reliée à la mécanique hamiltonienne et la stabilité des surfaces dites de flux est reliée à la théorie KAM. Enfin, l'étude de la dynamique du plasma fait appel à la magnétohydrodynamique. Les études de stabilité et de stabilisation sont des domaines actifs et complexes à l'interface avec les équations aux dérivées partielles.

#### I.4.4.1 Projet en cours avec Renaissance Fusion

Commençons par présenter quelques projets à court terme avec la start-up renaissance fusion.

**Optimisation sur de nouvelles surfaces** D'un point de vue conception, le *reach* d'une surface traduit mal les difficultés relatives à la construction cette surface. Renaissance Fusion souhaite par conséquent optimiser la CWS au sein de familles restreintes de formes admissibles sélectionnées pour être facilement manufacturables. L'adaptation de notre logiciel **Stellacode** est en cours au sein de Renaissance Fusion afin de répondre à ce besoin.

**Forces de Laplace et optimisation de forme** Une autre amélioration importante de **Stellacode** porte sur l'implémentation de l'optimisation de la CWS avec un coût prenant en compte les

forces de Laplace. L'optimisation numérique est alors plus complexe à mettre en place puisque résoudre le problème (I.27) nécessite lui-même une optimisation et une expression explicite, comme (I.4), n'est pas disponible. La difficulté est ainsi principalement concentrée sur le développement d'un code numériquement efficace et adapté au calcul haute performance.

**Injection de courant et décomposition de Hodge** Les densités de courant considérées dans (I.3) sont à divergence (surfactive) nulle car les injections de courant sur la surface sont situées au même emplacement que les prélèvements. En autorisant les densités de courant à ne plus être à divergence nulle, c'est-à-dire en raccordant la surface à des fils extérieurs, on élargit l'ensemble des champs de vecteurs admissibles. En utilisant la décomposition de Hodge, on peut caractériser simplement la partie à divergence non nulle et ainsi optimiser les positions des *sources* et *puits* de courant pour une généralisation du problème I.3. Ce travail réalisé avec Erol Balkovic<sup>9</sup>, Julien Fausty<sup>9</sup> et Francesco Volpe s'inscrit dans une logique exploratoire de dépassement de l'utilisation d'une CWS classique.

#### I.4.4.2 Optimisation du plasma

Un second aspect sur lequel la collaboration StellaCage porte son attention est l'optimisation de la forme du plasma et du champ magnétique cible dans ce plasma. Les critères utilisés pour l'étude de la stabilité des équilibres MHD sont complexes et une analyse mathématique à l'interface pourrait permettre des avancées dans ce domaine. Rappelons aussi que dans un domaine toroïdal donné, il existe une unique droite vectorielle dans l'espace des formes différentielles à divergence et rotationnelle nulle tout en étant tangente au bord du domaine (cf. [CDG01]). Cette droite correspond aux formes harmoniques sur le domaine. L'étude de ces formes est donc profondément reliée aux solutions des équations de Maxwell dans le vide et offre des perspectives de recherche intéressantes.

En conclusion, de nombreux axes de recherches restent à explorer dans ce domaine passionnant et hautement stratégique à l'interface entre physique, sciences de l'ingénieur, l'analyse numérique et les mathématiques.

## I.5 Contrôle de systèmes quantiques : motivations et outils

Introduisons à présent la deuxième partie de cette thèse consacrée au contrôle de systèmes quantiques.

### I.5.1 Motivations physiques générales

Depuis sa découverte au début du siècle dernier, la mécanique quantique a eu des impacts technologiques majeurs, que l'on regroupe généralement sous le nom de première révolution quantique. Par exemple les transistors développés dès le début de la deuxième moitié du siècle dernier, le laser mis au point expérimentalement en 1960 par T. Maiman, etc.

Après ces découvertes qui ont permis le développement de l'informatique moderne, une deuxième révolution est attendue avec pour figure de proue l'ordinateur quantique. Les prémices sont visibles et des premières applications industrielles sont déjà opérationnelles. Citons notamment les capteurs quantiques dont les *Superconducting Quantum Interference Devices* (SQUID) permettent déjà de mesurer des champs magnétiques de l'ordre de  $10^{-14}$  T, ou encore les détecteurs d'ondes gravitationnelles, qui utilisent les états compressés du vide. L'ingénierie quantique est en développement très actif et représente un enjeu stratégique majeur pour de nombreux



domaines : information quantique, communications quantiques, simulations quantiques, capteurs quantiques. . .

Le contrôle des systèmes quantiques est un élément clef dans le développement de ces nouvelles technologies. Notons que c'est un domaine extrêmement large en raison de :

- La nature des systèmes étudiés : atomes, molécules, matériaux (semi-conducteurs, supra-conducteurs, . . .), systèmes biologiques. . .
- La variété des applications : résonance magnétique et spectroscopie, contrôle et catalyse de réactions chimiques, technologies quantiques. . .
- la diversité des modèles : systèmes quantiques ouverts ou fermés, dimension finie ou infinie, avec ou sans opérations de mesures. . .

L'action sur le système quantique est généralement effectuée par couplage avec un champ électromagnétique externe que l'on contrôle. L'étude du contrôle quantique est ainsi celle de la manipulation précise et robuste des états quantiques.

Le contrôle de système quantique est ainsi un domaine émergent en mathématiques appliquées [DA107 ; BS12]. Une première question est celle de la **contrôlabilité**, c'est-à-dire la description des états atteignables par la dynamique du système. Nous étudierons la question de la contrôlabilité dans un cadre présentant de la dispersion sur les paramètres du système. Le design des contrôles, c'est-à-dire comment implémenter efficacement un contrôle permettant d'effectuer une action donnée sera l'objet d'une seconde contribution. Pour cela, nous utiliserons des outils de **contrôle optimal** qui permettent de caractériser l'optimalité des stratégies de contrôles en boucle ouverte.

La stabilisation d'un système quantique, ainsi que le contrôle en boucle fermée, sont également des questions très importantes et stratégiques. Le prix Nobel en 2012 de S. Haroche pour ses expériences illustre l'enjeu théorique, expérimental et technologique de ces questions.

Plusieurs difficultés apparaissent lorsque l'on cherche à stabiliser un système quantique. Premièrement, le flot d'une équation de Schrödinger, modélisant l'évolution d'un système quantique fermé, est un opérateur unitaire ; on ne peut donc pas créer de point fixe attractif. En utilisant des systèmes quantiques ouverts, par exemple modélisés une équation maîtresse de type Lindblad (cf. [Lin76]), il devient possible de stabiliser des systèmes grâce à de la dissipation. C'est le cas par exemple pour les *cat qubits* [Mir+14]. Cette stratégie appartient au *quantum feedback*, car l'évolution du système reste entièrement déterministe et déterminée par la mécanique quantique, sans faire appel à des mesures du système.

En effet, l'utilisation de mesures quantiques afin d'observer le système est considérablement différent de l'observation des systèmes physiques classiques qui est étudiée dans la théorie de l'observabilité, développée notamment par R. E. Kalman [Kal63]. En effet, les mesures quantiques perturbent le système de manière stochastique. Nous renvoyons par exemple à [WM93] pour une étude de *feedback* grâce à l'utilisation de mesures quantiques.

Il existe ainsi de nombreux modèles et équations sur lesquels des questions de contrôlabilité, stabilisation et design optimale peuvent être posées. Nous nous restreindrons dans ce manuscrit aux systèmes quantiques fermés sans opérations de mesures, régis par la célèbre équation de Schrödinger.

## I.5.2 Formalisation mathématique

### I.5.2.1 Systèmes quantiques fermés

Le formalisme de la mécanique quantique des systèmes fermés est désormais bien établi et se décrit avec l'aide de l'algèbre linéaire. Les états d'un système quantique sont représentés par des éléments de la sphère unité d'un  $\mathbb{C}$ -espace de Hilbert, noté  $\mathcal{H}$ . L'Hamiltonien  $H$  est un opérateur

autoadjoint sur  $\mathcal{H}$  qui décrit l'évolution du système quantique via l'équation de Schrödinger

$$i\partial_t|\psi\rangle = H|\psi\rangle. \quad (\text{S})$$

Les quantités physiques mesurables sont appelés les **observables**. Par exemple,  $H$  est l'observable représentant l'énergie du système donné. On appelle alors énergie moyenne de l'état  $|\psi\rangle$  la quantité  $\langle\psi|H|\psi\rangle$ , où  $\langle\psi|$  est la forme linéaire associée<sup>10</sup> à  $|\psi\rangle$ .

**Remarque I.1.** *Notons dès à présent que la mécanique quantique ainsi décrite présente une symétrie naturelle de jauge correspondant à un changement de phase globale. En d'autres termes, en multipliant les états par un nombre complexe de module 1, on ne change la physique du système. Il est ainsi rigoureux de formuler la dynamique non pas sur la sphère unité d'un espace de Hilbert mais sur son espace projectif.*

Lorsque l'Hamiltonien  $H$  ne dépend pas du temps (et pour simplifier, lorsque son spectre est discret), le théorème spectral permet de réduire complètement la dynamique. Soient  $(|\psi_n\rangle)_n$  les vecteurs propres normalisés (que l'on appelle aussi *états propres*) et  $(\lambda_n)_n$  les valeurs propres associées. La solution de l'équation d'évolution (S) est donnée par

$$|\psi(t)\rangle = \sum_n \langle\psi_n|\psi_0\rangle |\psi_n\rangle e^{-i\lambda_n t}, \quad (\text{I.28})$$

avec  $|\psi_0\rangle$  la condition initiale. En particulier d'après la Remarque I.1, un état propre est en quelque sorte un état stationnaire de l'équation de Schrödinger puisque son évolution temporelle se réduit à

$$|\psi(t)\rangle = e^{-i\lambda_n t} |\psi_n\rangle.$$

### I.5.2.2 Contrôle des systèmes quantiques fermés

Le contrôle (affine) d'un système quantique se formule alors comme suit : on se donne  $H_0$  un opérateur autoadjoint sur un  $\mathbb{C}$ -Hilbert  $\mathcal{H}$  et des opérateurs de contrôle  $(H_j)_{1 \leq j \leq m}$  également autoadjoints<sup>11</sup>. On se fixe enfin des domaines  $U_j \subset \mathbb{R}$  pour les contrôles et on étudie l'équation de Schrödinger

$$\begin{cases} i\partial_t|\psi\rangle &= (H_0 + \sum_j u_j(t)H_j)|\psi\rangle, \\ |\psi(0)\rangle &= |\psi_0\rangle, \\ u_j &\in L^\infty(0, T; U_j), \end{cases} \quad (\text{I.29})$$

où  $|\psi_0\rangle$  est de norme 1. L'équation (I.29) est dite *bilinéaire*, car elle est à la fois linéaire par rapport à l'état et affine par rapport aux contrôles. Ce dernier point peut être justifié physiquement, par exemple, dans le régime de l'approximation dipolaire. On notera qu'indépendamment des valeurs des  $(u_j)_j$ ,  $|\psi(t)\rangle$  reste sur la sphère unité de  $\mathcal{H}$ . Par conséquent, la théorie du contrôle linéaire n'est pas directement applicable pour répondre aux questions de contrôlabilité et/ou de contrôle optimal. L'espace atteignable ne sera jamais plus grand que la sphère unité de  $\mathcal{H}$  et on dira que le système (I.29) est *exactement contrôlable* si l'espace atteignable est la sphère unité de  $\mathcal{H}$  quotientée par une phase globale<sup>12</sup>.

10. Il s'agit de la notation Bra-Ket communément utilisée par la communauté physicienne et qui revient à identifier formes linéaires et vecteurs via la forme hilbertienne

11. Si les opérateurs de contrôle ne sont pas bornés, des hypothèses sur les domaines des opérateurs sont nécessaires.

12. c'est-à-dire que l'on a contrôlabilité dans l'espace projectif.

Nous attirons l'attention du lecteur sur le fait que de nombreuses difficultés peuvent apparaître dans l'étude de (I.29) lorsque l'espace de Hilbert considéré est de dimension infinie. P. Rouchon a ainsi prouvé dans [Rou02] (étendu ensuite avec M. Mirrahimi dans [MR04]), que toutes les approximations finies dimensionnelles de l'oscillateur harmonique quantique muni d'un contrôle de type dipôle sont contrôlables alors que le système de dimension infinie reste dans une variété de dimension finie. Cet exemple montre que le lien entre la contrôlabilité du système classique et de son équivalent quantique n'est systématique. Les relations entre ces deux notions qui existent via la théorie semi-classique restent encore assez inexplorées.

Une autre difficulté qui peut apparaître est reliée à la dégénérescence des valeurs propres qui est un phénomène profondément relié aux symétries. Le contrôle des rotations des molécules dans l'espace est ainsi un exemple où de nombreuses symétries discrètes et continues rendent l'étude du contrôle et de la contrôlabilité particulièrement délicate. Nous renvoyons aux travaux de E. Pozzoli et al. [BPS21 ; Poz22 ; Lei+22] pour l'étude théorique et expérimentale de ces systèmes. Notons également que des arguments permettant de passer de la contrôlabilité en dimension finie à la contrôlabilité approchée en dimension infinie avec des outils spectraux ont été développés par U. Boscain et al. dans [Bos+12a ; Bos+15].

Mentionnons aussi qu'il existe un résultat d'obstruction à la contrôlabilité exacte sur les systèmes bilinéaires énoncé par J. M. Ball, J. E. Marsden et M. Slemrod dans [BMS82], puis étendu notamment dans [ILT06 ; BCC19].

Cependant, dans la suite de ce manuscrit, nous nous restreindrons à des systèmes de dimension finie.

### I.5.2.3 Les qubits

Les systèmes quantiques à deux niveaux, c'est-à-dire lorsque  $\dim_{\mathbb{C}} \mathcal{H} = 2$ , sont appelés les *qubits*. Si certains systèmes comme le moment magnétique d'un fermion de spin 1/2 sont intrinsèquement des systèmes quantiques de dimension deux, on retrouve souvent les *qubits* comme une approximation d'un système quantique plus complexe : les *centre NV* qui sont des systèmes essentiels dans l'industrie des capteurs quantiques, les *Quantum dots* et les *jonctions Josephson* qui sont des candidats majeurs pour l'implémentation d'un ordinateur quantique, etc. Ces systèmes se comportent en effet dans certains régimes d'intérêts comme des systèmes à deux niveaux.

Commençons par classifier les opérateurs autoadjoints à trace nulle<sup>13</sup> agissant sur un espace de dimension deux. Ils se décomposent sur la base des matrices de Pauli

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \quad (\text{I.30})$$

Introduisons à présent  $w$  un contrôle à valeurs dans  $\mathbb{C}$  et considérons le système

$$i\partial_t|\psi\rangle = (E\sigma_z + \text{Re}(w(t))\sigma_x - \text{Im}(w(t))\sigma_y)|\psi\rangle \quad (\text{I.31})$$

où  $\text{Re}(w)$  (resp.  $\text{Im}(w)$ ) est la partie réelle (resp. imaginaire) de  $w$  et  $E > 0$  est l'énergie du système non contrôlé.

L'équation (I.31) décrit par exemple l'évolution du spin d'une particule dans le cadre de l'Imagerie par Résonance Magnétique (IRM) :  $E$  modélise un champ intense fixe selon l'axe  $Oz$ ,  $\text{Re}(w)$  (resp.  $\text{Im}(w)$ ) un champ contrôlé de plus faible amplitude selon l'axe  $Ox$  (resp.  $Oy$ ).

13. Les multiples de l'identité ne font qu'induire une dynamique relative à la phase globale de l'état, on peut donc les ignorer.

La contrôlabilité du système (I.31) avec  $w$  réel et borné peut être établie facilement. En effet, grâce aux techniques classiques de contrôle géométrique (voir par exemple [AS04]), on prouve que la dérive est récurrente et on applique le théorème de Rashevsky-Chow pour obtenir la contrôlabilité. D'autres questions de contrôle optimal peuvent être posées sur ce système. Par exemple la détermination du contrôle en temps minimal a été étudié par U. Boscain et P. Mason dans [BM06]. Nous verrons dans la Section I.6.2 un exemple de coût à minimiser amenant des problèmes dits de *chattering*.

### I.5.3 Contrôlabilité d'ensemble de qubits

Une question plus difficile que la contrôlabilité de l'équation (I.31) est celle de la contrôlabilité d'ensemble. Cela correspond à l'une des situations pratiques suivantes :

- On dispose de plusieurs qubits non identiques (l'énergie ou le couplage avec le contrôle sont légèrement différents), avec l'obligation d'utiliser un **unique contrôle** pour tous les qubits. Un exemple classique est l'IRM : il faut retourner un très grand nombre de spins avec un champ magnétique uniforme en espace et variable en temps : c'est donc le même contrôle pour toutes les particules.
- La connaissance du système est imparfaite, il est donc nécessaire de prendre un contrôle robuste face aux incertitudes.

Posons alors le problème suivant : on se donne des paramètres de dispersions  $(\alpha, \delta) \subset \mathbb{R} \times \mathbb{R}_+^*$  dans un compact  $D$  et on considère le système d'équations

$$\begin{cases} i\partial_t |\psi\rangle^{\alpha, \delta} &= \begin{pmatrix} E + \alpha & \delta w(t) \\ \delta w^*(t) & -E - \alpha \end{pmatrix} |\psi\rangle^{\alpha, \delta}, \\ |\psi(0)\rangle^{\alpha, \delta} &= |\psi_0\rangle^{\alpha, \delta}, \end{cases} \quad (\text{I.32})$$

où  $(\alpha, \delta) \mapsto |\psi_0\rangle^{\alpha, \delta}$  est une fonction (éventuellement constante) de  $D$  dans la sphère unité de  $\mathcal{H}$ , que l'on dénotera  $\mathcal{S}$ .

**Definition I.2** (Contrôlabilité d'ensemble). *On définit la contrôlabilité d'ensemble du système (I.32), avec des contrôles bornés par une constante  $K > 0$ , muni de la dispersion  $D$  par la propriété suivante :*

*Quelles que soient les distributions  $|\psi_0\rangle^{\alpha, \delta}, |\psi_f\rangle^{\alpha, \delta} \in \mathcal{C}(D, \mathcal{S})$ , et une précision  $\varepsilon > 0$ , il existe un temps  $T > 0$  et un contrôle  $w \in L^\infty([0, T])$  tels que  $|w| \leq K$  et que la solution du système (I.32) satisfasse*

$$\forall (\alpha, \delta) \in D, \exists \theta \in [0, 2\pi), \text{ tel que } \| |\psi(T)\rangle^{\alpha, \delta} - e^{i\theta} |\psi_f\rangle^{\alpha, \delta} \| \leq \varepsilon.$$

Le cas où  $D$  est fini se traite de manière assez similaire au cas de la contrôlabilité du système (I.31) : il suffit de considérer un nouvel état  $|\psi\rangle^{\text{tot}} = \prod_{(\alpha, \delta) \in D} |\psi\rangle^{\alpha, \delta} \in \mathcal{H}^{\#D}$ , où  $\#D$  est le cardinal de  $D$ . On obtient alors un système de dimension finie et les méthodes habituelles de contrôle géométrique peuvent être employées. Le cas où  $D$  est un ensemble continu est sensiblement différent, et nous nous focaliserons désormais sur ce cas.

Avant de faire une brève revue des travaux sur ce sujet récent, introduisons deux nouvelles notions de contrôlabilité d'ensemble. Une version plus faible est la contrôlabilité d'ensemble entre états propres. On se restreint alors à ne considérer comme distributions initiale et finale dans la définition I.2 que des états propres de la dérive  $|i\rangle^{\alpha, \delta}$  et  $|f\rangle^{\alpha, \delta}$ . Au contraire, une notion plus forte est la contrôlabilité d'ensemble sur le groupe (aussi appelé *propagateur*) associé à l'équation de Schrödinger. Introduisons pour ce faire le système d'équations différentielles sur le groupe de

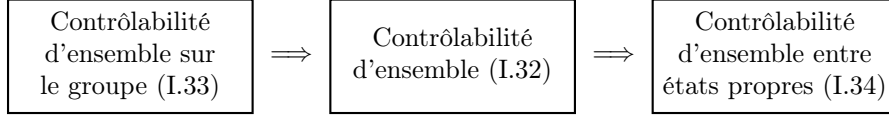


FIGURE I.8 – Implications entre les différents types de contrôlabilité d'ensemble.

Lie  $SU_2$  suivant :

$$\begin{cases} i \frac{d}{dt} M(\alpha, \delta, t) &= \begin{pmatrix} E + \alpha & \delta w(t) \\ \delta w^*(t) & -E - \alpha \end{pmatrix} M(\alpha, \delta, t), \\ M(\alpha, \delta, 0) &= \text{Id}. \end{cases} \quad (\text{I.33})$$

**Définition I.3** (Contrôlabilité d'ensemble sur le groupe). *On définit la contrôlabilité d'ensemble sur le groupe avec des contrôles bornés par la constante  $K > 0$  muni de la dispersion  $D$  par la propriété suivante :*

*Quelles que soient la distribution cible  $M_F \in \mathcal{C}(D, SU_2)$  et la précision  $\varepsilon > 0$ , il existe un temps  $T > 0$  et un contrôle  $w \in L^\infty([0, T])$  tels que  $|w| \leq K$  et que la solution du système (I.33) satisfasse  $\|M(\cdot, \cdot, T) - M_F(\cdot, \cdot)\|_{L^\infty(D, SU_2)} < \varepsilon$ .*

La Figure I.8 résume les implications entre ces trois notions.

Les travaux fondateurs de J.-S. Li et N. Khaneja [LK09; LK06] sont dans une large mesure à l'initiative du domaine. En utilisant précautionneusement les crochets de Lie de la dynamique, ils ont notamment prouvé le théorème suivant.

**Théorème I.4** (Li–Khaneja, 2009).<sup>14</sup> *Soit un espace de dispersion  $D \subset \mathbb{R} \times \mathbb{R}_+^*$  et une constante  $K > 0$  quelconque, alors (I.33) satisfait la propriété de contrôlabilité d'ensemble sur le groupe.*

La preuve est fondée sur des techniques de crochets de Lie en dimension infinie. Cependant, une des difficultés est de pouvoir effectuer des crochets de Lie avec la dérive. Nous y reviendrons dans la Section I.6.1.3. Notons que l'approche utilisée a été étendue dans le cadre sans dérive par A. Agrachev, Y. Baryshnikov et A. Sarychev à des systèmes plus généraux que ceux provenant de la mécanique quantique [ABS16].

K. Beauchard, J.-M. Coron et P. Rouchon dans [BCR10a] ont étudié la contrôlabilité d'ensemble sur les *qubits* à l'aide de l'analyse fonctionnelle. Le problème est équivalent à l'étude d'une EDP bilinéaire à spectre continu. La non-contrôlabilité exacte en temps fini pour des contrôles bornés  $L^2$  ainsi que la contrôlabilité approchée  $L^\infty$  en temps fini avec des contrôles non bornés y sont prouvées.

### I.5.3.1 Inversion adiabatique de populations

Un outil extrêmement puissant mais spécifique au contrôle quantique est le *contrôle adiabatique*. Il permet notamment de réaliser des contrôles robustes entre états propres à l'aide du théorème adiabatique. En effet, une propriété remarquable de l'équation de Schrödinger est la stabilité des états propres sous l'effet des variations lentes de l'Hamiltonien. Énonçons brièvement l'idée principale de ce théorème : considérons une fonction continue  $H(t)$  de  $[0, 1]$  à valeurs dans les opérateurs autoadjoints et satisfaisant des hypothèses de gaps spectraux. Considérons également une valeur propre  $\lambda(0)$  et un état propre  $|\psi_{\lambda(0)}\rangle$  de  $H(0)$ . On peut définir  $\lambda(t)$  et

14. Ce théorème est reproduit dans le Chapitre V sous le nom de Théorème V.7.

$|\psi_{\lambda(t)}\rangle$  les valeurs propres et états propres instantanés de  $H(t)$ . Alors la solution de l'équation

$$\begin{cases} i\varepsilon \frac{d}{dt} |\psi(t)\rangle = H(\varepsilon t) |\psi(t)\rangle & \text{pour } t \in [0, \frac{1}{\varepsilon}], \\ |\psi(0)\rangle = |\psi_{\lambda(0)}\rangle, \end{cases}$$

où  $\varepsilon \ll 1$  encode la lenteur du parcours, satisfait

$$1 - |\langle \psi_{\lambda(1)} | \psi(\frac{1}{\varepsilon}) \rangle| = O(\varepsilon).$$

Nous renvoyons à [Teu03] pour une introduction rigoureuse aux différentes versions du théorème adiabatique.

Pour le contrôle d'ensemble entre états propres, l'utilisation de la théorie adiabatique nécessite d'analyser la manière dont les valeurs propres s'intersectent afin de passer d'une valeur propre à l'autre. Nous renvoyons à [LSR11] ainsi qu'aux nombreux travaux de U. Boscain et al. sur le sujet, pour ne citer que [AB05; Bos+12b; Bos+15; ABS18; ABS20].

Dans le cas des qubits avec une dérive selon  $\sigma_z$ , les états propres sont  $(1, 0)$  et  $(0, 1)$ . On peut donc poser le problème dit d'inversion robuste de populations (ce qui revient à la contrôlabilité d'ensemble entre états propres) qui se formule comme suit : pour  $\varepsilon > 0$ , peut-on trouver  $T > 0$  et un contrôle borné  $w$  tel que la solution de

$$\begin{cases} i\partial_t |\psi\rangle^{\alpha, \delta} = \begin{pmatrix} E + \alpha & \delta w(t) \\ \delta w^*(t) & -E - \alpha \end{pmatrix} |\psi\rangle^{\alpha, \delta}, \\ |\psi(0)\rangle^{\alpha, \delta} = (0, 1), \end{cases} \quad (\text{I.34})$$

satisfasse

$$\forall (\alpha, \delta) \in D, \exists \theta \in [0, 2\pi), \text{ tel que } \| |\psi(T)\rangle^{\alpha, \delta} - (e^{i\theta}, 0) \| \leq \varepsilon? \quad (\text{I.35})$$

La réponse à cette question est positive si  $w$  est un contrôle à valeurs complexes. Pour cela, il est utile dans un premier temps d'effectuer le changement de variables suivant :

$$|\tilde{\psi}(t)\rangle^{\alpha, \delta} = e^{-i(Et + \frac{\Delta(t)}{2})\sigma_z} |\psi(t)\rangle^{\alpha, \delta}, \quad (\text{I.36})$$

où  $\Delta(t)$  est une fonction dérivable arbitraire pour le moment. L'application  $t \mapsto e^{-i(Et + \frac{\Delta(t)}{2})\sigma_z}$  est à valeurs dans les opérateurs unitaires. De plus  $|\tilde{\psi}(t)\rangle^{\alpha, \delta}$  est solution d'une nouvelle équation de Schrödinger :

$$\begin{aligned} i\partial_t |\tilde{\psi}\rangle^{\alpha, \delta} &= \left( \alpha - \frac{\Delta'(t)}{2} \right) \sigma_z |\tilde{\psi}\rangle^{\alpha, \delta} + \delta e^{-i(Et + \frac{\Delta(t)}{2})\sigma_z} (\text{Re}(w(t))\sigma_x - \text{Im}(w(t))\sigma_y) e^{i(Et + \frac{\Delta(t)}{2})\sigma_z} |\tilde{\psi}\rangle^{\alpha, \delta} \\ &= \begin{pmatrix} \alpha - \frac{\Delta'(t)}{2} & \delta w(t) e^{-i(2Et + \Delta(t))} \\ \delta w^*(t) e^{i(2Et + \Delta(t))} & -\alpha + \frac{\Delta'(t)}{2} \end{pmatrix} |\tilde{\psi}\rangle^{\alpha, \delta}. \end{aligned}$$

En exprimant le contrôle  $w(t)$  à l'aide de son amplitude et sa phase relative au repère d'interaction, on introduit  $u$  et on fixe  $\Delta$  par :

$$w(t) = u(t) e^{i(2Et + \Delta(t))}. \quad (\text{I.37})$$

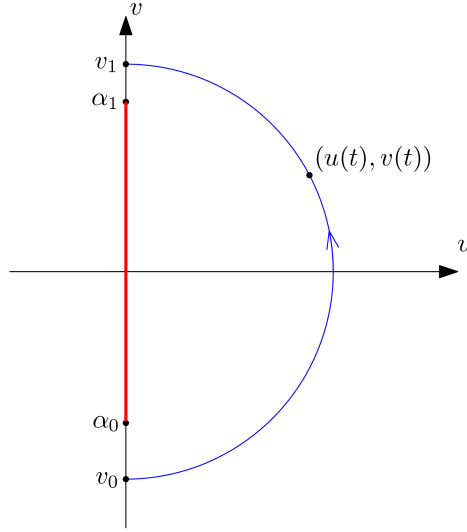


FIGURE I.9 – Reproduction de la Figure V.1. Un exemple de chemin adiabatique (en bleu) permettant l'inversion robuste de populations pour  $\alpha \in [\alpha_0, \alpha_1]$

En posant  $v(t) = \frac{\Delta'(t)}{2}$ , la dynamique de  $|\tilde{\psi}\rangle^{\alpha, \delta}$  s'exprime sous la forme suivante :

$$i\partial_t |\tilde{\psi}\rangle^{\alpha, \delta} = ((\alpha - v(t))\sigma_z + \delta u(t)\sigma_x) |\tilde{\psi}\rangle^{\alpha, \delta}. \quad (\text{I.38})$$

Remarquons dans un premier temps que, puisque  $|\psi\rangle$  et  $|\tilde{\psi}\rangle$  ne diffèrent que d'une phase relative, résoudre le problème (I.35) pour  $|\tilde{\psi}\rangle$  est équivalent à le résoudre pour  $|\psi\rangle$ . Nous allons ainsi utiliser un contrôle adiabatique pour  $|\tilde{\psi}\rangle$ .

Pour cela, nous utilisons un chemin adiabatique ayant une forme similaire à celui donné dans la Figure I.9. Notons  $\tilde{H}^{\alpha, \delta}(u, v) = (\alpha - v)\sigma_z + \delta u\sigma_x$  l'Hamiltonien du membre de droite de l'équation (I.38). Les valeurs propres de  $\tilde{H}^{\alpha, \delta}(u, v)$  sont  $\pm\sqrt{(\alpha - v)^2 + (\delta u)^2}$ . L'intersection des valeurs propres est ainsi conique et située en  $(0, \alpha)$ . Le chemin adiabatique proposé évite ainsi ces intersections. Notons également que les Hamiltoniens aux temps initial et final sont

$$\tilde{H}^{\alpha, \delta}(0, v_0) = \begin{pmatrix} \alpha - v_0 & 0 \\ 0 & -\alpha + v_0 \end{pmatrix}, \quad \tilde{H}^{\alpha, \delta}(0, v_1) = \begin{pmatrix} \alpha - v_1 & 0 \\ 0 & -\alpha + v_1 \end{pmatrix}. \quad (\text{I.39})$$

Ainsi, en supposons que  $|\tilde{\psi}\rangle$  est initialement dans l'état  $(0, 1)$ , qui est l'état propre associé à la valeur propre négative de  $\tilde{H}^{\alpha, \delta}$  au temps initial, alors  $|\tilde{\psi}\rangle$  restera proche de l'état propre instantané de  $\tilde{H}$  associé à la valeur propre négative dans la limite adiabatique. Or au temps final, l'état propre associé à la valeur propre négative est devenu  $(1, 0)$ .

Cette stratégie permet ainsi d'effectuer des transferts robustes de population, à condition de prendre un chemin adiabatique qui n'intersecte pas les intersections coniques et de le parcourir suffisamment lentement.

En revenant au système  $|\psi\rangle$ , on doit donc appliquer un contrôle complexe de la forme

$$w_\varepsilon(t) = u(\varepsilon t)e^{i(Et + \frac{\Delta(\varepsilon t)}{\varepsilon})}, \quad (\text{I.40})$$

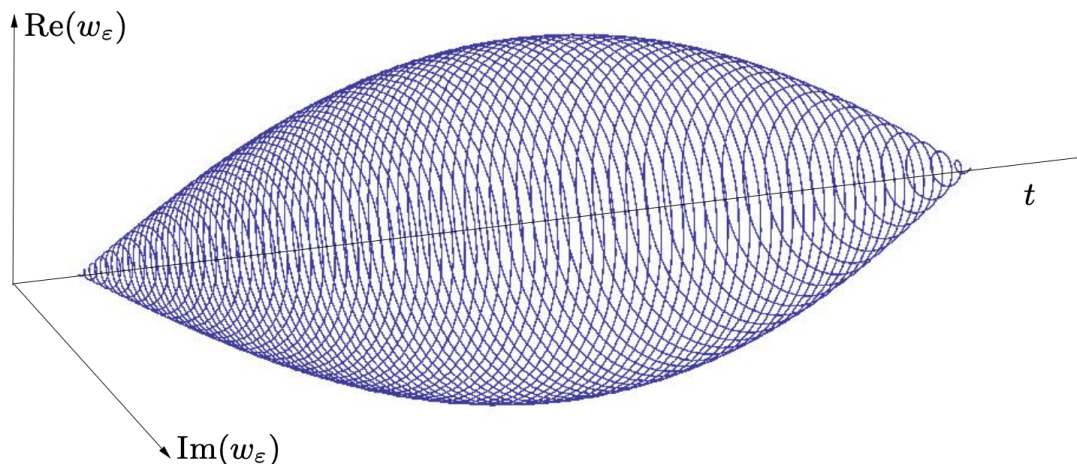


FIGURE I.10 – Un exemple de contrôle complexe de type *chirp*. L'amplitude augmente puis diminue alors que la fréquence baisse au cours du temps.

où  $u$  et  $\Delta(t) = 2 \int_0^t v(t') dt'$  sont des fonctions de  $[0, 1]$  vers  $\mathbb{R}$  définies à partir du chemin adiabatique  $(u, v)$  choisi. Ce type de contrôle est appelé *chirp* ou *adiabatic chirp*. Notons que  $u$  encode l'amplitude du contrôle, et  $v$  (qui détermine  $\Delta$ ) les fréquences balayées par le contrôle. La Figure I.10 représente un *chirp*.

Ainsi, ces contrôles assurent une inversion robuste de populations pour n'importe quel compact  $D \subset [\alpha_0, \alpha_1] \times \mathbb{R}_+^*$  avec une précision de l'ordre de  $\frac{1}{T}$ . Nous renvoyons à la Section V.1.2 pour plus de détails.

## I.6 Contributions au contrôle de systèmes quantiques

Dans la pratique, il est courant de ne pas pouvoir agir sur  $\sigma_x$  et  $\sigma_y$  mais uniquement sur un seul d'entre eux. Cela revient, si l'on choisit de mettre le contrôle uniquement devant  $\sigma_x$ , à imposer que le contrôle  $w$  soit à valeurs réelles. Pour des contrôles *résonants*, c'est-à-dire ayant une fréquence proche de celle du système, l'approximation de l'onde tournante permet de simuler un contrôle complexe résonant avec le système à l'aide d'un contrôle réel. Cependant, en regardant plus précisément les hypothèses, il n'est pas évident qu'il soit possible d'utiliser simultanément les approximations adiabatique et de l'onde tournante. Ceci est l'objet d'une des contributions de cette thèse.

### I.6.1 Compatibilité entre l'approximation de l'onde tournante et l'approximation adiabatique

Ce travail a été effectué en collaboration avec Nicolas Augier<sup>15</sup>, Ugo Boscain<sup>16</sup> et Mario Sigalotti<sup>6</sup>. Il est présenté en détails dans le Chapitre V.

15. Laboratoire d'analyse et d'architecture des systèmes, Toulouse

16. Laboratoire Jacques-Louis Lions, CNRS



### I.6.1.1 Approximation de l'onde tournante et différences avec l'adiabatique

L'approximation de l'onde tournante est un cas particulier de moyennisation temporelle, cette notion appartient à la théorie de l'*averaging* [SVM07]. Dans le cas de l'inversion de populations pour les qubits, elle revient à énoncer que les solutions de (I.34) avec les contrôles

$$w_\varepsilon(t) = 2\varepsilon u(\varepsilon t) \cos(2Et + \Delta(\varepsilon t)), \quad (\text{I.41})$$

$$w_\varepsilon^{\text{R}}(t) = \varepsilon u(\varepsilon t) e^{-i(2Et + \Delta(\varepsilon t))}, \quad (\text{I.42})$$

induisent des trajectoires dont la différence reste de l'ordre de  $\varepsilon$  sur une durée  $\frac{1}{\varepsilon}$ . La preuve repose sur le fait qu'en décomposant le cosinus en exponentielles complexes, on obtient

$$w_\varepsilon(t) = w_\varepsilon^{\text{R}}(t) + \varepsilon u(\varepsilon t) e^{i(2Et + \Delta(\varepsilon t))}. \quad (\text{I.43})$$

Il est alors possible de montrer que l'anti-résonance dans le référentiel d'interaction du second terme du membre de droite de l'équation (I.43) a peu d'effet sur la dynamique (voir Section V.1.1).

Notons dès à présent que si les contrôles (I.40) et (I.42) sont bien paramétrés par  $\varepsilon$  qui définit leur précision, et qu'ils doivent être appliqués sur un temps  $\frac{1}{\varepsilon}$ , des différences importantes entre ces deux contrôles existent :

- Pour l'approximation adiabatique, la bande de fréquences balayées par le *pulse* (I.40), c'est-à-dire la dérivée de l'exponentielle complexe, est indépendante du paramètre  $\varepsilon$  (grâce à l'utilisation de  $\frac{\Delta(\varepsilon t)}{\varepsilon}$  et non  $\Delta(\varepsilon t)$ ). Ce n'est pas le cas pour l'approximation de l'onde tournante qui réduit les fréquences balayées quand  $\varepsilon$  devient petit. En abandonnant la dispersion sur l'énergie (i.e., en fixant  $\alpha = 0$ ), N. Augier et al. ont prouvé qu'on pouvait faire fonctionner les deux approximations ensembles [ABS19].
- Pour l'approximation de l'onde tournante, il est nécessaire d'utiliser un contrôle petit<sup>17</sup>. Or, utiliser un contrôle de faible amplitude est doublement néfaste pour la théorie adiabatique : cela réduit le gap spectral et augmente la norme de la dérivée des projecteurs spectraux.

Il existe ainsi un antagonisme certain entre ces deux approximations.

### I.6.1.2 Compatibilité et idées de preuve

Afin d'utiliser au mieux chacune des approximations, nous utilisons deux paramètres positifs  $\varepsilon_1$  et  $\varepsilon_2$  dans le contrôle suivant :

$$w_{\varepsilon_1, \varepsilon_2}(t) = 2\varepsilon_1 u(\varepsilon_1 \varepsilon_2 t) \cos\left(2Et + \frac{\Delta(\varepsilon_1 \varepsilon_2 t)}{\varepsilon_1 \varepsilon_2}\right), \quad (\text{I.44})$$

sur un temps  $\frac{1}{\varepsilon_1 \varepsilon_2}$ . Nous avons alors prouvé qu'il était possible d'effectuer une inversion robuste de populations pour  $D = [\alpha_0, \alpha_1] \times [\delta_0, \delta_1]$ ,  $\delta_0 > 0$  sous l'hypothèse

$$3(E + \alpha_0) > (E + \alpha_1). \quad (\text{H})$$

---

<sup>17</sup>. Il y a  $\varepsilon$  en facteur du contrôle dans l'équation (I.42)

Le Théorème V.3 énonce alors que sous l'hypothèse (H), quel que soit  $N_0$ , il existe une constante  $C_{N_0}$  telle que l'erreur de l'inversion de populations soit contrôlée par

$$\left| \left| \psi_{\varepsilon_1, \varepsilon_2}^{\alpha, \delta} \left( \frac{1}{\varepsilon_1 \varepsilon_2} \right) \right\rangle - (e^{i\theta}, 0) \right| < C_{N_0} \max \left( \frac{\varepsilon_2}{\varepsilon_1}, \frac{\varepsilon_1^{N_0-1}}{\varepsilon_2} \right), \quad (\text{I.45})$$

pour un certain  $\theta \in \mathbb{R}$ . Le ratio  $\frac{\varepsilon_2}{\varepsilon_1}$  est informellement l'erreur due à l'approximation adiabatique, alors que  $\frac{\varepsilon_1^{N_0-1}}{\varepsilon_2}$  est assimilé à celle de l'onde tournante.

La preuve commence par l'utilisation de changements de variables successifs inspirés de la théorie classique de l'*averaging* afin de gérer l'approximation de l'onde tournante. Nous introduisons les outils d'algèbre nécessaires et formalisons un algorithme afin de comprendre les termes apparaissant. Afin de prouver l'inversion robuste de population, il est nécessaire d'aller au minimum jusqu'au troisième ordre de l'approximation de l'onde tournante.

L'utilisation de ces changements de variables nécessite l'hypothèse (H). On notera également que des simulations numériques (voir Figure V.2) suggèrent que cette hypothèse est nécessaire. Par ailleurs, ces changements de variables ont la propriété extrêmement intéressante de donner des systèmes restant très proche du système initial au temps initial et au temps final, mais pas au cours de la trajectoire. Ainsi, comme cela est illustré par les Figures V.9 et V.10, malgré une déviation non négligeable le long de la trajectoire entre l'évolution adiabatique –utilisant un contrôle complexe– et notre évolution –utilisant un contrôle réel–, l'écart entre les deux trajectoires redevient très faible au temps final.

La deuxième partie de la preuve utilise des propriétés fines de la théorie adiabatique pour estimer l'évolution de l'Hamiltonian réduit obtenu après la première étape. La difficulté principale est que le chemin adiabatique dépend de  $\varepsilon_1$ ; or plus  $\varepsilon_1$  est petit, plus le gap spectral le long du chemin est faible. Notons également que les simulations numériques, notamment la Figure V.7, semblent confirmer que le paramètre contrôlant l'erreur adiabatique obtenue par notre théorème,  $\frac{\varepsilon_2}{\varepsilon_1}$ , est optimal. Cependant, nous observons numériquement une convergence exponentielle  $e^{-\lambda \frac{\varepsilon_2}{\varepsilon_1}}$ , alors que notre borne est seulement linéaire. Cette amélioration de l'estimation reste ainsi ouverte.

### I.6.1.3 Spin-echos et application à la contrôlabilité sur le groupe

Nous avons mentionné que la preuve des résultats de J.-S. Li et N. Khaneja utilisait de manière cruciale la possibilité d'effectuer des crochets de Lie avec la dérive. Pour cela il faut pouvoir générer l'évolution en temps négatif de la dérive. En dimension finie, il suffit que la dérive soit récurrente, mais cela est rarement le cas en dimension infinie. Les *pulses* adiabatiques sont justement un outil idéal pour cela. En effet, sur la sphère de Bloch, un *pulse* adiabatique est, approximativement, une rotation d'angle  $\pi$  avec un axe de rotation quelque part sur l'équateur  $z = 0$ ; mais l'axe dépend a priori de la dispersion  $(\alpha, \delta)$ . Cependant en effectuant une première fois ce *pulse*, puis en attendant un temps  $t_0$  et en réappliquant ce même *pulse*, on obtient l'évolution  $e^{it_0(E+\alpha)\sigma_z}$ , c'est-à-dire l'évolution en temps négatif de la dérive pendant  $t_0$ . Cette technique permet par exemple de réaliser des expériences dites de *spin-echos* [Hah50].

Le Théorème V.3 nous permet ainsi de générer des *pulse* d'angle  $\pi$  à l'aide d'un seul contrôle. Cela nous a permis de généraliser le théorème de contrôlabilité d'ensemble sur le groupe I.4 en utilisant uniquement un contrôle réel. Notre résultat, le Théorème V.10, nécessite cependant l'hypothèse (H) sur la dispersion fréquentielle.

## I.6.2 Un exemple de *chattering* en contrôle quantique

Ce travail a été effectué en collaboration avec Ugo Boscain<sup>16</sup>, Mario Sigalotti<sup>6</sup> et Dominique Sugny<sup>18</sup>. Il est présenté en détails dans le Chapitre VI et a été soumis dans un journal de physique.

### I.6.2.1 Modélisation physique

La seconde contribution que nous présentons est reliée à un problème de contrôle optimal. Considérons un système à trois niveaux, notés  $|1\rangle$ ,  $|2\rangle$ ,  $|3\rangle$  avec un contrôle qui couple les états  $|1\rangle$  et  $|2\rangle$  ainsi qu'un couplage non contrôlable dit de pompage entre les états  $|2\rangle$  et  $|3\rangle$ . La dynamique peut se réduire (voir Section VI.2.1) à la dynamique suivante sur la demi sphère  $x_3 \geq 0$  de  $\mathbb{R}^3$  :

$$\dot{\mathbf{X}} = (\Delta\Omega_3 + u(t)\Omega_1)\mathbf{X}, \quad (\text{I.46})$$

où  $\mathbf{X}$  est un vecteur à coordonnées réelles  $(x_1, x_2, x_3)$  avec la condition d'unitarité  $x_1^2 + x_2^2 + x_3^2 = 1$ .

Les matrices anti-symétriques  $\Omega_1$  et  $\Omega_3$  sont les équivalents de  $\sigma_z$  et  $\sigma_x$  agissant sur la sphère de Bloch

$$\Omega_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \Omega_3 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

On souhaite à présent, à partir de n'importe quelle condition initiale sur la demi sphère  $x_3 \geq 0$ , atteindre l'état  $|3\rangle$  en minimisant la présence dans l'état  $|1\rangle$ . Cela peut être motivé par exemple dans le cas où l'état  $|1\rangle$  est soumis à un processus de relaxation non désiré ; il est alors naturel de minimiser la présence dans cet état, c'est-à-dire

$$\int_0^T x_1^2(s) ds. \quad (\text{I.47})$$

Nous cherchons ainsi la trajectoire optimale pour le coût (I.47) atteignant  $(0, 0, 1)$  avec le temps final  $T$  libre et en imposant que  $u(t) \in [-1, 1]$ .

### I.6.2.2 La théorie du *chattering*

Un outil très puissant pour résoudre les problèmes de contrôle optimaux est le Principe du Maximum de Pontryagin (PMP), étudié par le célèbre mathématicien du même nom [Pon+74]. Cependant, de nombreuses subtilités et difficultés existent dans l'étude du contrôle optimal, y compris dans le cadre de la dimension finie et du contrôle linéaire. Un des problèmes relativement incompris est le *chattering*. Ce phénomène a été découvert par A. T. Fuller [Ful60]. L'exemple historique est le suivant : on fixe une condition initiale  $(x_0, y_0) \neq (0, 0)$  et on souhaite minimiser le coût intégral  $\int_0^T x^2(s) ds$  avec le temps final  $T$  libre et  $u \in L^\infty([0, T], [-1, 1])$  avec la dynamique

$$\dot{x} = y, \quad \dot{y} = u. \quad (\text{I.48})$$

La trajectoire optimale atteint  $(0, 0)$  en temps fini, le contrôle optimal est *bang-bang*, c'est-à-dire que  $|u| = 1$  presque partout, mais une accumulation d'un nombre infini de *switch* se produit avant d'atteindre  $(0, 0)$ . C'est ainsi un cas d'*sur-regulation* puisque la trajectoire optimale coupe l'axe  $y = 0$  en suivant une progression géométrique vers  $(0, 0)$ .

---

18. Laboratoire Interdisciplinaire Carnot de Bourgogne, Université de Bourgogne-Franche-Comté

Ce phénomène peut apparaître dans de nombreux domaines d'applications du contrôle optimal, H. Schättler et U. Ledzewicz l'ont par exemple identifié en médecine [SL09], M. I. Zelikin et V. F. Borisov dans de nombreux systèmes mécaniques [ZB94], J. Zhu, E. Trélat et M. Cerf dans des systèmes de contrôle spatial [ZTC16] . . . I. Kupka a par ailleurs prouvé des conditions suffisantes pour que les extrémales du PMP présentent du *chattering*. Il en a déduit qu'en dimension élevée, ce phénomène est générique [Kup90].

Ainsi, le *chattering* n'est pas uniquement une curiosité mathématique, c'est un problème concret qui se pose dans de nombreuses applications. Ces accumulations de *switch* sont extrêmement néfastes pour les résolutions numériques à double titre ; non seulement il n'est pas possible d'utiliser des méthodes de tir rétropropagé car le covecteur est nul au temps final<sup>19</sup>, mais de surcroît les méthodes directes souffrent également d'une grande instabilité au voisinage du point où le *chattering* se produit. Pour résoudre ce problème, M. Caponigro et al. ont étudié une méthode de régularisation en pénalisant la variation totale du contrôle et empêchant de ce fait le *chattering* de se produire [Cap+18].

À notre connaissance, ce phénomène commun en mécanique classique n'avait jamais été observé dans le cadre de la mécanique quantique [Gla+15 ; BCR10b ; Koc+22].

### I.6.2.3 Identification et résolution numérique

Nous avons mis en lumière la présence de *chattering* pour le problème de contrôle optimal de la dynamique (I.46) munie du coût (I.47) au voisinage de  $(0, 0, 1)$ . Rappelons que l'étude des extrémales en contrôle optimale nécessite de considérer des trajectoires sur le fibré cotangent de l'espace des états. Dans notre cas, l'espace d'intérêt est ainsi le fibré cotangent de la sphère. Les trajectoires optimales sont alors des projections des solutions d'une dynamique hamiltonienne sur ce fibré.

Afin de prouver que le système que nous considérons présente du *chattering*, nous avons montré la présence d'une symétrie anisotrope spécifique dans l'espace cotangent. En utilisant des résultats de M. I. Zelikin et V. F. Borisov [ZB94, Chapitre 3], nous avons alors prouvé que les trajectoires optimales atteignent l'état  $(0, 0, 1)$  en temps fini, et obtenu l'asymptotique de la courbe de *switch* donnant la synthèse optimale. Nous insistons sur le fait que notre preuve concerne les trajectoires optimales (et pas uniquement extrémales).

En particulier, notre exemple prouve que la structure des équations de la mécanique quantique n'empêche pas le *chattering* de se produire en contrôle optimal de système quantique.

Nous avons ensuite caractérisé partiellement la courbe de *switch* et l'avons approché numériquement. La stratégie est alors la suivante : En utilisant l'asymptotique de la courbe de *switch* au voisinage de  $(0, 0, 1)$  obtenue grâce aux résultats de M. I. Zelikin et V. F. Borisov, nous avons reconstruit numériquement la synthèse optimale en intégrant en temps inversé les équations hamiltonienne sur le fibré cotangent.

Cette approche est justifiée théoriquement (au moins localement) par l'argument suivant. Notons  $S_0$  l'ensemble des points du fibré cotangent satisfaisant  $H = 0$  et  $\varphi = 0$  où  $H$  est l'Hamiltonien de Pontryagin et  $\varphi$  la fonction de *switch*. La dynamique de notre système restant sur une variété de dimension 2, le fibré cotangent est de dimension 4 et  $S_0$  est une sous-variété de dimension 2. Définissons à présent sur  $S_0$  le système dynamique discret  $\Phi$  consistant à suivre la dynamique hamiltonienne<sup>20</sup>. Alors,  $x = (0, 0, 1), p = 0$  est un point fixe de l'application  $\Phi^2$ . Cette dernière est hyperbolique (dégénérée) au voisinage de ce point et admet une unique variété stable qui correspond à la courbe de *switch* qui nous intéresse. Nous renvoyons à [ZB94, Chapitre 3] pour les preuves de ces propriétés et au Chapitre VI pour la preuve que le système (I.46) satisfait

19. on ne sait donc pas comment initialiser la dynamique *backward*.

20. Voir [ZB94, Remarque 3.1] pour la définition rigoureuse de ce système dynamique non lisse

les hypothèses nécessaires. Ainsi, en simulant la dynamique *backward*, nous utilisons la propriété de stabilité de la courbe de *switch*, car elle correspond à la variété stable de la dynamique de  $\Phi^2$ , pour assurer sa reconstruction fidèle.

### I.6.3 Quelques problèmes ouverts et perspectives

Les deux contributions décrites précédemment ouvrent de nombreuses questions qui nous semblent pertinentes et intéressantes à étudier.

**Amélioration de la borne liée à l'onde tournante dans (I.45).** L'erreur sur le *pulse* adiabatique réel se borne partiellement par  $C_{N_0} \frac{\varepsilon_1^{N_0-1}}{\varepsilon_2}$  correspondant à l'utilisation de l'approximation de l'onde tournante jusqu'à l'ordre  $N_0$ . Nous n'avons réussi ni à borner uniformément la constante  $C_{N_0}$ , ni à prouver le contraire. Nous posons alors deux questions : y a-t-il convergence exponentielle en  $\varepsilon_1$  de l'erreur présente dans le terme de gauche de l'inégalité (I.35) ? A-t-on de plus le résultat plus fort suivant : dans l'hypothèse où  $\varepsilon_1$  est pris assez petit mais fixé et sous les hypothèses du Théorème V.3, peut-on espérer une convergence de cette même erreur avec  $\varepsilon_2$  allant vers zéro ?

**Extension à des dimensions plus élevées.** Dans [ABS19 ; ABS22], N. Augier, U. Boscain et M. Sigalotti prouvent des résultats en dimensions plus élevées sur la compatibilité entre l'approximation de l'onde tournante et l'approximation adiabatique. Cependant, ces approches ne sont pas robustes aux incertitudes sur les énergies du système et n'utilisent pas les deux échelles  $(\varepsilon_1, \varepsilon_2)$  que nous avons introduites. Il nous semble raisonnable de penser que cette robustesse pourrait être récupérée en appliquant les techniques développées au Chapitre V au prix de conditions plus complexes sur les plages d'incertitudes autorisées.

**Extension à des opérateurs non auto-adjoints.** L'étude du transfert de populations à l'aide de l'approximation adiabatique et de l'approximation de l'onde tournante dans le cas d'un système ouvert serait également une extension intéressante.

**Contrôle optimal d'ensemble.** Nous avons prouvé un résultat de contrôlabilité d'ensemble mais la question du temps nécessaire pour obtenir une précision donnée est évidemment d'une grande importance. À la connaissance de l'auteur, ce domaine est encore largement ouvert dans le cas d'une dispersion continue.

**Étude du *chattering* en contrôle quantique.** Ce domaine est encore à un stade préliminaire. Par exemple, nous ne savons pas si l'ubiquité du phénomène qui est connue pour des systèmes classiques reste valide pour des dynamiques contraintes de respecter la structure de l'équation de Schrödinger. De même, peut-on exhiber des conditions vérifiables assurant l'optimalité des extrémales présentant un phénomène de *chattering* dans le cas quantique ? Les réponses à ces questions permettraient de mieux comprendre ce phénomène et pouvoir l'identifier plus facilement.

## I.7 Contrôlabilité globale à zéro en temps petit des équations de Burgers généralisées

La dernière partie de cette thèse se concentre sur un problème de contrôlabilité d'équations aux dérivées partielles non linéaires en dimension un d'espace. Ce travail a été effectué sous les conseils de Jean-Michel Coron<sup>21</sup> et est présenté en détails dans le Chapitre VII.

### I.7.1 Description du système étudié

Pour  $T > 0$  fixé et  $\gamma \geq 1$  un réel, nous nous intéressons à l'équation de Burgers généralisée sur le segment  $[0, 1]$  :

$$\begin{cases} y_t + (|y|^\gamma)_x - y_{xx} = u(t) & \text{on } (0, T) \times (0, 1), \\ y(t, 0) = v(t) & \text{on } (0, T), \\ y(t, 1) = 0 & \text{on } (0, T), \\ y(0, x) = y_0(x) & \text{on } (0, 1), \end{cases} \quad (E_\gamma)$$

où  $u(t)$  est un contrôle interne uniforme en espace, et  $v(t)$  un contrôle au bord. La question qui nous intéresse est celle de la contrôlabilité globale à zéro en temps court, c'est-à-dire de savoir si de n'importe quel état initial  $y_0$  (éventuellement très grand) et pour n'importe quel temps final  $T$  (éventuellement très court), on peut trouver des contrôles  $u$  et  $v$  telle que la solution de l'équation  $(E_\gamma)$  soit amenée à zéro au temps  $T$ .

### I.7.2 Motivations et résultats existants

Notre motivation pour étudier le système  $(E_\gamma)$  provient de questions relatives à la contrôlabilité d'équations de la mécanique des fluides en dimensions deux et trois. Rappelons brièvement quelques résultats majeurs obtenus au cours des trente dernières années. J.-M. Coron et O. Glass ont respectivement établi la contrôlabilité en dimension deux d'espace [Cor93] et en dimension trois [Gla97] des équations d'Euler. Quant à l'équation de Navier–Stokes, le cas bidimensionnel dans une variété sans frontière a été résolu par J.-M. Coron et A. V. Fursikov dans [CF96]. Suite à ce dernier résultat, A. V. Fursikov et O. Imanuvilov ont prouvé un résultat de contrôlabilité exacte globale en dimension trois dans [FI99] avec un contrôle agissant sur toute la frontière.

Plus récemment, J.-M. Coron, F. Marbach et F. Sueur ont prouvé dans [CMS20] la contrôlabilité globale à zéro en temps court en s'appuyant sur un contrôle agissant sur une partie de la frontière et une condition de Navier de glissement avec frottement ailleurs. Le même problème avec une condition limite de Dirichlet, énoncé par Lions dans [Lio91], reste un problème ouvert et est considéré comme un défi majeur dans le domaine. La plupart des difficultés proviennent de l'interaction des forces inertielles et visqueuses au voisinage de la frontière non contrôlée, ce qui crée une couche limite difficile à contrôler. Cette question a motivé l'étude de plusieurs modèles à géométrie plus simple en deux dimensions, par exemple [Cha09c], [GIP06] et [GIP12].

Dans ce travail, nous considérons les équations de Burgers généralisées  $(E_\gamma)$ . Ces équations sont une famille d'équations aux dérivées partielles d'évolution non linéaires et unidimensionnelles en espace, dont les solutions présentent un comportement de couche limite au voisinage d'une condition limite de type Dirichlet. D'un point de vue historique, l'équation classique dite de Burgers visqueuse ( $\gamma = 2$  dans  $(E_\gamma)$ ) a été introduite et étudiée par J. M. Burgers en [Bur48]. L'équation de Burgers apparaît naturellement en physique des plasmas, en dynamique des fluides

21. Laboratoire Jacques-Louis Lions, Sorbonne Université

et en écoulement de trafic. Les équations de Burgers généralisées sont une généralisation de l'équation de Burgers classique tout en restant des cas particuliers de lois de conservations visqueuses. Une étude générale de ces équations ainsi que les motivations physiques qui les sous-tendent peuvent être trouvées par exemple dans [Mur70b; Mur70a; SN87; SW99; EVZ93].

Rappelons maintenant quelques-uns des principaux résultats obtenus concernant la contrôlabilité de l'équation de Burgers. A. V. Fursikov et O. Imanuvilov dans [FI96] ont prouvé la contrôlabilité exacte locale en petit temps au voisinage des trajectoires. Leur résultat s'appuie sur des estimées de Carleman et n'utilise qu'un seul contrôle aux limites. Nous étendrons ce résultat aux équations de Burgers généralisées dans la Section VII.5.

La contrôlabilité globale en temps fini vers des états stables avec des contrôles aux limites a été établie dans [FI95]. Plusieurs généralisations ont été prouvées : citons par exemple les travaux de M. Léautaud qui a trouvé une extension pour une large classe de lois de conservation visqueuses dans [Léa12] en s'intéressant à la limite de viscosité évanescence.

Une obstruction à la contrôlabilité globale à zéro en temps petit avec des contrôles aux limites a été trouvée dans [GI07]. Par conséquent, l'utilisation du contrôle interne  $u(t)$  est nécessaire dans  $(E_\gamma)$ . À l'aide de ce nouveau contrôle et de deux contrôles aux limites, M. Chapouly a prouvé la contrôlabilité globale à zéro en temps petit en utilisant des résultats sur l'équation de Burgers non visqueuse dans [Cha09b].

Par la suite, F. Marbach a étendu dans [Mar14] la preuve de la contrôlabilité globale à zéro en temps petit sans le contrôle de la limite droite, c'est-à-dire dans notre cadre décrit par  $(E_\gamma)$ . Un élément clé de la preuve est la transformation de Cole-Hopf ([Col51; Hop50]). Cette transformation réduit l'équation de Burgers à l'équation de la chaleur grâce à un changement de variable.

Dans cette contribution, nous généralisons le travail de F. Marbach aux équations de Burgers généralisées. À notre connaissance, ces équations n'ont pas de transformation équivalente. Par conséquent, dériver des estimations précises sur la couche limite représente l'une des difficultés principales.

### I.7.3 Contribution et idées de preuves

Notre résultat principal est le suivant :

**Théorème I.5.** <sup>22</sup> Soient  $\gamma > 1.5$ ,  $y_0 \in L^\infty(0, 1)$  et  $T > 0$ . Alors, il existe des contrôles  $u \in L^\infty(0, T)$  et  $v \in H^{1/4}(0, T) \cap L^\infty(0, T)$  tels que la solution  $y$  de  $(E_\gamma)$  atteigne zéro au temps  $T$ .

Le schéma de preuve utilise la méthode du retour introduite par J.-M. Coron dans [Cor92] (voir aussi [Cor09]). Plus précisément, nous utiliserons la stratégie en trois étapes développée par F. Marbach dans [Mar14].

- Étape hyperbolique, première partie : nous introduisons  $\vartheta$  l'état propre de  $(E_\gamma)$  avec  $u = 0$  Et  $v = \theta \gg 1$  :

$$\begin{cases} \vartheta_{xx} = (\vartheta^\gamma)_x, \\ \vartheta(0) = \theta, \quad \vartheta(1) = 0. \end{cases} \quad (\text{I.49})$$

Un élément important est que  $\vartheta$  présente une couche limite au voisinage de  $x = 1$ . Ainsi, en utilisant la nature hyperbolique de l'équation de Burgers généralisée lorsqu'elle est gouvernée par le terme non linéaire, nous prouvons que l'on peut se rapprocher de  $\vartheta$  en temps petit.

---

22. Ce théorème est reproduit dans le Chapitre VII sous le nom de Théorème VII.1

- Étape hyperbolique, deuxième partie : en utilisant le contrôle  $v$  qui joue un rôle similaire à la pression, nous prouvons que l'on peut atteindre approximativement zéro à un résidu de couche limite près.
- Étape passive : nous n'appliquons pas de contrôle et attendons la dissipation du résidu de couche limite. L'hypothèse  $\gamma > 1.5$  intervient de manière cruciale dans la preuve.
- Étape parabolique : nous prouvons un résultat de contrôlabilité locale exacte au voisinage de zéro à l'aide d'un argument classique de point fixe.

#### I.7.4 Perspectives et conclusion

Commençons par commenter le fait que l'hypothèse  $\gamma > 1.5$  soit nécessaire pour notre preuve. Ceci peut paraître d'autant plus étrange que le système linéaire ( $\gamma = 1$ ) est globalement contrôlable à zéro en temps petit. Ce point surprenant est lié à la technique de preuve qui consiste à utiliser la méthode du retour en passant par l'état  $\vartheta$ . Afin d'approcher l'état nul, la couche limite de  $\vartheta$  doit se dissiper rapidement lors de l'*étape passive*. Cependant, la taille de la couche limite augmente lorsque  $\gamma$  diminue ; et jusqu'à ne plus exister dans le cas linéaire. Par conséquent, nous pensons que notre méthode de preuve ne peut pas s'étendre à tout l'intervalle  $\gamma \in (1, 1.5]$ . D'autres méthodes nous semblent nécessaires. Par exemple en utilisant des contrôles hautement oscillants, on peut *préparer* la dissipation en annulant certains moments de l'état. Cette méthode a été employée avec succès dans [CMS20] pour les équations de Navier–Stokes avec des conditions de Navier.

Une autre extension intéressante consiste à considérer des fonctions flux  $(f(u))_x$  comme non-linéarité à la place de  $(|u|^\gamma)_x$  et de se demander sous quelles hypothèses le résultat de contrôlabilité globale à zéro en temps court reste valide. L'extension aux fonctions strictement convexes nous semble accessible, le cas général demanderait cependant une analyse précise des solutions stationnaires de (I.49).

Il nous semble également très intéressant de regarder des équations dispersives. M. Chapouly dans [Cha09a] a prouvé que l'équation de Korteweg–De Vries est globalement contrôlable à zéro en temps petit en utilisant deux contrôles aux bords et un contrôle interne uniforme en espace. À notre connaissance, savoir si ce résultat persiste sans l'utilisant d'un contrôle de type Dirichlet à droite (et en utilisant éventuellement un contrôle sur la dérivée à droite) est toujours une question ouverte.





# First Part: Optimizations for Stellarators



## Chapter II

# Optimal shape of stellarators for magnetic confinement fusion

This chapter is taken from the following article (also referred as [PRS22b]):

Y. Privat, R. Robin, and M. Sigalotti. “Optimal shape of stellarators for magnetic confinement fusion”. In: *Journal de Mathématiques Pures et Appliquées* 163 (2022), pp. 231–264

We are interested in the design of stellarators, devices for the production of controlled nuclear fusion reactions alternative to tokamaks. The confinement of the plasma is entirely achieved by a helical magnetic field created by the complex arrangement of coils fed by high currents around a toroidal domain. Such coils describe a surface called “coil winding surface” (CWS). In this chapter, we model the design of the CWS as a shape optimization problem, so that the cost functional reflects both optimal plasma confinement properties, through a least square discrepancy, and also manufacturability, thanks to geometrical terms involving the lateral surface or the curvature of the CWS.

We completely analyze the resulting problem: on the one hand, we establish the existence of an optimal shape, prove the shape differentiability of the criterion, and provide the expression of the differential in a workable form. On the other hand, we propose a numerical method and perform simulations of optimal stellarator shapes. We discuss the efficiency of our approach with respect to the literature in this area.

## II.1 Introduction

### II.1.1 Motivations: towards a shape optimization problem

Nuclear fusion is a nuclear reaction involving the use of light nuclei. In order to produce energy by nuclear fusion, high temperature plasmas<sup>1</sup> must be produced and confined. For these reactions to occur, the nuclei must get close to each other at very small distances. They must therefore overcome the Coulomb repulsion. This happens naturally in a plasma during collisions if the

---

1. This is a particular state of matter when it becomes totally *ionized*, i.e., when all its atoms have lost one or more peripheral electrons. This is the most common state of matter in the universe because it is found (at 99%) in the stars, the interstellar medium, and earth’s ionosphere.

energy of the nuclei is sufficient. This is the objective of devices called tokamaks, steel magnetic confinement chambers that allow a plasma to be controlled in order to study and experiment with energy production by nuclear fusion. The magnetic confinement technique allows to maintain a sufficient temperature and density of the plasma, in an intense magnetic field. The simplest configuration for the magnetic field is the toroidal solenoid; this is the configuration found in most current experiments.

Unfortunately, the magnetic field is not uniform in general, which causes a vertical drift of the particles, in opposite directions for the ions and for the electrons. This charge separation creates a vertical electric field which, in turn, causes the particles to drift out of the torus. This phenomenon dramatically reduces the confinement. To get around this obstacle, the effect of such drifts is canceled by giving a poloidal component<sup>2</sup> to the magnetic field: the field lines are wound on nested toroids. Thus, the particles, following the magnetic field lines, have their vertical drift cancelled at each turn. In a tokamak, the poloidal magnetic field is created by a toroidal electric current circulating in the plasma. This current is called *plasma current*.

A possible alternative to correct the problems of drift of magnetically confined plasma particles in a torus is to modify the toroidal shape of the device, by breaking the axisymmetry, yielding to the concept of stellarator. A stellarator is analogous to a tokamak except that it does not use a toroidal current flowing inside the plasma to confine it. The poloidal magnetic field is generated by external coils, or by a deformation of the coils responsible for the toroidal magnetic field. This system has the advantage of not requiring plasma current and therefore of being able to operate continuously; but it comes at the cost of more complex coils (non-planar coils) and of a more important neoclassical transport [HS05].

The confinement of the plasma is then entirely achieved by a helical magnetic field created by the complex arrangement of coils around the torus, supplied with strong currents and called poloidal coils.

Despite the promise of very stable steady-state fusion plasmas, stellarator technology also presents significant challenges related to the complex arrangement of magnetic field coils. These magnetic field coils are particularly expensive and especially difficult to design and fabricate due to the complexity of their spatial arrangement.

In this chapter, we are interested in the search for the optimal shape of stellarators, i.e., the best coil arrangement (provided that it exists) to confine the plasma. In general, two steps are considered: first, the shape of the plasma boundary is determined in order to optimize the physical properties, among which the neoclassical transport and the magnetohydrodynamic (MHD) stability. In a second step, we search for the coil shapes producing approximately the "target" plasma shape resulting from the previous step.

In this chapter, we focus entirely on the second step, assuming that the target magnetic field  $B_T$  is known. It is then convenient to define a coil winding surface (CWS) on which the coils will be located (see Figure II.1). The optimal arrangement of stellarator coils corresponds then to the determination of a closed surface (the CWS) chosen to guarantee that the magnetic field created by the coils is as close as possible to the target magnetic field  $B_T$ . Of course, it is necessary to consider feasibility and manufacturability constraints. We will propose and study several relevant choices of such constraints in what follows.

---

2. The terms toroidal and poloidal refer to directions relative to a torus of reference. The poloidal direction follows a small circular ring around the surface, while the toroidal direction follows a large circular ring around the torus, encircling the central void.

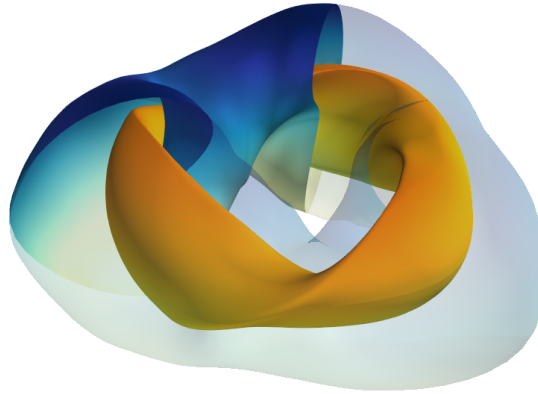


Figure II.1 – CWS (blue and white) and plasma surface (orange) of the National Compact Stellarator Experiment (NCSX) designed by the Princeton Plasma Physics Laboratory. There is a three-folds discrete symmetry in the design.

### II.1.2 State of the art and main contributions of this chapter

The question of determining the best location of coils around a stellarator, reformulated as an optimal surface problem, is a major issue for the construction of stellarators with efficient confinement properties. The physical and mathematical literature dedicated to plasmas is rich of references on this issue. We mention hereafter a non-exhaustive list of various important contributions around this problem. Let us first mention [IPW19], where all the basic theoretical elements to understand the modeling of stellarator magnetic fields are gathered.

Regarding optimal design issues, let us distinguish between several optimization/optimal control approaches and modeling choices. Each discrete stellarator coil can be represented as a closed one-dimensional curve embedded in  $\mathbb{R}^3$  [Zhu+18c; Zhu+18b; Zhu+18a]. In these references, several optimization methods are tested among which the steepest descent and Newton like methods.

Another common choice consists in using the aforementioned CWS, in other words to define a closed toroidal winding surface enclosing the plasma surface on which all coils lie. Two kinds of issues related to the optimal design of stellarators can then be addressed. The simplest is to assume the CWS to be given, and to look for currents on this surface generating the desired magnetic field for confining the plasma. Indeed, in the limit of a large number of coils, a set of discrete coils can be described by a continuous current density on the CWS. Let us mention NESCOIL [Mer86; Pom+01], where the current potential representing a surface current distribution is sought such that the normal component of the magnetic field vanishes in a least-squares sense at the plasma boundary. In the same vein, REGCOIL [Lan17] improves the NESCOIL approach by adding a Tikhonov regularization term in the minimization functional whereas COILOPT [SBH02] uses an explicit representation of modular coils on a toroidal winding surface. A review of such approaches can be found in [Gat+17]. Recently, a similar problem where an extra Laplace forces penalization term is taken into account has been investigated in [RV22].

A much more difficult problem is to determine the CWS and the density current distribution at the same time. This is expected to improve the performances of the resulting device. On

the other hand, this approach requires solving a dual optimization problem, including a rather challenging surface optimization problem. This is the main purpose of this chapter. In the following we mention some of the many contributions on this topic and position our contribution through this literature. In [Pau+18], this problem is modeled using a cost functional written as the weighted sum of four terms: the first one is the surface-integrated-squared normal magnetic field on the desired plasma surface. The second is the opposite of the total volume enclosed by the coil-winding surface, acting to enforce the coil-plasma separation. The third one is a measure of the spectral width of the Fourier series describing the coil-winding surface. This allows to overcome the non-uniqueness of the Fourier series representation of the coil-winding surface. The last one is the  $L^2$  norm of the current density, allowing to obtain coils with good manufacturing properties. It is important to note here, and this is related to the motivation for this chapter, that the approach developed in [Pau+18] rests upon a (truncated) Fourier series parameterization of the surface equation. The authors thus compute derivatives of their cost with respect to these Fourier coefficients.

In [Pau+19], a more complex model involving a drift kinetic equation is considered and similar shape optimization issues are investigated.

In what follows, we propose a continuous approach, which does not rely on any parameterization of the surfaces involved. We use the notion of Hadamard variation and shape derivative. We rigorously analyze, in a continuous framework, the sensitivity with respect to the domain of a REGCOIL-type cost. We thus obtain intrinsic expressions with respect to any parametrization. This makes our approach flexible and the formulas obtained by using developments of the parametric equation of the surfaces in Fourier series can be adapted without any difficulty to other choices of parametrization. We also propose several choices of manufacturability terms in the cost functional and discuss their relevance.

The issues addressed in the following as well as our main contributions are summed-up hereafter:

- **Modeling of the problem (Section II.1.4).** Using the CWS concept, we propose a continuous formulation of the question of the best coil arrangement as a shape optimization problem, regardless of any surface parametrization. In particular, several choices of manufacturing constraints are proposed. They are integrated to the cost functional using a penalization/regularization term. From the mathematical point of view, the main issue comes to minimize a functional involving the trace of the solution of an elliptic partial differential equation (PDE) on a manifold, under geometrical constraints involving the distance to this manifold.
- **Analysis of the shape optimization problem (Sections II.2 and II.3).** Having in mind the determination of an efficient algorithm for finding an optimal shape for the above problem, we focus mainly on two questions. The first one is dedicated to the existence of an optimal shape (Section II.2.2). In this context, the developed approach is not completely standard and requires to carefully establish semicontinuity properties of the trace of the solution of the PDE on manifolds satisfying a uniform regularity property. The second one concerns the establishment of optimality conditions using the notion of shape derivative (Section II.3.1). Here again, due to the particular nature of the PDE at stake, the classical approach cannot be used in a direct way and many adaptations are necessary. We establish a workable expression of this derivative, which is the basis of the numerical approaches developed in the next section.
- **Numerical implementation (Section II.4).** A relevant aspect of this chapter is that the study of the sensitivity of the studied criterion to a variation of the shape of the stellarator is carried out without using any parameterization of the surface to be designed. As a result, the sensitivity relations obtained at the end of the previous step are totally

intrinsic with respect to any parameterization of the surface. As a consequence, we can apply the more robust “optimize then discretize” approach, instead of a “discretize then optimize” procedure as in most of the methods implemented for this application. The shape derivatives constitute the basis of a quasi-Newton optimization method that we implement by using a parametric representation of the surface in terms of Fourier series.

### II.1.3 Notations

In what follows, the notation  $S$  is used to denote a  $\mathcal{C}^{1,1}$  toroidal surface<sup>3</sup> in  $\mathbb{R}^3$ , equipped with the Riemannian metric induced by the canonical embedding  $i_S : S \hookrightarrow \mathbb{R}^3$ , i.e., the scalar product between two vectors  $v$  and  $w$  tangent to  $S$  at a common point is equal to  $\langle v, w \rangle$ , where  $\langle \cdot, \cdot \rangle$  denotes the Euclidean scalar product in  $\mathbb{R}^3$ . We also denote by  $\mu_S$  the associated Riemannian volume form, which coincides with the two-dimensional Hausdorff measure on  $S$  ([Fed69, Theorem 2.10.10] and [EG92, Theorem 2 in Section 2.2]). We write  $V$  to denote the bounded domain of  $\mathbb{R}^3$  such that  $S = \partial V$ .

Throughout this chapter, we use the following notation:

- For  $n \in \mathbb{N}^* = \{1, 2, \dots\}$  and any integer  $m \geq n$ ,  $\mathcal{H}^n$  denotes the  $n$ -dimensional Hausdorff measure in  $\mathbb{R}^m$ ;
- $P$  denotes a smooth toroidal domain<sup>4</sup> of  $\mathbb{R}^3$  standing for the plasma domain;
- $\mathfrak{X}(S)$  denotes the set of smooth tangent vector fields on  $S$ ;
- $\mathcal{F}_S = L^2(\Gamma(TS))$  denotes the completion of  $\mathfrak{X}(S)$  for the inner product

$$\forall (X_1, X_2) \in \mathfrak{X}(S)^2, \quad \langle X_1, X_2 \rangle_{\mathcal{F}_S} = \int_S \langle X_1, X_2 \rangle d\mu_S;$$

- $\times$  denotes the cross product in  $\mathbb{R}^3$ ;
- Given a function  $F : \mathbb{R}^{n_1} \rightarrow \mathbb{R}^{n_2}$  and  $x \in \mathbb{R}^{n_1}$ ,  $DF(x)$  denotes the  $n_1 \times n_2$  Jacobian matrix of  $F$ . In the case where  $n_1 = n_2$ ,  $|DF|$  stands for the absolute value of the determinant of  $DF$ . The symbol  $D_x$  is used to denote the Jacobian operator with respect to the (vector) variable  $x$ ;
- $\nabla_S$  denotes the *tangential gradient* to  $S$  in  $\mathbb{R}^3$ , defined for every differentiable function  $f : \mathbb{R}^3 \rightarrow \mathbb{R}$  by  $\nabla_S f = \nabla f - \langle \nabla f, \nu \rangle \nu$  on  $S$ , where  $\nu$  stands for the outward normal vector to  $V$ . Similarly, the notation  $\operatorname{div}_S$  stands for the *tangential divergence* given by  $\operatorname{div}_S \theta = \operatorname{div} \theta - \langle D\theta \nu, \nu \rangle$  on  $S$ , where  $\theta$  is a vector field on  $\mathbb{R}^3$ ;
- The norm on  $\mathfrak{X}(S)$  induced by the inner product  $\langle \cdot, \cdot \rangle_{\mathcal{F}_S}$  is denoted  $\| \cdot \|_{\mathcal{F}_S}$ ;
- $\mathcal{F}_S^0$  is the closure under the norm  $\| \cdot \|_{\mathcal{F}_S}$  of the (tangential) divergence-free vectors of  $\mathfrak{X}(S)$ ;
- The flat two-dimensional torus is denoted by  $T = (\mathbb{R}/\mathbb{Z})^2$ .  $\mathfrak{X}(T)$ ,  $\mathcal{F}_T$  and  $\mathcal{F}_T^0$  are defined similarly to what has been done above;
- The Hausdorff distance  $d_V$  and the signed distance  $b_V$  from  $V$  are defined as:

$$d_V(x) = \inf_{y \in V} |x - y|, \quad b_V(x) = d_V(x) - d_{\mathbb{R}^3 \setminus V}(x);$$

- If  $h > 0$ , the  $h$ -*tubular neighborhood*  $U_h(V)$  of  $V$  is the level set

$$U_h(V) = \{x \in \mathbb{R}^3 \mid d_V(x) < h\}$$

3. By *toroidal surface*, we mean here the range of the toroidal solenoid by a homeomorphism. In what follows, we will rather consider smooth toroidal surfaces, where the wording “smooth” refers to at least  $\mathcal{C}^{1,1}$  regularity.

4. *toroidal domain* stands for any three-dimensional domain whose boundary is a toroidal surface



of  $d_V$ ;

- The *reach* of  $V$  [Fed69] is given by

$$\text{Reach}(V) = \sup\{h > 0 \mid d_V \text{ is differentiable on } U_h(V) \setminus \bar{V}\}.$$

More explanations are provided in Appendix II.A.2. For more exhaustive informations about this notion, we refer to [Fed69] and [DZ11, Sect. 6.6];

- For two Banach spaces  $E$  and  $F$ , we denote by  $\mathcal{L}(E, F)$  the Banach space of continuous linear maps from  $E$  to  $F$  and by  $\mathcal{L}(E)$  the Banach space of continuous endomorphisms;
- The adjoint of a linear operator  $L$  is denoted by  $L^\dagger$ ;
- If  $A$  and  $B$  denote two matrices in  $\mathbb{M}_3(\mathbb{R})$ , we define their *doubly contracted product* as

$$A : B = \sum_{i,j=1}^3 A_{ij} B_{ij};$$

- $I_3$  denotes the identity matrix in  $\mathbb{R}^3$ .

#### II.1.4 Modeling: towards a shape optimization problem

Since we are interested in solving a shape optimization problem whose unknown is the coil winding surface  $S$ , we are led to make some assumptions on  $S$  motivated by the application under consideration. In particular, we assume in what follows that the distance  $d(S, P)$  between  $S$  and the plasma domain  $P$  is uniformly bounded from below, namely, we fix  $\delta > 0$  and we require that

$$d(S, P) = \inf_{x \in S, y \in P} |x - y| = \inf_{x \in S} d_P(x) \geq \delta. \quad (\mathcal{H}_{\text{dist}, P, \delta})$$

We now introduce the main operator we will deal with, which plays a crucial role in electromagnetism: the so-called *Biot and Savart* operator. This operator associates with each current distribution on  $S$  the corresponding magnetic field in  $P$ . It can be considered as a kind of inverse of the curl operator.

**Definition II.1** (The Biot and Savart operator  $\text{BS}_S$  [EGP18]). *Let  $S$  be a smooth two-dimensional manifold. Let  $\delta_S$  denote the single layer distribution supported on  $S$  defined by*

$$\forall \varphi \in \mathcal{C}_c^\infty(\mathbb{R}^3, \mathbb{R}^3), \quad \forall X \in \mathcal{F}_S, \quad \langle X \delta_S, \varphi \rangle = \int_S \langle \varphi, X \rangle d\mu_S.$$

Let us fix  $X \in \mathcal{F}_S$  and denote by  $u$  the unique distributional solution of the PDE

$$\begin{cases} \nabla \times u = X \delta_S & \text{in } \mathcal{D}'(\mathbb{R}^3) \\ \langle \nabla, u \rangle = 0 \end{cases}$$

that falls off at infinity, i.e.,

$$u(y) = \int_S \frac{(x - y) \times X(x)}{|x - y|^3} d\mu_S(x), \quad y \in \mathbb{R}^3 \setminus S.$$

Then the Biot and Savart operator is defined as the map  $\text{BS}_S : \mathcal{F}_S \rightarrow L^2(P, \mathbb{R}^3)$  associating

with  $X$  the restriction of  $u$  to the plasma domain  $P$ . By introducing the kernel  $K$  given by

$$K : [\mathbb{R}^3]^2 \setminus \{(x, x) \mid x \in \mathbb{R}^3\} \longrightarrow \mathbb{R}^3$$

$$(x, y) \longmapsto \frac{x - y}{|x - y|^3},$$

one has

$$\text{BS}_S(X)(y) = \int_S K(x, y) \times X(x) d\mu_S(x), \quad y \in P. \quad (\text{II.1})$$

**Remark II.2.** According to  $(\mathcal{H}_{\text{dist}, P, \delta})$ , the restriction of  $K$  to  $S \times P$  is uniformly bounded. By standard regularity results for parameterized integrals, the mapping  $P \ni y \mapsto \text{BS}_S(X)(y)$  is smooth and the operator  $\text{BS}_S$ , seen as going from  $\mathcal{F}_S$  to  $\mathcal{C}^k(P, \mathbb{R}^3)$  with  $k \in \mathbb{N} \cup \{+\infty\}$ , is continuous. As a consequence, the operator  $\text{BS}_S : \mathcal{F}_S \rightarrow L^2(P, \mathbb{R}^3)$  is compact.

In what follows, we will use several times that for every  $x, y, h \in \mathbb{R}^3$  with  $x \neq y$  we have

$$D_x K(x, y)(h) = \lim_{\varepsilon \searrow 0} \frac{K(x + \varepsilon h, y) - K(x, y)}{\varepsilon} = \frac{h}{|x - y|^3} - \frac{3\langle (x - y), h \rangle (x - y)}{|x - y|^5}. \quad (\text{II.2})$$

**Computation of the optimal current  $j$ .** In view of modeling the optimal design problem we will deal with, let us now introduce a target magnetic field  $B_T \in L^2(P, \mathbb{R}^3)$ .

The target magnetic field  $B_T$  being given, we model the optimal design of a stellarator problem as a kind of regularized least square problem, where one aims at determining both the current  $j$  and the manifold shape  $S$  leading to the magnetic field closest to  $B_T$  on  $P$ . To this aim, and according to the REGCOIL procedure [Lan17], we introduce the shape functional  $C$  defined, for every closed smooth two-dimensional manifold  $S$ , as

$$C(S) = \inf_{j \in \mathcal{F}_S^0} \|\text{BS}_S j - B_T\|_{L^2(P, \mathbb{R}^3)}^2 + \lambda \|j\|_{\mathcal{F}_S}^2, \quad (\mathcal{P}_S)$$

where  $\lambda > 0$  denotes a regularization parameter.

**The shape optimization problem.** To state the shape optimization problem that we will consider, let us first define the set of admissible manifolds. We gather hereafter several conditions evoked previously that we will take into account in the search of the CWS.

- *Topology and uniform boundedness.* To preserve the topology of the device (see Footnote 3), we will only consider CWSs that are two-dimensional closed toroidal manifolds. Moreover, we will fix a compact set  $D$  of  $\mathbb{R}^3$  and we require the CWS to be contained in  $D$ .
- *Uniform distance constraint of the coils to the plasma.* To build the vacuum vessel around the plasma, we will assume that the CWS satisfies assumption  $(\mathcal{H}_{\text{dist}, P, \delta})$ .
- *Manufacturing cost.* In order to avoid irregular shapes that are too difficult to build, we will assume that the CWS has a minimal regularity, say  $\mathcal{C}^{1,1}$ , and a minimal reach condition. More precisely, we will assume that the reach of the CWS is uniformly bounded from below by some  $r_{\min} > 0$ . We recall that this condition imposes that the curvature radii (where they can be defined) are larger than  $r_{\min}$  and that there is no bottleneck of distance smaller than  $2r_{\min}$  (see, e.g., [Aam+19, Figure 3]). To sum-up,

$$S \text{ is a } \mathcal{C}^{1,1} \text{ closed toroidal surface such that } \text{Reach}(S) \geq r_{\min} > 0. \quad (\mathcal{H}_{\text{reach}, r_{\min}})$$

Notice that we would have a weaker constraint if the lower bound was imposed on  $\text{Reach}(V)$  instead of  $\text{Reach}(S)$ , where the volume  $V$  is such that  $S = \partial V$ .

As it will be emphasized in what follows, the regularity assumption is actually a consequence of the reach constraint: indeed, the class of sets satisfying a ‘‘Reach’’ constraint is closed in a sense to be specified later and all elements are of class  $\mathcal{C}^{1,1}$ .

Other constraints such as a bound on the two-dimensional Hausdorff measure  $\mathcal{H}^2(S)$  of  $S$  (in other words the perimeter of the stellarator in  $\mathbb{R}^3$ ) will also be considered:

$$\mathcal{H}^2(S) \leq P_{\max}. \quad (\mathcal{H}_{\text{Perim}, P_{\max}})$$

To sum-up, let us introduce the admissible set of shapes we will deal with in what follows:

$$\mathcal{O}_{\text{ad}} = \{S = \partial V \subset D \mid P \subset V \text{ and } S \text{ satisfies } (\mathcal{H}_{\text{dist}, P, \delta}), (\mathcal{H}_{\text{reach}, r_{\min}}), (\mathcal{H}_{\text{Perim}, P_{\max}})\}.$$

The resulting shape optimization problem we will consider reads

$$\boxed{\inf_{S \in \mathcal{O}_{\text{ad}}} C(S)}. \quad (\mathcal{P}_{\text{shape}})$$

In the two following sections, we investigate two important aspects of the shape optimization problem ( $\mathcal{P}_{\text{shape}}$ ). The first one concerns the existence of optimal shapes and is investigated in Section II.2. The second one is related to the derivation of first order optimality conditions, at the heart of the algorithms implemented in the last section of this article. To this aim, we apply in Section II.3.1 the so-called Hadamard boundary variation method recalled at the beginning of Section II.3.

## II.2 Existence issues for Problem ( $\mathcal{P}_{\text{shape}}$ )

### II.2.1 Existence of an optimal current for a given shape (Solving of Problem ( $\mathcal{P}_S$ ))

We first establish that the infimum defining ( $\mathcal{P}_S$ ) is in fact a minimum. Moreover, the minimizer is unique.

**Lemma II.3.** *Let  $S \in \mathcal{O}_{\text{ad}}$ . The optimization problem ( $\mathcal{P}_S$ ) has a unique minimizer  $j_S$ . Moreover, one has*

$$\begin{aligned} j_S &= (\lambda \text{Id} + \text{BS}_S^\dagger \text{BS}_S)^{-1} \text{BS}_S^\dagger B_T, \\ C(S) &= \lambda \|(\lambda \text{Id} + \text{BS}_S^\dagger \text{BS}_S)^{-1} \text{BS}_S^\dagger B_T\|_{\mathcal{F}_S}^2 + \|\text{BS}_S(\lambda \text{Id} + \text{BS}_S^\dagger \text{BS}_S)^{-1} \text{BS}_S^\dagger B_T - B_T\|_{L^2(P, \mathbb{R}^3)}^2. \end{aligned} \quad (\text{II.3})$$

*Proof.* First observe that  $\mathcal{F}_S^0$  is a Hilbert space and that the mapping  $\mathcal{F}_S^0 \ni j \mapsto \|\text{BS}_S j - B_T\|_{L^2(P, \mathbb{R}^3)}^2 + \lambda \|j\|_{\mathcal{F}_S}^2$  is strongly convex, since it is the sum of the convex functional  $j \mapsto \|\text{BS}_S j - B_T\|_{L^2(P, \mathbb{R}^3)}^2$  and the strongly convex one  $j \mapsto \lambda \|j\|_{\mathcal{F}_S}^2$ . Furthermore, we claim that the functional  $\mathcal{F}_S^0 \ni j \mapsto \|\text{BS}_S j - B_T\|_{L^2(P, \mathbb{R}^3)}^2 + \lambda \|j\|_{\mathcal{F}_S}^2$  is lower semicontinuous for the strong topology of  $\mathcal{F}_S^0$ . Indeed, let  $(j_n)_{n \in \mathbb{N}}$  denote a sequence of  $\mathcal{F}_S^0$  converging to  $j \in \mathcal{F}_S^0$ . According

to  $(\mathcal{H}_{\text{dist},P,\delta})$ ,  $K$  is uniformly bounded in  $S \times P$ , yielding

$$\begin{aligned} \forall y \in P, \quad \left| \int_S K(x, y) \times j_n(x) d\mu_S(x) - \int_S K(x, y) \times j(x) d\mu_S(x) \right| &\leq \|K(\cdot, y)\|_{\mathcal{F}_S} \|j_n - j\|_{\mathcal{F}_S} \\ &\leq C \|j_n - j\|_{\mathcal{F}_S}, \end{aligned}$$

for some constant  $C$  independent of  $y$ . It follows that the functional to minimize is lower semi-continuous (and even continuous) in  $\mathcal{F}_S^0$ . By convexity it is also lower semicontinuous for the weak topology, whence the existence of a unique minimizer  $j_S$  for Problem ( $\mathcal{P}_S$ ).

Since  $\text{BS}_S$  is continuous, its adjoint  $\text{BS}_S^\dagger$  is well defined on  $\mathcal{F}_S^0$ . It is hence standard that the first order optimality condition for this problem reads

$$\forall v \in \mathcal{F}_S^0, \quad \langle \text{BS}_S v, \text{BS}_S j_S - B_T \rangle_{L^2(P, \mathbb{R}^3)} + \lambda \langle v, j_S \rangle_{\mathcal{F}_S^0} = 0 \quad (\text{II.4})$$

which also rewrites

$$\forall v \in \mathcal{F}_S^0, \quad \langle v, \lambda j_S + \text{BS}_S^\dagger (\text{BS}_S j_S - B_T) \rangle_{\mathcal{F}_S^0} = 0.$$

Since  $v$  is arbitrary in  $\mathcal{F}_S^0$ , we thus infer that  $\lambda j_S + \text{BS}_S^\dagger \text{BS}_S j_S = \text{BS}_S^\dagger B_T$ . The operator  $\text{BS}_S^\dagger \text{BS}_S$  is compact and symmetric. Besides its spectrum is positive and we can therefore consider its resolvent for negative real numbers  $-\lambda$  with  $\lambda > 0$ , so that

$$j_S = (\lambda \text{Id} + \text{BS}_S^\dagger \text{BS}_S)^{-1} \text{BS}_S^\dagger B_T.$$

The expression of  $C(S)$  given in (II.3) follows from a straightforward computation.  $\square$

**Remark II.4.** *When confronted with the numerical implementation of the shape optimization, motivated by the structure of  $\mathcal{F}_S^0$  and the properties of the in vacuo Maxwell-equations, we will find it useful to:*

- optimize on a closed affine subset  $j_S^a + \hat{\mathcal{F}}_S^0 \subset \mathcal{F}_S^0$  instead of the entire set  $\mathcal{F}_S^0$ . We refer to Section II.4.1.3 for further details;
- consider only divergence-free and curl-free target magnetic fields  $B_T$ . Not only does the image of the Biot and Savart operator satisfies these properties, but also the orthogonal Hodge decomposition (see, e.g., [CDG01]);
- replace the target magnetic field  $B_T : P \rightarrow \mathbb{R}^3$  by its normal component on the plasma surface (thus, by an object in  $L^2(\partial P, \mathbb{R})$ ). Indeed, a divergence-free and curl-free vector field on a 3D domain (in absence of electric currents in the plasma) is entirely characterized, up to fixing the circulation of the vector field along a toroidal loop, by its normal component on the boundary. Further details are given in Section II.4.1.2 and Appendix II.A.1.3.

Nevertheless, such changes have a minor impact on the theoretical discussion on the shape optimization process and we believe that, for the sake of clarity, it is better to postpone the details about such modifications to Section II.4.

## II.2.2 Existence of an optimal shape

**Theorem II.5.** *The shape optimization problem ( $\mathcal{P}_{\text{shape}}$ ) has at least one solution.*

The proof follows the direct method of the calculus of variation. Most of the compactness on our set of admissible shapes comes from the bounded reach assumption ( $\mathcal{H}_{\text{reach}, r_{\min}}$ ). In particular, the following Lipschitz estimate is crucial.

**Lemma II.6** (Theorem 2.8 of [Dal18]). *Let  $V \subset \mathbb{R}^n$  be a nonempty set such that  $\text{Reach}(\partial V) \geq r_{\min}$  and  $\mathcal{H}^n(\partial V) = 0$ . Then, for every  $r \in (0, r_{\min})$ , the gradient  $\nabla b_V$  of the signed distance function is  $\frac{2}{r_{\min}-h}$ -Lipschitz on the tubular neighborhood  $U_r(\partial V)$ .*

With this estimate, we can state the following compactness result.

**Lemma II.7.** *Let  $(S_n)_{n \in \mathbb{N}} = (\partial V_n)_{n \in \mathbb{N}}$  be a sequence in  $\mathcal{O}_{\text{ad}}$ . Then, there exists  $S_\infty = \partial V_\infty \in \mathcal{O}_{\text{ad}}$  such that, up to a subsequence, for any  $r$  in  $(0, r_{\min})$*

- $b_{V_\infty}$  is in  $\mathcal{C}^{1,1}(\overline{U_r(S_\infty)})$  and  $(b_{V_n})_{n \in \mathbb{N}}$  converges to  $b_{V_\infty}$  in  $\mathcal{C}^1(\overline{U_r(S_\infty)})$ ;
- $(b_{V_n})_{n \in \mathbb{N}}$  converges to  $b_{V_\infty}$  in  $\mathcal{C}(\overline{D})$ ;
- $(d_{S_n})_{n \in \mathbb{N}}$  converges to  $d_{S_\infty}$  in  $\mathcal{C}(\overline{D})$ ;
- $(\mathcal{H}^2(S_n))_{n \in \mathbb{N}}$  converges to  $\mathcal{H}^2(S_\infty)$ .

*Proof.* Compactness properties among Hausdorff distances from sets of uniformly positive reach are well known and remain valid for the signed distance (see, e.g., [DZ11, Chapter 6]). Besides, as stated in [Dal18], the convergence property holds true for the strong topology of  $\mathcal{C}^{1,\alpha}$  (for  $\alpha < 1$ ) and for the weak topology of  $W^{2,\infty}$  in a tubular neighborhood of  $S_\infty$ . As a consequence,  $d(S_\infty, P) \geq \delta$  and  $\text{Reach}(\partial V_\infty) \geq r_{\min}$ . In particular, thanks to Lemma II.6,  $b_{V_\infty}$  is  $\mathcal{C}^{1,1}$  on  $\overline{U_r(S_\infty)}$ . The fact that  $S_\infty$  remains a toroidal domain is proved, e.g., in [Dal18, Proposition 4.22]. Finally, the convergence of  $\mathcal{H}^2(S_n)$  to  $\mathcal{H}^2(S_\infty)$  follows from standard results on the continuity of  $S \mapsto \mathcal{H}^2(S)$  (see [Dal18] or [GY13]).  $\square$

The end of this section is devoted to the proof of Theorem II.5. Let  $(S_n)_{n \in \mathbb{N}} = (\partial V_n)_{n \in \mathbb{N}}$  be a minimizing sequence for Problem  $(\mathcal{P}_{\text{shape}})$ . Denote by  $S_\infty$  a closure point of this sequence in the sense of Lemma II.7. In what follows, we will still denote by  $(S_n)_{n \in \mathbb{N}}$  the converging subsequence introduced in Lemma II.7.

We will proceed by showing a semicontinuity property of the criterion, namely that

$$\liminf_{n \rightarrow +\infty} C(S_n) \geq C(S_\infty).$$

Let  $j_n$  denote the minimizer of Problem  $(\mathcal{P}_S)$  for the surface  $S_n$ , whose existence is provided by Lemma II.3.

The idea is to consider a volume integral as approximation of the surface integral in the same spirit as in [Del00]. For this purpose, we need to extend locally  $j_n$  to a volume around  $S_n$ . Notice that, without loss of generality,  $S_n$  is contained in  $U_{r_{\min}}(S_\infty)$  for every  $n$ , which implies, in particular, that  $\nabla b_{V_\infty}$  is everywhere defined and Lipschitz continuous on  $S_n$ . Let  $r > 0$  be a small constant to be fixed later and define the map

$$\begin{aligned} T_n : (-r, r) \times S_n &\rightarrow A_r(S_n) \subset U_h(S_n) \\ (t, x) &\mapsto x + t \nabla b_{V_\infty}(x), \end{aligned}$$

where  $A_r(S_n)$  denote the image of  $T_n$ . Notice that  $T_n$  is a bijection between  $(-r, r) \times S_n$  and  $A_r(S_n)$  if the latter is contained in  $U_{r_{\min}}(S_\infty)$  (see Figure II.2).

The differential of  $T_n$  at  $(t_0, x_0) \in (-r, r) \times S_n$  reads

$$\begin{aligned} DT_n(t_0, x_0) : \mathbb{R} \times T_{x_0} S_n &\rightarrow \mathbb{R}^3 \\ (s, y) &\mapsto s \nabla b_{V_\infty}(x_0) + y + t_0 \nabla_{S_n}(\nabla b_{V_\infty})(x_0) y, \end{aligned}$$

where  $\nabla_{S_n}(\nabla b_{V_\infty})(x_0)$  is a  $3 \times 3$  matrix, according to the notation introduced in Section II.1.3, and  $T_{x_0} S_n$  is identified with a linear subspace of  $\mathbb{R}^3$ . We can identify  $DT_n(t_0, x_0)$  with a  $3 \times 3$  matrix by choosing an orthogonal basis on  $T_{x_0} S_n$  and its determinant, denoted  $|DT_n(t_0, x_0)|$  in what follows, is independent of such a choice.

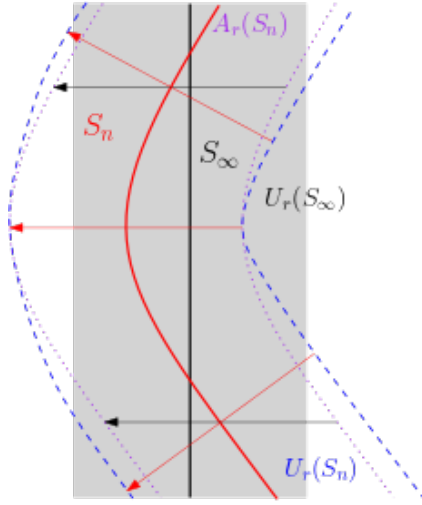


Figure II.2 – This figure illustrates the difference between  $U_r(S_\infty)$  filled in grey,  $U_r(S_n)$  (resp.,  $A_r(S_n)$ ) delimited by the blue dashed (resp., purple dotted) curves. The black arrows represent the field  $\nabla b_{V_\infty}$  and the red ones represent  $\nabla b_{V_n}$ . Note that both  $V_n$  and  $V_\infty$  are on the right of the figure.

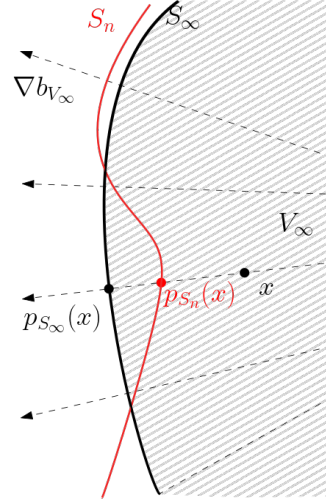


Figure II.3 –  $p_{S_n}(x)$  is obtained by taking the intersection of the flow of  $\nabla b_{V_\infty}$  and  $S_n$ . Whereas the standard projector (in the sense of shortest distance) on  $S_n$  is obtained by using the flow of  $\nabla b_{V_n}$ .

Using the regularity of  $\nabla b_{V_\infty}$  near the surface  $S_\infty$ , one can prove the following crucial estimate.

**Lemma II.8.** *For every  $\varepsilon > 0$ , there exists  $r = r_\varepsilon > 0$  such that*

$$1 - \varepsilon \leq |DT_n(y)| \leq 1 + \varepsilon, \quad \text{for a.e. } (t_0, x_0) \in (-r, r) \times S_n$$

for every  $n \in \mathbb{N} \cup \{\infty\}$  large enough.

*Proof.* By Lemma II.7,  $b_{V_\infty}$  is  $\mathcal{C}^{1,1}$  in a neighborhood of  $S_\infty$ , thus  $\nabla b_{V_\infty}|_{S_n}$  is Lipschitz continuous for  $n$  large enough with a Lipschitz constant independent of  $n$  and, in particular, there exists  $C > 0$  such that for all  $n$  large enough

$$|\nabla_{S_n}(\nabla b_{V_\infty})|_{L^\infty(S_n)} \leq C.$$

Besides, the linear mapping  $R_{x_0} : (-r, r) \times T_{x_0}S_n \ni (s, y) \mapsto s\nabla b_{V_\infty}(x_0) + y$  is direct and orthogonal since  $\nabla b_{V_\infty}(x_0)$  is the unit normal outward vector. Hence, its determinant is equal to 1. Since the determinant is  $\mathcal{C}^\infty$ , we have

$$\sup_{(t_0, x_0) \in (-r, r) \times S_n} ||DT_n(t_0, x_0)| - \det R_{x_0}| = \mathcal{O}(h),$$

where the reminder term is uniformly bounded with respect to  $n$  for  $n$  large enough. This concludes the proof.  $\square$

In what follows  $\varepsilon > 0$  is a small parameter to be fixed and  $h$  is as in the statement of Lemma II.8. We shall also assume that  $A_r(S_n) \subset U_{r_{\min}}(S_\infty)$  for every  $n$ . Notice that, as soon as

$\varepsilon < 1$ , for  $n$  large enough  $T_n$  is a diffeomorphism. Thus we can define on  $A_r(S_n)$  the projector  $p_{S_n}$  onto  $S_n$  along the field  $\nabla b_{V_\infty}$  by requiring that  $p_{S_n}$  coincides with the  $S_n$ -component of the inverse of  $T_n$  (see Figure II.3).

This allows us to introduce  $\tilde{j}_n$ , defined on  $A_h(S_n)$  by

$$\tilde{j}_n = j_n \circ p_{S_n}.$$

Using [DZ11, Chap. 7, theorem 8.5], for any  $m \in \mathbb{N} \cup \{\infty\}$  large enough, we get

$$p_{S_n} = p_{S_n} \circ p_{S_m} \tag{II.5}$$

on  $A_h(S_n) \cap A_h(S_m)$ . Moreover,  $p_{S_n}$  converges uniformly to  $p_{S_\infty}$  in a neighborhood of  $S_\infty$ , since for every  $x \in U_{r_{\min}}(S_\infty)$  one has

$$|p_{S_n}(x) - p_{S_\infty}(x)| = d_{S_\infty}(p_{S_n}(x)) \leq \|d_{S_\infty} - d_{S_n}\|_{L^\infty(\bar{D})} \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

where the limit is a consequence of Lemma II.7.

Using the change of variable formula (also known as area formula for Lipschitz functions), one gets for  $n \in \mathbb{N} \cup \{\infty\}$  large enough, every  $f \in L^1(S_n)$ , and every  $\kappa \in (0, h)$ ,

$$\int_{-\kappa}^{\kappa} \int_{S_n} f(x) d\mu_{S_n}(x) dt = \int_{A_\kappa(S_n)} f \circ p_{S_n}(y) |DT_n(T_n^{-1}(y))| dy,$$

which also rewrites

$$\int_{S_n} f(x) d\mu_{S_n}(x) = \frac{1}{2\kappa} \int_{A_\kappa(S_n)} f \circ p_{S_n}(y) |DT_n(T_n^{-1}(y))| dy. \tag{II.6}$$

Let  $\kappa$  be in  $(0, r)$  and  $\eta > 0$  be small enough so that  $\kappa - \eta > 0$  and  $\kappa + \eta < r$ , that is  $\eta < \min(\kappa, r - \kappa)$ . Since  $d_{S_n} \rightarrow d_{S_\infty}$  in  $\mathcal{C}(\bar{D})$ , there exists  $N$  such that for all  $n > N$ ,

$$A_{\kappa-\eta}(S_n) \subset A_\kappa(S_\infty) \subset A_{\kappa+\eta}(S_n). \tag{II.7}$$

Using that with Equation (II.5), we obtain

$$\begin{aligned} \int_{S_\infty} |\tilde{j}_n(x)|^2 d\mu_{S_\infty}(x) &= \frac{1}{2\kappa} \int_{A_\kappa(S_\infty)} |\tilde{j}_n(y)|^2 |DT_\infty(T_\infty^{-1}(y))| dy \\ &\leq \frac{1}{2\kappa} \int_{A_{\kappa+\eta}(S_n)} |\tilde{j}_n(y)|^2 |DT_\infty(T_\infty^{-1}(y))| dy \\ &= \frac{1}{2\kappa} \int_{A_{\kappa+\eta}(S_n)} |\tilde{j}_n(y)|^2 \frac{|DT_\infty(T_\infty^{-1}(y))|}{|DT_n(T_n^{-1}(y))|} |DT_n(T_n^{-1}(y))| dy \\ &\leq \frac{\kappa + \eta}{\kappa} \frac{1 + \varepsilon}{1 - \varepsilon} \|j_n\|_{\mathcal{F}_{S_n}}^2, \end{aligned} \tag{II.8}$$

which ensure that  $\tilde{j}_n$  belongs to  $L^2(S_\infty, \mathbb{R}^3)$ . (Notice however that  $\tilde{j}_n$  is not necessarily in  $\mathcal{F}_{S_\infty}^0$ , as it is neither, in general, a tangent vector field nor a divergence free one.) Equation (II.8) actually shows that  $(\tilde{j}_n)_{n \in \mathbb{N}}$  is bounded in  $L^2(S_\infty, \mathbb{R}^3)$ . Up to subsequence, it converges weakly to  $j_\infty \in L^2(S_\infty, \mathbb{R}^3)$  with

$$\|j_\infty\|_{L^2(S_\infty, \mathbb{R}^3)}^2 \leq \liminf_{n \rightarrow +\infty} \|j_n\|_{\mathcal{F}_{S_n}}^2.$$

The remaining two steps of the proof consist first in showing the semicontinuity property

$$\|BS_{S_\infty} j_\infty - B_T\|_{L^2(P, \mathbb{R}^3)}^2 \leq \liminf_{n \rightarrow +\infty} \|BS_{S_n} j_n - B_T\|_{L^2(P, \mathbb{R}^3)}^2 \quad (\text{II.9})$$

and then in checking that  $j_\infty$  belongs to  $\mathcal{F}_{S_\infty}^0$ . Notice that, even if we have defined the operator  $BS_{S_\infty}$  only among the vector fields tangent to  $S_\infty$ , by a slight abuse of notation it still makes sense to consider  $BS_{S_\infty} j_\infty$ , defined using formula (II.1), even without having checked that  $j_\infty$  is in  $\mathcal{F}_{S_\infty}$ .

Both steps rely on the following lemma.

**Lemma II.9.** *Given  $C > 0$  and  $\varepsilon' > 0$ , there exists  $N \in \mathbb{N}$  such that for every  $r > 0$  and every  $f : U_r(S_\infty) \rightarrow \mathbb{R}$  such that  $\|f\|_{L^\infty(U_r(S_\infty))} \leq C$  and  $f$  is  $C$ -Lipschitz continuous on  $U_r(S_\infty)$ , we have*

$$\left| \int_{S_n} f(x) j_n(x) d\mu_{S_n}(x) - \int_{S_\infty} f(x) \tilde{j}_n(x) d\mu_{S_\infty}(x) \right| < \varepsilon'$$

for  $n > N$ .

*Proof.* Let  $r > 0$  and  $f$  be such that  $\|f\|_{L^\infty(U_r(S_\infty))} \leq C$  and  $f$  is  $C$ -Lipschitz continuous on  $U_r(S_\infty)$ . Up to taking  $h$  small enough, we can assume that

$$A_h(S_n) \subset U_r(S_\infty) \quad \text{for } n \text{ large enough.} \quad (\text{II.10})$$

As above, consider  $\kappa \in (0, h)$ ,  $0 < \eta < \min(\kappa, h - \kappa)$ , and  $n$  large enough so that (II.7) holds true. By (II.6), we have

$$\left| \int_{S_n} f(x) j_n(x) d\mu_{S_n}(x) - \int_{S_\infty} f(x) \tilde{j}_n(x) d\mu_{S_\infty}(x) \right| = \left| \frac{1}{2(\kappa - \eta)} \int_{A_{\kappa - \eta}(S_n)} f(p_{S_n}(x)) \tilde{j}_n(x) |DT_n| dx - \frac{1}{2\kappa} \int_{A_\kappa(S_\infty)} f(p_{S_\infty}(x)) \tilde{j}_n(x) |DT_\infty| dx \right|,$$

where, for notational simplicity, we write  $|DT_n|$  for  $|DT_n(T_n^{-1}(x))|$  and  $|DT_\infty|$  for  $|DT_\infty(T_\infty^{-1}(x))|$ . Hence,

$$\left| \int_{S_n} f(x) j_n(x) d\mu_{S_n}(x) - \int_{S_\infty} f(x) \tilde{j}_n(x) d\mu_{S_\infty}(x) \right| \leq A_1 + A_2$$

where we added and subtracted  $\frac{1}{2\kappa} \int_{A_{\kappa - \eta}(S_n)} f(p_{S_\infty}(x)) \tilde{j}_n(x) |DT_\infty| dx$  to get

$$A_1 = \left| \int_{A_{\kappa - \eta}(S_n)} \left( \frac{1}{2(\kappa - \eta)} f(p_{S_n}(x)) \tilde{j}_n(x) |DT_n| - \frac{1}{2\kappa} f(p_{S_\infty}(x)) \tilde{j}_n(x) |DT_\infty| \right) dx \right|,$$

$$A_2 = \left| \frac{1}{2\kappa} \int_{A_\kappa(S_\infty) \setminus A_{\kappa - \eta}(S_n)} f(p_{S_\infty}(x)) \tilde{j}_n(x) |DT_\infty| dx \right|.$$

We are going to show that  $A_1$  and  $A_2$  can be made arbitrarily small by suitably choosing  $\kappa$  and  $\eta$  (depending only on  $C$  and not on the specific function  $f$ ) and letting  $n$  be large enough.



The term  $A_2$  can be estimated using the inequality  $\|f\|_{L^\infty(U_r(S_\infty))} \leq C$ , as follows:

$$\begin{aligned} A_2 &\leq C \frac{1+\varepsilon}{1-\varepsilon} \int_{A_\kappa(S_\infty) \setminus A_{\kappa-\eta}(S_n)} |\tilde{j}_n(x)| |DT_n| dx \leq C \frac{1+\varepsilon}{1-\varepsilon} \int_{A_{\kappa+\eta}(S_n) \setminus A_{\kappa-\eta}(S_n)} |\tilde{j}_n(x)| |DT_n| dx \\ &= 4\eta C \frac{1+\varepsilon}{1-\varepsilon} \|j_n\|_{L^1(S^n)}, \end{aligned}$$

where the factor 4 comes from the fact that the measure of  $(-\kappa - \eta, -\kappa + \eta) \cup (\kappa - \eta, \kappa + \eta)$  is equal to  $4\eta$ . Notice that  $\|j_n\|_{L^1(S^n)} \leq \|j_n\|_{\mathcal{F}_{S_n}} \sqrt{\mathcal{H}^2(S_n)}$  is bounded uniformly with respect to  $n$ , so that  $A_2$  can be made arbitrarily small by choosing  $\eta$  small enough (depending only on  $C$ ).

Let us now focus on the term  $A_1$ . Since

$$|f(x_1) - f(x_2)| \leq C|x_1 - x_2|, \quad \forall x_1, x_2 \in U_r(S_\infty),$$

and because of (II.10), it follows that

$$\sup_{x \in A_{\kappa-\eta}(S_n)} |f(p_{S_n}(x)) - f(p_{S_\infty}(x))| \leq C \|p_{S_n} - p_{S_\infty}\|_{L^\infty(A_{\kappa-\eta}(S_n))} \leq C \|p_{S_n} - p_{S_\infty}\|_{L^\infty(A_h(S_\infty))}$$

for  $n$  large enough.

Hence, we have the estimates

$$\begin{aligned} A_1 &\leq \frac{1}{2(\kappa - \eta)} \left| \int_{A_{\kappa-\eta}(S_n)} (f(p_{S_n}(x))\tilde{j}_n(x)|DT_n| - f(p_{S_\infty}(x))\tilde{j}_n(x)|DT_n|) dx \right| \\ &\quad + \left| \int_{A_{\kappa-\eta}(S_n)} \left( \frac{1}{2(\kappa - \eta)} f(p_{S_\infty}(x))\tilde{j}_n(x)|DT_n| - \frac{1}{2\kappa} f(p_{S_\infty}(x))\tilde{j}_n(x)|DT_n| \right) dx \right| \\ &\quad + \frac{1}{2\kappa} \left| \int_{A_{\kappa-\eta}(S_n)} (f(p_{S_\infty}(x))\tilde{j}_n(x)|DT_n| - f(p_{S_\infty}(x))\tilde{j}_n(x)|DT_\infty|) dx \right| \\ &\leq C \|p_{S_n} - p_{S_\infty}\|_{L^\infty(A_h(S_\infty))} \|j_n(x)\|_{L^1(S_n)} + C \|j_n(x)\|_{L^1(S_n)} \left( 1 - \frac{\kappa - \eta}{\kappa} \right) \\ &\quad + C \frac{\kappa - \eta}{\kappa} \|j_n(x)\|_{L^1(S_n)} \left( 1 - \frac{1 + \varepsilon}{1 - \varepsilon} \right). \end{aligned}$$

Hence  $A_1$  can be made arbitrarily small choosing  $\varepsilon$  and then  $\eta$  small enough, and letting  $n$  large enough.  $\square$

Let us start the proof of (II.9) by comparing  $\text{BS}_{S_n}$  and  $\text{BS}_{S_\infty}$ . Given  $y \in P$ , one has

$$|\text{BS}_{S_n}(j_n)(y) - \text{BS}_{S_\infty}(\tilde{j}_n)(y)| = \left| \int_{S_n} K(x, y) \times j_n(x) dx - \int_{S_\infty} K(x, y) \times \tilde{j}_n(x) dx \right|.$$

Notice that  $|K(\cdot, y)|$  is bounded in a neighborhood of  $S_\infty$ , uniformly with respect to  $y \in P$ , since  $\sup_{(x,y) \in S_n \times P} |K(x, y)| \leq \frac{1}{\delta^2}$ . Moreover, for every  $y \in P$  and every  $\rho > 0$ , the map  $x \mapsto \|D_x K(x, y)\|$  is upper bounded by  $\frac{4}{\rho^3}$  outside  $U_\rho(P)$  according to (II.2). Assume that  $h < \delta$ , so that  $A_h(S_n)$  is at distance at least  $\delta - h$  from  $P$  for every  $n$ . Consider  $\rho < \delta - h$  and a Lipschitz neighborhood  $\mathcal{N}$  of  $\mathbb{R}^3 \setminus U_{\delta-h}(P)$  not intersecting  $U_\rho(P)$ . Since the geodesic distance in  $\mathcal{N}$  is equivalent to the restriction to  $\mathcal{N}$  of the standard Euclidean distance, we deduce that

there exists  $\tilde{C} > 0$  independent of  $y$  such that

$$|K(x_1, y) - K(x_2, y)| \leq \tilde{C}|x_1 - x_2|, \quad \forall x_1, x_2 \in \mathbb{R}^3 \setminus U_{\delta-h}(P).$$

We deduce from Lemma II.9 that for every  $\varepsilon' > 0$  there exists  $N > 0$  such that for any integer  $n > N$ ,

$$\|\text{BS}_{S_n}(j_n) - \text{BS}_{S_\infty}(\tilde{j}_n)\|_{L^2(P, \mathbb{R}^3)} \leq \varepsilon',$$

and, in particular,

$$\|\text{BS}_{S_\infty} \tilde{j}_n - B_T\|_{L^2(P, \mathbb{R}^3)} \leq \|\text{BS}_{S_n} j_n - B_T\|_{L^2(P, \mathbb{R}^3)} + \varepsilon'.$$

Using the compactness of  $\text{BS}_{S_\infty}$ , we have

$$\|\text{BS}_{S_\infty} j_\infty - B_T\|_{L^2(P, \mathbb{R}^3)} \leq \liminf_{n \rightarrow +\infty} \|\text{BS}_{S_n} j_n - B_T\|_{L^2(P, \mathbb{R}^3)} + \varepsilon'.$$

This concludes the proof of (II.9), since  $\varepsilon'$  is arbitrary.

To conclude the proof, it remains to check that  $j_\infty$  belongs to  $\mathcal{F}_{S_\infty}^0$ . By weak convergence of  $\tilde{j}_n$  to  $j_\infty$  and according to Lemma II.9,

$$\begin{aligned} \|\langle j_\infty, \nabla b_{V_\infty} \rangle\|_{L^2(S_\infty, \mathbb{R}^3)} &= \lim_{n \rightarrow \infty} \|\langle \tilde{j}_n, \nabla b_{V_\infty} \rangle\|_{L^2(S_\infty, \mathbb{R}^3)} = \lim_{n \rightarrow \infty} \|\langle j_n, \nabla b_{V_\infty} \rangle\|_{L^2(S_n, \mathbb{R}^3)} \\ &\leq \limsup_{n \rightarrow \infty} \|\langle j_n, \nabla b_{V_\infty} - \nabla b_{V_n} \rangle\|_{L^2(S_n, \mathbb{R}^3)}, \end{aligned}$$

where we used that  $j_n$  is orthogonal to  $\nabla b_{V_n}$  everywhere on  $S_n$ . According to Lemma II.6, moreover,

$$\lim_{n \rightarrow \infty} \|\nabla b_{V_\infty} - \nabla b_{V_n}\|_{L^\infty(S_\infty, \mathbb{R}^3)} = 0,$$

and we conclude that  $\|\langle j_\infty, \nabla b_{V_\infty} \rangle\|_{L^2(S_\infty, \mathbb{R}^3)} = 0$  since the sequence  $(\|j_n\|_{L^2(S_\infty, \mathbb{R}^3)})_{n \in \mathbb{N}}$  is bounded. This proves that  $j_\infty$  is a vector field tangent to  $S_\infty$ .

To prove that  $j_\infty$  is divergence free (in distributional sense), we have to check that  $j_\infty$  is orthogonal to  $\{\nabla_{S_\infty} f \mid f \in \mathcal{C}^1(S_\infty)\}$ . Indeed, this characterization of divergence-free vector fields follows from the Hodge decomposition (see Appendix II.A.1.1). For  $g \in \mathcal{C}^\infty(\mathbb{R}^3)$ , since  $\text{div } j_n = 0$  on  $S_n$ , one has

$$0 = \int_{S_n} \langle j_n, \nabla_{S_n} g \rangle d\mu_{S_n} = \int_{S_n} \langle j_n, (\nabla g - \langle \nabla g, \nabla b_{V_n} \rangle \nabla b_{V_n}) \rangle d\mu_{S_n}.$$

Set  $G_n := \nabla g - \langle \nabla g, \nabla b_{V_n} \rangle \nabla b_{V_n}$  for  $n \in \mathbb{N} \cup \{\infty\}$ . Notice that  $G_n$  converges uniformly to  $G_\infty$  in a neighborhood of  $S_\infty$ . Hence, again using Lemma II.9,

$$\begin{aligned} \int_{S_\infty} \langle j_\infty, G_\infty \rangle d\mu_{S_\infty}(x) &= \lim_{n \rightarrow \infty} \int_{S_\infty} \langle \tilde{j}_n, G_\infty \rangle d\mu_{S_\infty}(x) = \lim_{n \rightarrow \infty} \int_{S_n} \langle \tilde{j}_n, G_\infty \rangle d\mu_{S_n}(x) \\ &= \lim_{n \rightarrow \infty} \int_{S_n} \langle \tilde{j}_n, G_n \rangle d\mu_{S_n}(x) = 0. \end{aligned}$$

This concludes the proof of Theorem II.5.

### II.3 Shape differentiation for Problem $(\mathcal{P}_{\text{shape}})$

In the analysis of shape optimization problems, it is often convenient to consider particular perturbations called *identity perturbations*. The latter are of the form  $\tau = \text{Id} + \theta$ , where  $\theta$  is small enough in a suitable sense. More precisely, according to the approach developed in [MS76a; MS76b], if  $\Omega_0$  denotes an open bounded subset of  $\mathbb{R}^3$  such that  $\partial\Omega_0$  is of class  $\mathcal{C}^{1,1}$  and if  $\|\theta\|_{W^{2,\infty}} < 1$ , then  $\tau$  is a  $W^{2,\infty}$ -diffeomorphism and  $\tau(\Omega_0)$  is an open bounded domain whose boundary is of class  $\mathcal{C}^{1,1}$ .

Let us now recall the notion of shape differentiability.

**Definition II.10.** A shape functional  $\Omega \mapsto J(\Omega)$  is said to be shape differentiable at  $\Omega$  (in the sense of Hadamard) in the class of domains with  $\mathcal{C}^{1,1}$  boundary whenever the underlying mapping

$$W^{2,\infty}(\mathbb{R}^3, \mathbb{R}^3) \ni \theta \mapsto J(\Omega_\theta) \in \mathbb{R},$$

with  $\Omega_\theta = (\text{Id} + \theta)(\Omega)$ , is differentiable in the sense of Fréchet at  $\theta = 0$ . The corresponding differential  $\langle dJ(\Omega), \cdot \rangle$  is the so-called shape derivative of  $J$  at  $\Omega$  and, by definition of Fréchet differential, the following expansion holds:

$$J(\Omega_\theta) = J(\Omega) + \langle dJ(\Omega), \theta \rangle + o(\theta), \quad \text{where } \frac{o(\theta)}{\|\theta\|_{W^{2,\infty}(\mathbb{R}^3, \mathbb{R}^3)}} \xrightarrow{\theta \rightarrow 0} 0.$$

In the next section we study the shape differentiability of the cost  $C$ . In order to fit Definition 2,  $C$  is implicitly identified with a functional  $V \mapsto C(\partial V) = C(S)$  on the set of  $\mathcal{C}^{1,1}$  toroidal domains.

#### II.3.1 Shape derivative of the cost functional $C$

This section and the next one are devoted to the computation of the shape derivative of the functional  $C$ .

**Theorem II.11.** Let  $S = \partial V \in \mathcal{O}_{ad}$ . Let  $Z_P \in \mathcal{L}(L^2(P, \mathbb{R}^3), \mathcal{F}_S)$  and  $\widehat{Z}_P$ , a bilinear mapping from  $L^2(P, \mathbb{R}^3) \times \mathcal{F}_S^0$  into  $\mathcal{F}_S$ , defined by

$$\begin{aligned} Z_P(k) &= \int_P K(\cdot, y) \times k(y) d\mu_P(y), \\ \widehat{Z}_P(k, j)(x) &= \int_P D_x \left( \frac{x-y}{|x-y|^3} \right)^T (k(y) \times j(x)) d\mu_P(y), \quad \forall x \in S. \end{aligned}$$

The functional  $C$  defined by  $(\mathcal{P}_S)$  is shape differentiable at  $S$ . Moreover, for every  $\theta \in W^{2,\infty}(\mathbb{R}^3, \mathbb{R}^3)$  one has

$$\langle dC(S), \theta \rangle = \int_S \langle \theta, (X_1 - \overrightarrow{\text{div}}_S(X_2)) \rangle d\mu_S$$

with

$$\begin{aligned} X_1 &= -2\widehat{Z}_P(\text{BS}_S j_S - B_T, j_S), \\ X_2 &= -2Z_P(\text{BS}_S j_S - B_T)j_S^T + 2\lambda j_S j_S^T - \lambda |j_S|^2 (I_3 - \nu\nu^T), \end{aligned}$$

where  $\overrightarrow{\text{div}}_S(X_2)$  is the vector field with  $i$ -th component  $\overrightarrow{\text{div}}_S(X_2)_i = \text{div}_S((X_2)_i)$  for  $i \in \{1, 2, 3\}$ ,

$(X_2)_i$ : denotes the  $i$ -th line of  $X_2$  seen as a column vector, and  $\nu$  denotes the outward normal vector to  $S = \partial V$ .

**Remark II.12.** *The proof of this result relies crucially on the expression of the magnetic field provided through the Biot and Savart operator  $\text{BS}_S$  (see Definition II.1). In general, in many shape optimization problems involving PDEs on bounded domains, PDEs are interpreted as implicit equations on the deformation variable  $\theta$  and on the state variable. They are in general taken into account by applying the implicit function theorem which also provides an expression for the material (or Lagrangian) derivative of the state with respect to the deformation (see, e.g., [HP18, Chapter 5]). In the present case, dealing with the Biot and Savart operator comes to consider a PDE on an unbounded domain. The approach we have chosen here, instead, is based on the integral representation of the state variable (the magnetic field here). To establish the above result, we use suitable changes of variables that allows us to rewrite the criterion as an integral over a fixed domain and derive it as a parameterized integral with respect to  $\theta$ . Although the principle of this calculation is simple, its implementation is not straightforward.*

### II.3.2 Proof of Theorem II.11

For the sake of notational simplicity, the inverse of a group element  $\vartheta^\varepsilon$  will be denoted with a slight abuse of notation by  $\vartheta^{-\varepsilon} := (\vartheta^\varepsilon)^{-1}$ .

Let  $S$  and  $\theta$  be as in the statement of the theorem. Assume for now that the criterion  $C$  is shape differentiable at  $S$ . We will comment on this assumption at the end of the proof. In what follows, we concentrate on the computation of the shape derivative in the direction  $\theta$ .

Since  $C$  is shape differentiable at  $S$ , we infer that

$$\langle dC(S), \theta \rangle = \left. \frac{d}{d\varepsilon} C(S^\varepsilon) \right|_{\varepsilon=0}, \quad \text{with } S^\varepsilon = (\text{Id} + \varepsilon\theta)S.$$

**Step 1: a change of variable.** In order to compute  $C(S^\varepsilon)$ , we need to compute some kind of derivative of  $\text{BS}_{S^\varepsilon}$  and its adjoint. Nevertheless, we aim to overcome the fact that the domain of  $\text{BS}_{S^\varepsilon}$  depends on  $\varepsilon$ .

Let us introduce the diffeomorphism from  $S$  to  $S^\varepsilon$  defined by

$$\begin{aligned} \vartheta^\varepsilon : S &\longrightarrow S^\varepsilon \\ x &\longmapsto (\text{Id} + \varepsilon\theta)(x). \end{aligned}$$

Notice that, according to the discussion at the beginning of Section II.3, the mapping  $\vartheta^\varepsilon$  induces a bijection between  $\mathfrak{X}(S)$  and  $\mathfrak{X}(S^\varepsilon)$ . Nevertheless this bijection does not map  $\mathcal{F}_S^0$  into  $\mathcal{F}_{S^\varepsilon}^0$ . This leads us to introduce the linear mapping

$$\begin{aligned} \Phi^\varepsilon : \mathcal{F}_S &\longrightarrow \mathcal{F}_{S^\varepsilon} \\ X &\longmapsto \frac{1}{[J(\mu_S, \mu_S^\varepsilon)\vartheta^\varepsilon] \circ \vartheta^{-\varepsilon}} (\text{Id} + \varepsilon D\theta)X \circ \vartheta^{-\varepsilon}, \end{aligned} \tag{II.11}$$

where  $J(\mu_S, \mu_S^\varepsilon)\vartheta^\varepsilon$  denotes the Jacobian determinant<sup>5</sup> of  $\vartheta^\varepsilon$  (see Appendix II.A.3 for further details and the explicit expression of  $J(\mu_S, \mu_S^\varepsilon)\vartheta^\varepsilon$ ).

The following result will be crucial in what follows since it confirms that  $\Phi^\varepsilon$  is indeed a diffeomorphism preserving divergence-free vector fields.

5. Note that  $J(\mu_S, \mu_S^\varepsilon)\vartheta^\varepsilon$  is not the determinant of the three-dimensional mapping  $(\text{Id} + \varepsilon D\theta)$  but the determinant of the restriction of this application from  $T_x S$  (the tangent space of  $S$  at  $x$ ) into  $T_{(\text{Id} + \varepsilon\theta)(x)} S^\varepsilon$ .

**Lemma II.13.** *For every  $\varepsilon$  small enough,  $\Phi^\varepsilon$  is a diffeomorphism from  $\mathcal{F}_S^0$  to  $\mathcal{F}_{S^\varepsilon}^0$ .*

*Proof.* Since  $\vartheta^\varepsilon$  is an orientation preserving diffeomorphism, one has  $J(\mu_S, \mu_{S^\varepsilon}^\varepsilon)\vartheta^\varepsilon > 0$ . Besides,

$$\Phi^{-\varepsilon}(X) = \frac{1}{[J(\mu_S^\varepsilon, \mu_S)\vartheta^{-\varepsilon}] \circ \vartheta^\varepsilon} D[(\text{Id} + \varepsilon\theta)^{-1}]X \circ \vartheta^\varepsilon, \quad X \in \mathfrak{X}(S^\varepsilon).$$

As a consequence,  $\Phi^\varepsilon$  defines a diffeomorphism from  $\mathcal{F}_S$  to  $\mathcal{F}_{S^\varepsilon}$ . We are left to prove that it preserves divergence-free vector fields. According to the Hodge decomposition (see Appendix II.A.1.1), it is enough to check that  $\Phi^\varepsilon(\mathcal{F}_S)$  is orthogonal to  $\{\nabla_{S^\varepsilon} f \mid f \in \mathcal{C}^\infty(S^\varepsilon)\}$ . Using the change of variables formula (cf. (II.28)), one has, for every  $X \in \mathfrak{X}(S)$ ,

$$\begin{aligned} \int_{S^\varepsilon} \langle df, \Phi^\varepsilon(X) \rangle d\mu_{S^\varepsilon} &= \int_{S^\varepsilon} \langle df, \vartheta_*^\varepsilon(X) \rangle \frac{1}{[J(\mu_S, \mu_{S^\varepsilon}^\varepsilon)\vartheta^\varepsilon] \circ \vartheta^{-\varepsilon}} d\mu_{S^\varepsilon} = \int_S \langle \vartheta^{\varepsilon,*} df, X \rangle d\mu_S \\ &= \int_S \langle d(f \circ \vartheta^\varepsilon), X \rangle d\mu_S, \end{aligned}$$

where the notation  $\vartheta^{\varepsilon,*}$  stands for the conormal derivative of  $\vartheta^\varepsilon$ . Then  $X$  is divergence-free if and only if  $\Phi^\varepsilon(X)$  is. The lemma is thus proved.  $\square$

**Step 2: computation of the variation of  $j$ .** Since we prefer to avoid dealing with operators defined on  $S^\varepsilon$ , we will use  $\Phi^\varepsilon$  to relate  $\mathcal{F}_S^0$  and  $\mathcal{F}_{S^\varepsilon}^0$ .

Let us first compute the adjoint  $(\Phi^\varepsilon)^\dagger$ . Let  $j \in \mathcal{F}_S$  and  $g \in \mathcal{F}_{S^\varepsilon}$ . One has

$$\begin{aligned} \langle \Phi^\varepsilon j, g \rangle &= \int_{S^\varepsilon} \frac{1}{[J(\mu_S, \mu_{S^\varepsilon}^\varepsilon)\vartheta^\varepsilon] \circ \vartheta^{-\varepsilon}} \langle (\text{Id} + \varepsilon D\theta)j(\vartheta^{-\varepsilon}(x)), g(x) \rangle d\mu_{S^\varepsilon}(x) \\ &= \int_S \langle (\text{Id} + \varepsilon D\theta)j(x), g(\vartheta^\varepsilon(x)) \rangle d\mu_S(x) \\ &= \int_S \langle j(x), (\text{Id} + \varepsilon D\theta)^T g(\vartheta^\varepsilon(x)) \rangle d\mu_S(x). \end{aligned}$$

We thus infer that  $(\Phi^\varepsilon)^\dagger$  is given by

$$\begin{aligned} (\Phi^\varepsilon)^\dagger : \mathcal{F}_{S^\varepsilon} &\longrightarrow \mathcal{F}_S \\ g &\longmapsto (\text{Id} + \varepsilon D\theta^T)g \circ \vartheta^\varepsilon. \end{aligned}$$

Let  $j^\varepsilon := \Phi^{-\varepsilon}(j_{S^\varepsilon})$ . According to Lemmas II.3 and II.13,  $j^\varepsilon$  is well defined and belongs to  $\mathcal{F}_S^0$ . To compute the differential of  $j^\varepsilon$ , it is convenient to introduce the operators

$$\begin{aligned} Q^\varepsilon : \mathcal{F}_S^0 &\longrightarrow \mathcal{F}_S^0 & \text{and} & \quad L_\varepsilon : \mathcal{F}_S^0 &\longrightarrow L^2(P, \mathbb{R}^3) \\ j &\longmapsto (\Phi^\varepsilon)^\dagger \Phi^\varepsilon j & & \quad j &\longmapsto \text{BS}_{S^\varepsilon} \Phi^\varepsilon j \end{aligned}$$

so that

$$\forall j, k \in \mathcal{F}_S, \quad \|\Phi^\varepsilon(j)\|_{\mathcal{F}_{S^\varepsilon}}^2 = \langle j, Q^\varepsilon j \rangle_{\mathcal{F}_S} \quad \text{and} \quad \langle Q^\varepsilon j, k \rangle_{\mathcal{F}_S} = \langle j, Q^\varepsilon k \rangle_{\mathcal{F}_S}.$$

According to the optimality condition (II.4) on  $j_{S^\varepsilon}$ ,  $j^\varepsilon$  is uniquely characterized by the identity

$$\forall v \in \mathcal{F}_S^0, \quad 0 = \langle L_\varepsilon v, L_\varepsilon j^\varepsilon - B_T \rangle_{L^2(P, \mathbb{R}^3)} + \lambda \langle v, Q^\varepsilon j^\varepsilon \rangle_{\mathcal{F}_S^0}$$

which also rewrites

$$\forall v \in \mathcal{F}_S^0, \quad 0 = \langle v, \lambda Q^\varepsilon j^\varepsilon + L_\varepsilon^\dagger(L_\varepsilon j^\varepsilon - B_T) \rangle_{\mathcal{F}_S^0}.$$

It follows that

$$j^\varepsilon = (\lambda Q^\varepsilon + L_\varepsilon^\dagger L_\varepsilon)^{-1} L_\varepsilon^\dagger B_T. \quad (\text{II.12})$$

Let us now compute the first order variation of  $j^\varepsilon$ . To this aim, we use the expansion

$$J(\mu_S, \mu_S^\varepsilon) \vartheta^\varepsilon = 1 + \varepsilon \operatorname{div}_S \theta + o(\varepsilon) \quad (\text{II.13})$$

obtained in [HP18, Lemma 5.4.15]. Recall that the notation  $\operatorname{div}_S \theta$  stands for the tangential divergence of  $\theta$  on  $S$ .

**Lemma II.14.** *Let  $S$  and  $\theta$  be chosen as above. Then, one has*

$$\begin{aligned} L_\varepsilon &= \text{BS}_S + \left. \frac{dL_\varepsilon}{d\varepsilon} \right|_{\varepsilon=0} \varepsilon + o(\varepsilon) \text{ in } \mathcal{L}(\mathcal{F}_S^0, L^2(P, \mathbb{R}^3)), \\ Q^\varepsilon &= I + \left. \frac{dQ^\varepsilon}{d\varepsilon} \right|_{\varepsilon=0} \varepsilon + o(\varepsilon) \text{ in } \mathcal{L}(\mathcal{F}_S^0), \end{aligned}$$

where, for every  $j \in \mathcal{F}_S^0$  and  $y \in P$ ,

$$\begin{aligned} \left( \left. \frac{dL_\varepsilon}{d\varepsilon} \right|_{\varepsilon=0} j \right) (y) &= \int_S (K(x, y) \times (D\theta(x)j(x)) + (D_x K(x, y)\theta(x)) \times j(x)) d\mu_S(x), \\ \left. \frac{dQ^\varepsilon}{d\varepsilon} \right|_{\varepsilon=0} &= D\theta + D\theta^T - \operatorname{div}_S \theta \operatorname{Id} = e(\theta) - \operatorname{div}_S \theta \operatorname{Id}, \end{aligned} \quad (\text{II.14})$$

and  $e(\theta)$  is twice the symmetric part of the Jacobian matrix  $D\theta$ , that is,

$$e(\theta) = D\theta + (D\theta)^T.$$

*Proof of Lemma II.14.* Let us start with  $L_\varepsilon$ . Given  $j \in \mathcal{F}_S^0$  and  $y \in P$ , we have

$$\begin{aligned} L_\varepsilon(j)(y) &= \int_{S^\varepsilon} \frac{1}{[J(\mu_S, \mu_S^\varepsilon)\vartheta^\varepsilon] \circ \vartheta^{-\varepsilon}} K(x, y) \times [(\operatorname{Id} + \varepsilon D\theta)j(\vartheta^{-\varepsilon}(x))] d\mu_{S^\varepsilon}(x) \\ &= \int_S K(\vartheta^\varepsilon(x), y) \times [(\operatorname{Id} + \varepsilon D\theta)j(x)] d\mu_S(x) \\ &= \text{BS}_S(j)(y) + \varepsilon \int_S (K(x, y) \times (D\theta(x)j(x)) + [D_x K(x, y)\theta(x)] \times j(x)) d\mu_S(x) + o(\varepsilon). \end{aligned} \quad (\text{II.15})$$

Moreover, it can be easily checked that the reminder term of this expansion grows at most linearly with respect to  $\|j\|_{\mathcal{F}_S}$ . Regarding  $Q^\varepsilon$ , a similar reasoning using (II.13) yields

$$\begin{aligned} Q^\varepsilon &= \frac{1}{[J(\mu_S, \mu_S^\varepsilon)\vartheta^\varepsilon] \circ \vartheta^{-\varepsilon}} (\operatorname{Id} + \varepsilon D\theta^T)(\operatorname{Id} + \varepsilon D\theta) \\ &= \operatorname{Id} + \varepsilon (D\theta + D\theta^T - \operatorname{div}_S \theta \operatorname{Id}) + o(\varepsilon), \end{aligned}$$

concluding the proof.  $\square$

Combining all the results above, we now compute the sensitivity of  $j^\varepsilon$  with respect to  $\varepsilon$ . The following result is an immediate consequence of Lemma II.14 and (II.12).

**Proposition II.15.** *One has  $j^\varepsilon = j_S + \frac{dj^\varepsilon}{d\varepsilon}\Big|_{\varepsilon=0} \varepsilon + o(\varepsilon)$  with*

$$\begin{aligned} \frac{dj^\varepsilon}{d\varepsilon}\Big|_{\varepsilon=0} &= \left( \lambda \text{Id} + \text{BS}_S^\dagger \text{BS}_S \right)^{-1} \frac{dL_\varepsilon^\dagger}{d\varepsilon}\Big|_{\varepsilon=0} B_T \\ &\quad - \left( \lambda \text{Id} + \text{BS}_S^\dagger \text{BS}_S \right)^{-1} \left( \lambda \frac{dQ^\varepsilon}{d\varepsilon}\Big|_{\varepsilon=0} + \frac{dL_\varepsilon^\dagger}{d\varepsilon}\Big|_{\varepsilon=0} \text{BS}_S + \text{BS}_S^\dagger \frac{dL_\varepsilon}{d\varepsilon}\Big|_{\varepsilon=0} \right) \left( \lambda \text{Id} + \text{BS}_S^\dagger \text{BS}_S \right)^{-1} \text{BS}_S^\dagger B_T. \end{aligned}$$

**Step 3: computation of the cost functional derivative.** Recall that

$$\begin{aligned} C(S^\varepsilon) &= \| \text{BS}_{S^\varepsilon} j_{S^\varepsilon} - B_T \|_{L^2(P, \mathbb{R}^3)}^2 + \lambda \| j_{S^\varepsilon} \|_{\mathcal{F}_{S^\varepsilon}}^2 \\ &= \| L_\varepsilon j^\varepsilon - B_T \|_{L^2(P, \mathbb{R}^3)}^2 + \lambda \langle j^\varepsilon, Q^\varepsilon j^\varepsilon \rangle_{\mathcal{F}_S}. \end{aligned} \quad (\text{II.16})$$

By differentiating this expression and according to Proposition II.15, we get

$$\begin{aligned} \frac{dC(S^\varepsilon)}{d\varepsilon}\Big|_{\varepsilon=0} &= \lambda \left( \left\langle j_S, \frac{dQ^\varepsilon}{d\varepsilon}\Big|_{\varepsilon=0} j_S \right\rangle_{\mathcal{F}_S} + 2 \left\langle j_S, \frac{dj^\varepsilon}{d\varepsilon}\Big|_{\varepsilon=0} \right\rangle_{\mathcal{F}_S} \right) \\ &\quad + 2 \left\langle \text{BS}_S j_S - B_T, \frac{dL_\varepsilon}{d\varepsilon}\Big|_{\varepsilon=0} j_S + \text{BS}_S \frac{dj^\varepsilon}{d\varepsilon}\Big|_{\varepsilon=0} \right\rangle_{L^2(P, \mathbb{R}^3)}. \end{aligned}$$

Note that

$$2 \left\langle (\lambda \text{Id} + \text{BS}_S^\dagger \text{BS}_S) j_S - \text{BS}_S^\dagger B_T, \frac{dj^\varepsilon}{d\varepsilon}\Big|_{\varepsilon=0} \right\rangle_{\mathcal{F}_S} = 0.$$

Thus

$$\frac{dC(S^\varepsilon)}{d\varepsilon}\Big|_{\varepsilon=0} = \lambda \left\langle j_S, \frac{dQ^\varepsilon}{d\varepsilon}\Big|_{\varepsilon=0} j_S \right\rangle_{\mathcal{F}_S} + 2 \left\langle \text{BS}_S j_S - B_T, \frac{dL_\varepsilon}{d\varepsilon}\Big|_{\varepsilon=0} j_S \right\rangle_{L^2(P, \mathbb{R}^3)}. \quad (\text{II.17})$$

**Remark II.16.** *The previous expression can be understood as follows: writing  $C(S) =: \tilde{C}(S, j_S)$  with the natural choice of  $\tilde{C}$  and assuming that  $(C, j) \mapsto \tilde{C}$  and  $S \mapsto j_S$  are sufficiently regular, one has*

$$\frac{\partial \tilde{C}(S, j_S)}{\partial S} = \frac{\partial \tilde{C}}{\partial S}(S, j_S) + \frac{\partial \tilde{C}}{\partial j} \frac{\partial j_S}{\partial S}(S, j_S).$$

Using the fact that  $\frac{\partial \tilde{C}}{\partial j}(j_S) = 0$  since  $j_S$  is the minimizer of  $j \mapsto \tilde{C}(S, j)$ , we get

$$\frac{\partial \tilde{C}(S, j_S)}{\partial S} = \frac{\partial \tilde{C}}{\partial S}(S, j_S).$$

In what follows, we will use the identity stated in the following lemma.

**Lemma II.17.** *Let  $j \in \mathcal{F}_S^0$ ,  $k \in L^2(P, \mathbb{R}^3)$ , and  $\theta$  be as in the statement of Theorem II.11. Then*

$$\left\langle \frac{dL_\varepsilon}{d\varepsilon}\Big|_{\varepsilon=0} j, k \right\rangle_{L^2(P, \mathbb{R}^3)} = - \langle D\theta j, Z_P(k) \rangle_{\mathcal{F}_S} - \left\langle \theta, \widehat{Z}_P(k, j) \right\rangle_{\mathcal{F}_S}.$$

*Proof.* The proof follows from straightforward computations, by combining the Fubini theorem

with standard properties of the scalar triple product<sup>6</sup>. □

By combining (II.14), (II.17), and Lemma II.17, one computes

$$\begin{aligned} \left. \frac{dC(S^\varepsilon)}{d\varepsilon} \right|_{\varepsilon=0} = & \lambda \langle j_S, e(\theta)j_S - \operatorname{div}_S \theta j_S \rangle_{\mathcal{F}_S} - 2 \langle D\theta j_S, Z_P(\operatorname{BS}_S j_S - B_T) \rangle_{\mathcal{F}_S} \\ & - 2 \left\langle \theta, \widehat{Z}_P(\operatorname{BS}_S j_S - B_T, j_S) \right\rangle_{\mathcal{F}_S}. \end{aligned}$$

To conclude this computation, observe that for all vectors  $u$  and  $v$  in  $\mathbb{R}^3$ ,

$$\begin{aligned} \langle D\theta u, v \rangle &= \sum_{i,j=1}^3 (D\theta)_{ij} u_j v_i = D\theta : (uv^T), & \langle (D\theta)^T u, v \rangle &= D\theta : (vu^T), \\ \langle e(\theta)u, v \rangle &= D\theta : (uv^T + vu^T) \end{aligned}$$

so that

$$\operatorname{div}_S \theta = \sum_{i=1}^3 \partial_{x_i} \theta_i - D\theta : (\nu\nu^T) = D\theta : (I_3 - \nu\nu^T).$$

We thus obtain

$$\left. \frac{dC(S^\varepsilon)}{d\varepsilon} \right|_{\varepsilon=0} = \int_S (\langle \theta, X_1 \rangle + D\theta : X_2) d\mu_S,$$

where  $X_1$  and  $X_2$  have been introduced in the statement of the theorem.

Note that each line of  $X_2$  is tangential to  $S$  (in other words, normal to  $\nu$ ). This can be written as  $\langle (X_2)_{i\cdot}, \nu \rangle = 0$  where, for  $i \in \{1, 2, 3\}$ ,  $(X_2)_{i\cdot}$  denotes the  $i$ -th line of  $X_2$  seen as a column vector. Indeed, this follows from the definitions of the mapping  $Z_P$ , the function  $j_S$ , and the fact that  $I_3 - \nu\nu^T$  corresponds to the matrix of the orthogonal projection onto  $S$ .

Now, according to [HP18, Prop. 5.4.9], the above cost functional derivative can be recast as

$$\begin{aligned} \left. \frac{dC(S^\varepsilon)}{d\varepsilon} \right|_{\varepsilon=0} &= \int_S \langle \theta, X_1 \rangle + \sum_{i=1}^3 \int_S \langle \nabla_S \theta_i, (X_2)_{i\cdot} \rangle d\mu_S \\ &= \int_S \langle \theta, X_1 \rangle - \sum_{i=1}^3 \int_S \theta_i \operatorname{div}_S (X_2)_{i\cdot} d\mu_S. \end{aligned}$$

To conclude this proof, it remains to investigate the shape differentiability of  $S \mapsto C(S)$ . Let us introduce  $\tau_\theta = \operatorname{Id} + \theta$  where  $\theta$  is chosen as in the statement of the theorem. It is straightforward to show that the real number  $C(\tau_\theta(S))$  can be written as a smooth function of integrals written on the fixed domain  $S$ , for which the integrand depends regularly on  $\theta$ . Indeed, this can be straightforwardly obtained by replacing  $\operatorname{Id} + \varepsilon\theta$  by  $\operatorname{Id} + \theta$  in the reasoning above, and mimicking the associated computations leading to (II.12), (II.15) and (II.16). This yields to the expansion

$$C(\tau_\theta(S)) = C(S) + \left. \frac{dC(S^\varepsilon)}{d\varepsilon} \right|_{\varepsilon=0} + o(\|\theta\|_{W^{2,\infty}(\mathbb{R}^d, \mathbb{R}^d)}),$$

with  $S^\varepsilon = (\operatorname{Id} + \varepsilon\theta)S$  and the shape differentiability of  $C$  hence follows.

---

6. Recall that the scalar triple product of three vectors  $a, b, c \in \mathbb{R}^3$  is given by  $\langle a, (b \times c) \rangle$  and coincides with the (signed) volume of the parallelepiped defined by the three vectors. Therefore, the scalar triple product is preserved by a circular shift of the triple  $(a, b, c)$ .



## II.4 Numerical implementation

The results obtained above are intrinsic in the sense that they do not depend on a specific parametrization of the objects (surfaces, magnetic field, electric current, ...). There are several ways to represent them numerically. We have chosen to use what is, to the best of the authors' knowledge, the classical approach in the stellarator community. In particular:

- Surfaces, vector fields and magnetic fields are represented by Fourier coefficients. We detail the parametrization in Section II.4.1.
- Stellarator symmetry is imposed on all the objects. We refer to [IPW19, Section 12.3] and [DH98] for details and justifications of the stellarator symmetry.
- As mentioned in Remark II.4,  $BS_S$  is slightly modified. Not only the optimization space  $\mathcal{F}_S^0$  is replaced by a suitable affine subspace of it, but also we restrict the image of  $BS_S$  to the plasma boundary. We provide further details in Sections II.4.1.2 and II.4.1.3 and Section II.A.1.3.

### II.4.1 Parametrization issues

#### II.4.1.1 Surface representation

We represent a toroidal surfaces  $S$  as the image of the two-dimensional flat torus  $T = (\mathbb{R}/\mathbb{Z})^2$  by an embedding

$$\begin{aligned} \psi : T &\rightarrow \mathbb{R}^3 \\ (u, v) &\mapsto \psi(u, v). \end{aligned}$$

Stellarators often exhibits a discrete symmetry by rotation. For example W7X is invariant by the rotation of angle  $2\pi/5$  along the vertical axis and NCSX has an invariance by the rotation of angle  $2\pi/3$ . To reduce the complexity, we only represent one module of the surface and we denote by  $N_p$  the number of modules needed to generate the entire surface (using rotations of angle  $2\pi/N_p$ ). We introduce the cylindrical coordinates  $(R, \varphi, Z)$ . We will make the assumption of no toroidal folding, i.e., that the intersection of each half plane  $\{\varphi = \text{constant}\}$  with  $S$  is a single loop. We express  $\psi$  in cylindrical coordinates  $(R(u, v), \frac{2\pi v}{N_p}, Z(u, v))$  as

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} R(u, v) \cos(\frac{2\pi v}{N_p}) \\ R(u, v) \sin(\frac{2\pi v}{N_p}) \\ Z(u, v) \end{pmatrix}, \quad (u, v) \in T.$$

Then we develop  $R$  and  $Z$  in Fourier components and we impose the stellarator-symmetry

$$R(u, v) = \sum_{m \geq 0} \sum_{n \in \mathbb{Z}} R_{m,n} \cos(2\pi(mu + nv)), \quad (\text{II.18})$$

$$Z(u, v) = \sum_{m \geq 0} \sum_{n \in \mathbb{Z}} Z_{m,n} \sin(2\pi(mu + nv)). \quad (\text{II.19})$$

Note the absence of sin terms for  $R$  and cos terms for  $Z$ . For the numerical simulation, we truncate the number of Fourier components in (II.18) and (II.19).

**Remark II.18.** *The cost considered in this chapter only depends on the surface (and is independent of its parametrization  $\psi$ ). On the toroidal direction, we have already imposed that  $\varphi = 2\pi v/N_p$ . On the other hand, we can compose  $\psi$  with any diffeomorphism  $f_v : \mathbb{R}/\mathbb{Z} \rightarrow \mathbb{R}/\mathbb{Z}$*

on the poloidal direction  $u$ . Namely,  $\psi(f_v(u), v)$  and  $\psi(u, v)$  have the same image for a fixed  $v$ . Thus our problem is invariant under the action of a smooth family of diffeomorphisms. This extra degree of freedom has two consequences:

- If we use a regular discretization for the surface  $(\psi(\frac{i}{n_u}, \frac{j}{n_v}))_{i,j}$  of size  $n_u \times n_v$ , we need  $|\partial_u \psi|$  and  $|\partial_v \psi|$  to be as regular as possible.
- As we take a finite number of harmonics, we would like to “compress” as much as possible the information on the shape by using low harmonics.

This problem has been studied in the plasma community in [HB98] and gave rise to the notion of spectrally optimized Fourier series. Nevertheless, we would like to highlight that this approach is extrinsic (it depends on the parametrization) and should not be used for other purposes than fixing the gauge invariance. We have not implemented it since our numerical results empirically already provided a reasonable regularity on the poloidal parametrization.

### II.4.1.2 Magnetic field representation

In the previous sections, we represented the target magnetic field as a three-dimensional vector field in the plasma domain. Let  $\Gamma_t$  be a toroidal loop inside the plasma domain. As proved in Appendix II.A.1.3, thanks to the structure of Maxwell’s equations, the magnetic field inside  $P$  is entirely determined by its normal component along  $\partial P$  and its line integral (also called circulation) along  $\Gamma_t$ . By Stoke’s theorem (also known as Ampere circuital law in electromagnetism), the line integral of the magnetic field along  $\Gamma_t$  is equal to the total flux of the electric current across any surface enclosed by  $\Gamma_t$ . This quantity is called the *total poloidal current* and is denoted by  $I_p$ .

As a consequence, it is reasonable to minimize

$$\chi_B^2(j) = \int_{\partial P} \langle (\text{BS}_S j - B_T), \nu \rangle^2 d\mu_{\partial P} \quad (\text{II.20})$$

with the total poloidal current of  $j$  fixed, where  $\nu$  denotes the outward normal unit vector to  $\partial P$ .

This idea has been used by physicists for a long time, for example [Mer86; Lan17]. Besides, if we consider two currents distribution  $j$  and  $\tilde{j}$  on two toroidal surfaces  $S$  and  $\tilde{S}$  outside of  $P$  with the same total poloidal currents, there exists  $C > 0$  (independent of  $j$  and  $\tilde{j}$ ) such that the induced magnetic field in  $P$  satisfies

$$\| \text{BS}_S j - \text{BS}_{\tilde{S}} \tilde{j} \|_{L^2(P, \mathbb{R}^3)}^2 \leq C \int_{\partial P} \langle (\text{BS}_S j - \text{BS}_{\tilde{S}} \tilde{j}), \nu \rangle^2 d\mu_{\partial P}.$$

We provide mathematical proofs of these facts in Appendix II.A.1.3.

We also use a normal target magnetic field that respects the stellarator symmetry, that is,

$$\langle B_T, \nu \rangle(\psi(u, v)) = \sum_{m \geq 0} \sum_{n \in \mathbb{Z}} B_{m,n} \sin(2\pi(mu + nv)).$$

As before, we truncate the Fourier series to obtain a numerically tractable expression.

### II.4.1.3 Current-sheet representation

As mentioned in the previous section, we need to parameterized all divergence-free vector field on  $S$  with a fixed total poloidal current  $I_p$ . In Appendix II.A.1.2 we prove that

$$\mathcal{F}_T^0 = \{\nabla^\perp \Phi + \lambda_1 \partial_u + \lambda_2 \partial_v \mid \Phi \in H^1(T), (\lambda_1, \lambda_2) \in \mathbb{R}^2\} \quad (\text{II.21})$$

with  $\nabla^\perp \Phi = \frac{\partial \Phi}{\partial u} \partial_v - \frac{\partial \Phi}{\partial v} \partial_u$ .

The following lemma describes how embeddings induce isomorphisms between  $\mathcal{F}_T^0$  and  $\mathcal{F}_S^0$ .

**Lemma II.19.** *Let  $\psi : T \rightarrow \mathbb{R}^3$  be an embedding with  $S = \psi(T)$  and consider*

$$\begin{aligned} \Psi : \mathfrak{X}(T) &\rightarrow \mathfrak{X}(S) \\ X &\mapsto \frac{D\psi X}{\left| \frac{\partial \psi}{\partial u} \times \frac{\partial \psi}{\partial v} \right|}. \end{aligned}$$

*Then  $\Psi$  induces an isomorphism between  $\mathcal{F}_T^0$  and  $\mathcal{F}_S^0$ .*

The proof is completely similar to that of Lemma II.13.

Let us suppose now that  $(u, v)$  are poloidal and toroidal coordinates for the parameterization  $\psi$ , that is,

- $\Gamma_p : \mathbb{R}/\mathbb{Z} \ni t \mapsto \psi(t, 0) \in S$  is a loop doing exactly one poloidal turn (and 0 toroidal ones);
- $\Gamma_t : \mathbb{R}/\mathbb{Z} \ni t \mapsto \psi(0, t) \in S$  is a loop doing exactly one toroidal turn (and 0 poloidal ones).

Besides, as is it in general the convention in the dedicated literature, we assume that  $\psi$  is orientation reversing, meaning that

$$(\Psi(\partial_u), \Psi(\partial_v), -\nu) \text{ is direct}, \quad (\text{II.22})$$

with  $\nu$  the outward normal vector field.

**Lemma II.20.** *Let  $X = \nabla^\perp \Phi + I_p \partial_u + I_t \partial_v$ . Then the poloidal (respectively, toroidal) flux of  $\Psi(X)$ , i.e., the flux of  $\Psi(X)$  across  $\Gamma_t$  (respectively,  $\Gamma_p$ ), is given by  $I_p$  (respectively,  $I_t$ ).*

*Proof.* Remark that  $\text{div } \Psi(X) = 0$  ensure that the flux across any loop depends only on the isotopic class of the loop considered. Recall that the flux of  $\Psi(X)$  across some loop  $\Gamma$  is given by

$$\oint_{\Gamma} \langle \Psi(X), \left( \frac{\Gamma'}{|\Gamma'|} \times -\nu \right) \rangle d\mu_{\Gamma} = \int_0^1 \langle \Psi(X), (\Gamma' \times -\nu) \rangle (\Gamma(t)) dt,$$

where the choice of the sign in  $-\nu$  is due to the convention (II.22) and  $\Gamma'$  denotes the derivative

of  $\Gamma$ . Thus, the flux across  $\Gamma_t$  (the poloidal flux) is

$$\begin{aligned}
\int_0^1 \langle \Psi(X), \left( \frac{\partial \psi}{\partial v} \times -\nu \right) \rangle (\Gamma_t(t)) dt &= - \int_0^1 \langle \nu, (\Psi(X) \times \frac{\partial \psi}{\partial v}) \rangle (\Gamma_t(t)) dt \\
&= \int_0^1 \frac{1}{\left| \frac{\partial \psi}{\partial u} \times \frac{\partial \psi}{\partial v} \right|^2} \langle \left( \frac{\partial \psi}{\partial u} \times \frac{\partial \psi}{\partial v} \right), (D\psi(\nabla^\perp \Phi + I_p \partial_u + I_t \partial_v) \times \frac{\partial \psi}{\partial v}) \rangle (\Gamma_t(t)) dt \\
&= \int_0^1 \frac{1}{\left| \frac{\partial \psi}{\partial u} \times \frac{\partial \psi}{\partial v} \right|^2} \langle \left( \frac{\partial \psi}{\partial u} \times \frac{\partial \psi}{\partial v} \right), \left( I_p - \frac{\partial \Phi}{\partial v} \right) \frac{\partial \psi}{\partial u} \times \frac{\partial \psi}{\partial v} \rangle (\Gamma_t(t)) dt \\
&= \int_0^1 \left( I_p - \frac{\partial \Phi}{\partial v}(0, t) \right) dt \\
&= I_p.
\end{aligned}$$

The computation of the flux across  $\Gamma_p$  is analogous.  $\square$

Thanks to this lemma, in order to minimize on  $\mathcal{F}_S^0$  we fix  $I_p$  and  $I_t$  and minimize with respect to  $\Phi$ . Indeed,  $I_p$  is fixed by the toroidal circulation of the target magnetic field (see II.A.1.3), whereas  $I_t$  is usually set to 0. This second condition is necessary to ensure the existence of "poloidal coils". Otherwise, no closed field lines would realize one poloidal turn and zero toroidal ones. Thus, the set of admissible currents is described by

$$J_{\text{adm}}(S) = \{ \Psi(\nabla^\perp \Phi + I_p \partial_u + I_t \partial_v) \mid \Phi \in H^1(T) \}.$$

We say that  $\Phi$  is the *scalar current potential*. By stellarator symmetry, its expansion in Fourier series is

$$\Phi(u, v) = \sum_{m \geq 0} \sum_{n \in \mathbb{Z}} \Phi_{m,n} \sin(2\pi(mu + nv)).$$

Let us denote

$$j_{a,S} = \Psi(I_p \partial_u + I_t \partial_v) \quad \text{and} \quad \hat{\mathcal{F}}_S^0 = \{ \Psi(\nabla^\perp \Phi) \mid \Phi \in H^1(T) \}.$$

It is straightforward that the affine decomposition  $J_{\text{adm}}(S) = j_{a,S} + \hat{\mathcal{F}}_S^0$  is compatible with  $\Phi^\varepsilon$  (cf. (II.11)), meaning that for any shape deformation  $\theta$ ,

$$\Phi^\varepsilon(j_{a,S}) = j_{a,S^\varepsilon} \quad \text{and} \quad \Phi^\varepsilon(\hat{\mathcal{F}}_S^0) = \hat{\mathcal{F}}_{S^\varepsilon}^0.$$

Thus, we can consider the restriction of  $\text{BS}_S$  to  $\hat{\mathcal{F}}_S^0$  that we will denote  $\widehat{\text{BS}}_S$ . Let  $\widehat{\text{BS}}_S^\dagger$  be its adjoint (in  $\hat{\mathcal{F}}_S^0$ ) and  $\hat{\pi}_S$  the orthogonal projector defined in  $\mathcal{F}_S^0$  onto  $\hat{\mathcal{F}}_S^0$ . Then Lemma II.3 holds and the expression of the unique minimizer is given by

$$\begin{aligned}
\hat{j}_S &= (\lambda \text{Id} + \widehat{\text{BS}}_S^\dagger \widehat{\text{BS}}_S)^{-1} \left( \widehat{\text{BS}}_S^\dagger (B_T - \text{BS}_S j_S^a) - \lambda \hat{\pi}_S j_S^a \right), \quad j_S = j_S^a + \hat{j}_S \\
C(S) &= \lambda \|j_S\|_{\mathcal{F}_S}^2 + \|\text{BS}_S j_S - B_T\|_{L^2(P, \mathbb{R}^3)}^2.
\end{aligned}$$

## II.4.2 Implementation

We wrote our implementation in python using several scientific computing open source libraries and, in particular:

- Numpy [Har+20] for array computation,
- Scipy [Vir+20] for the implementation of the Broyden-Fletcher-Goldfarb-Shanno (BFGS) minimization algorithm,
- Opt\_einsum [SG18] for optimizing tensor construction,
- Dask [Das16] for large array and efficient scientific computing parallelization,
- Matplotlib [Hun07] and Mayavi [RV11] for plotting and graphic representations.

The full code is available on our gitlab<sup>7</sup> under MPL 2 license.

The constraints on the perimeter, the reach and the plasma-CWS distance are implemented as a nonlinear penalization cost which blows up rapidly once the values exceed (or subceed) a given threshold. We refer to the code documentation for further details<sup>8</sup>.

## II.4.3 Numerical results

In what follows, the data used for the simulations come from the NCSX stellarator equilibrium known as LI383 [Zar+01]. We will also use as reference CWS the one used in the original REGCOIL paper [Lan17].

We present here four simulations. We used either  $\lambda = 2.5e^{-16}$  or  $\lambda = 5.1e^{-19}$  as regularization parameter in the expression of the cost  $C$ . We mesh the CWS and the plasma surface with  $64 \times 64$  grids. The scalar current potential  $\Phi$  is developed in Fourier series up to order 12 in both directions. The optimization is performed with up to 2000 steps of the BFGS algorithm. In our laptop<sup>9</sup>, each step takes about 50 seconds. Hence each simulation is performed in about 30 hours. In every simulation we implemented a penalization on the perimeter of the CWS (penalization above  $56m^2$ ) and on plasma-CWS distance (penalization under 20cm). We also implemented a reach penalization for two simulations (penalization under 7.69cm). Let us call Ref the initial CWS. We use DP to refer to the simulations with distance and perimeter penalization and DPR for those with additional reach penalization.

In Table II.1, we observe that our optimized surface DPR achieves an important reduction of the cost function ( $\chi_B^2$  nearly divided by 4 and a reduction of one third of  $\|j\|_{\mathcal{F}_S}^2$ ) while preserving a comparable distance to the plasma and a perimeter very close to the reference shape. Such an optimized surface is plotted in Figures II.5 and II.6. Similarly, in Table II.2 the result of the DPR simulation is a reduction of 38% of  $\chi_B^2$  and a division by 3.8 of  $\|j\|_{\mathcal{F}_S}^2$ .

Figure II.7 illustrates the convergence history of the implemented optimization algorithm.

<i>type</i>	$\chi_B^2$	$\ j\ _{\mathcal{F}_S}^2$	$C(S)$	Distance ( <i>m</i> )	Perim. ( $m^2$ )	Reach ( <i>m</i> )	nb of iter.
Ref	$4.80e^{-03}$	$1.43e^{+14}$	$4.06e^{-02}$	$1.92e^{-01}$	$5.57e^{+01}$	$8.40e^{-02}$	
DPR	$1.23e^{-03}$	$9.48e^{+13}$	$2.49e^{-02}$	$1.99e^{-01}$	$5.60e^{+01}$	$7.69e^{-02}$	775
DP	$1.05e^{-03}$	$7.36e^{+13}$	$1.95e^{-02}$	$2.00e^{-01}$	$5.60e^{+01}$	$4.33e^{-06}$	2000

Table II.1 – Numerical results for  $\lambda = 2.5e^{-16}$ ,  $\chi_B$  is defined in (II.20).

7. <https://plmlab.math.cnrs.fr/rrobin/stellacode>

8. <https://rrobin.pages.math.cnrs.fr/stellacode/>

9. equipped with a 6 cores i7-9850H CPU

<i>type</i>	$\chi_B^2$	$\ j\ _{\mathcal{F}_S}^2$	$C(S)$	Distance ( <i>m</i> )	Perim. ( $m^2$ )	Reach ( <i>m</i> )	nb of iter.
Ref	$1.44e^{-04}$	$4.91e + 14$	$3.94e^{-04}$	$1.92e^{-01}$	$5.57e^{+01}$	$8.40e^{-02}$	
DPR	$9.05e^{-06}$	$1.26e + 14$	$7.34e^{-05}$	$2.00e^{-01}$	$4.17e^{+01}$	$7.69e^{-02}$	2000
DP	$7.16e^{-06}$	$1.21e + 14$	$6.90e^{-05}$	$2.00e^{-01}$	$5.60e^{+01}$	$8.33e^{-05}$	2000

Table II.2 – Numerical results for  $\lambda = 5.1e^{-19}$ ,  $\chi_B$  is defined in (II.20).

**Remark II.21.** *Without penalization on the reach, one naturally obtains better results (as less constraints are applied on the set of admissible shapes). Nevertheless, such an approach seems a very bad idea:*

- *theoretically, since the existence of an optimal shape is guaranteed only for bounded reach,*
- *numerically, because sharper and sharper “spikes” appear, as shown in Figure II.4. Those spikes can be arbitrary long while still keeping a finite perimeter (and encapsulated volume).*

## II.A Appendix

### II.A.1 Some differential geometry

In this section, we recall some basics fact about differential geometry and vector fields on toroidal surfaces and domains.

#### II.A.1.1 Hodge decomposition

We recall in this part some notions of differential geometry and in particular of Hodge theory. We refer to [Jos17, Chapter 3] and [Lee12] for details and precise definitions in the smooth setting. Although we are only interested in  $\mathcal{C}^{1,1}$  manifolds in this chapter, note for the sake of completeness that details on Hodge theory for Lipschitz manifolds can be found for instance in [Tel83].

The Hodge decomposition is a powerful tool which gives an orthogonal decomposition of the space of square integrable  $p$ -forms on a Riemannian closed manifold  $M$  as

$$L_p^2(M) = B_p \oplus B_p^* \oplus \mathcal{H}_p,$$

where  $B_p$  is the  $L^2$ -closure of  $\{d\alpha \mid \alpha \in \Omega^{p-1}(M)\}$ ,  $B_p^*$  is the  $L^2$ -closure of  $\{d^*\beta \mid \beta \in \Omega^{p+1}(M)\}$  ( $d^*$  is the coderivative), and  $\mathcal{H}_p$  is the set  $\{\omega \in \Omega^p(M) \mid \Delta_H \omega = 0\}$  of harmonic  $p$ -forms with  $\Delta_H$  the Hodge Laplacian.

We apply this result to the simple case of 1-forms on a two-dimensional closed Riemannian manifold  $S$ . We recall a few basics facts:

- 1-forms and vector fields can be identified thanks to the Riemannian metric. This isomorphism is called the *musical isomorphism* and we denote by  $X^b$  the 1-form defined as the image of a vector field  $X$  thought the musical isomorphism. Conversely  $w^\#$  denotes the vector field which is the image of the 1-form  $\omega$ .
- The divergence of a vector field  $X$  is  $-d^*X^b$ .
- $d \circ d = 0$  and  $d^* \circ d^* = 0$ .

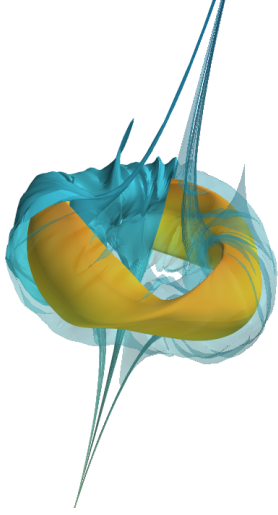


Figure II.4 – Main pattern of the optimal CWS for the DP simulation with  $\lambda = 2.5e^{-16}$ , top and bottom spikes have been truncated.

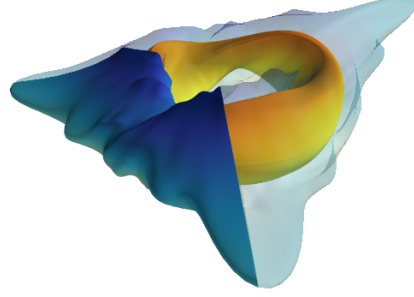


Figure II.5 – Main pattern of the CWS (blue and white) for the DPR simulation with  $\lambda = 2.5e^{-16}$ .

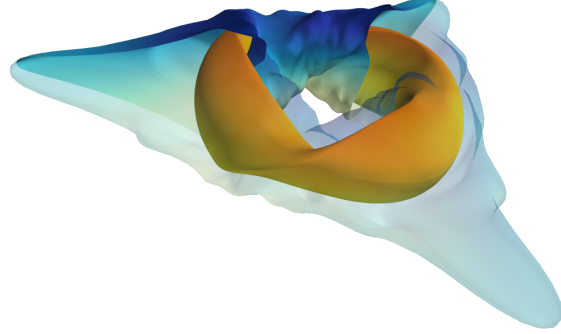


Figure II.6 – Main pattern of the optimal CWS (blue and white) for the DPR simulation with  $\lambda = 2.5e^{-16}$ .

—  $\Delta_H \alpha = 0$  is equivalent to the system of equations 
$$\begin{cases} d\alpha = 0, \\ d^* \alpha = 0. \end{cases}$$

We want to show that the space of “divergence-free” 1-forms (i.e.,  $\ker d^*$ ) coincides with  $B_1^* \oplus \mathcal{H}_p$ . It is clear that the latter space is contained in  $\ker d^*|_{\Omega^1(M)}$ . Conversely, for every exact form  $\omega$ , i.e., such that  $\omega = df$  with  $f \in \mathcal{C}^\infty(M)$ , one has  $d^* \omega = d^* df = \Delta_H f$ . We recall that the Hodge Laplacian coincides with the Laplace–Beltrami operator on 0-forms. But  $\Delta_H f = 0$  implies that  $f$  is constant on each connected component, thus  $d^* \omega = 0$  implies that  $\omega = 0$ . As a result the space of divergence-free 1-forms is  $B_1^* \oplus \mathcal{H}_p = (B_1)^\perp$ .

Equivalently, the space of divergence-free vector fields coincides with the orthogonal to  $\{\nabla f \mid f \in \Omega^0(M)\}$ . In Appendix II.A.1.2 we give an explicit description of  $(B_1^* \oplus \mathcal{H}_p)^\#$  for the two-dimensional flat torus.

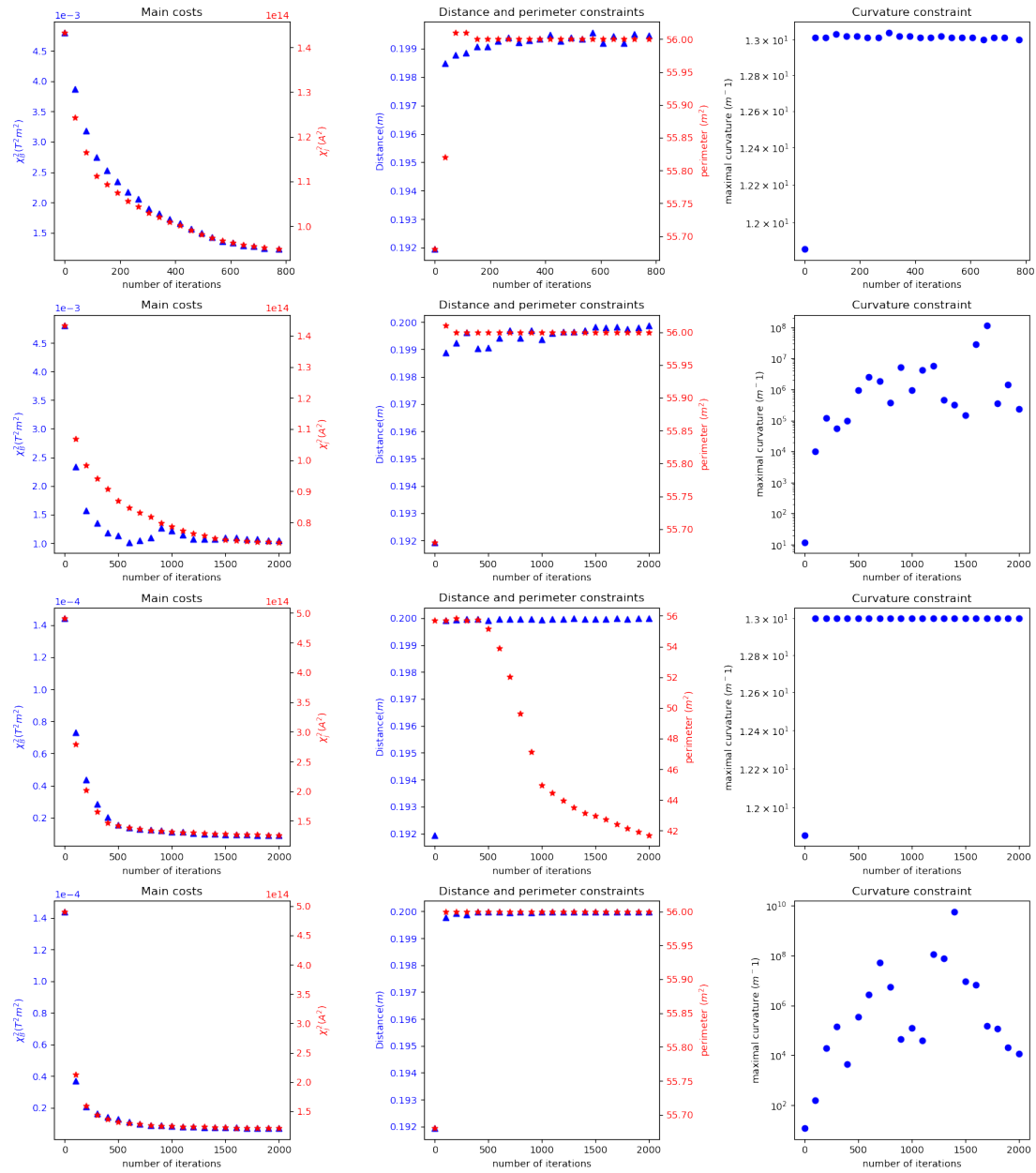


Figure II.7 – History of convergence for the implemented optimization algorithm. From left to right, evolution of the costs (left), distance and perimeter constraints (middle) and the curvature constraint (right) along the optimization process. From top to bottom: Table II.1 ( $\lambda = 2.5e^{-16}$ ) configurations DPR and DP, then table II.2 ( $\lambda = 5e^{-19}$ ) configurations DPR and DP.



### II.A.1.2 Divergence-free vector field on a flat torus

Let  $T = (\mathbb{R}/\mathbb{Z})^2$  be the flat torus with Cartesian parametrization  $(u, v)$ . We want to characterize the set of divergence-free vector fields on  $T$ .

As explained in II.A.1.1, we only need to characterize  $B_1^*(T)$  and  $\mathcal{H}_1(T)$ .

—  $B_1^*(T)$  is the  $L^2$ -closure of the 1-forms  $\frac{\partial \Phi}{\partial u} dv - \frac{\partial \Phi}{\partial v} du$  for  $\Phi \in \mathcal{C}^\infty(T)$ .

—  $\mathcal{H}_1(T)$  is a two-dimensional vector space as the first Betti number for a 2D torus satisfies  $b_1 = 2$ . We easily compute  $\mathcal{H}_1(T) = \{\lambda_1 du + \lambda_2 dv \mid (\lambda_1, \lambda_2) \in \mathbb{R}^2\}$ .

Using the musical isomorphism, we deduce that all divergence-free vector fields in  $L^2$  have the form given in Equation (II.21).

### II.A.1.3 Poisson equation on a toroidal 3D domain

Given a toroidal 3D domain  $P$ , we want to study the Maxwell equations in vacuum inside  $P$ . We introduce a toroidal loop  $\Gamma$  inside  $P$  and denote by  $I_p$  the electric current-flux across any surface enclosed by  $\Gamma$ . By the conservation of charges ( $\operatorname{div} j = 0$ ), this quantity is well defined. By smoothness of the Biot and Savart operator and of the plasma boundary  $\partial P$ , all functions considered in this appendix may be assumed to be  $\mathcal{C}^\infty$ .

**Lemma II.22.** *Let  $g$  be the normal magnetic field on  $\partial P$  (i.e., the normal component of  $B|_{\partial P}$ ). Then  $g$  and  $I_p$  determine completely the magnetic field  $B$  in  $P$ . Besides, there exists a constant  $C > 0$  such that for every other magnetic field  $\tilde{B}$  with the same total poloidal current,  $|B - \tilde{B}|_{L^2(P, \mathbb{R}^3)} \leq C|g - \tilde{g}|_{L^2(\partial P)}$  where  $\tilde{g}$  is the normal component of  $\tilde{B}|_{\partial P}$ .*

Before going to the proof of this statement, we emphasize that the structure of the space  $L^2(P, \mathbb{R}^3)$  is well understood and admits a generalized Hodge decomposition ( $P$  is not a closed manifold, thus part II.A.1.1 does not apply) from which the lemma follows easily. Such a decomposition is proved, for example, in [CDG02]. For completeness, we provide the following proof.

*Proof.* We have the cochain complex (meaning  $\operatorname{Im}(\nabla) \subset \ker(\operatorname{curl})$  and  $\operatorname{Im}(\operatorname{curl}) \subset \ker(\operatorname{div})$ )

$$\mathcal{C}^\infty(P) \xrightarrow{\nabla} \mathcal{C}^\infty(P, \mathbb{R}^3) \xrightarrow{\operatorname{curl}} \mathcal{C}^\infty(P, \mathbb{R}^3) \xrightarrow{\operatorname{div}} \mathcal{C}^\infty(P).$$

For simply connected domains of  $\mathbb{R}^3$ , the complex is an exact sequence, meaning that  $\operatorname{Im}(\nabla) = \ker(\operatorname{curl})$  and  $\operatorname{Im}(\operatorname{curl}) = \ker(\operatorname{div})$ . For a 3D toroidal domain, the dimension of the quotient space  $\frac{\ker(\operatorname{curl})}{\operatorname{Im}(\nabla)}$  is always one. This is a consequence of the De Rham cohomology of  $P$ . We refer to [Lee12, Diagram 16.15] for further details.

Thus  $\ker(\operatorname{curl}) \not\subset \operatorname{Im}(\nabla)$ , i.e., there exists  $X \in \mathcal{C}^\infty(P, \mathbb{R}^3)$  such that  $X \notin \operatorname{Im}(\nabla)$  and  $\operatorname{curl} X = 0$ . Without loss of generality, we can suppose that  $\operatorname{div} X = 0$ . Indeed, it is enough to consider  $X' = X - \nabla \zeta$  with  $\zeta$  solution of the Poisson equation

$$\begin{aligned} \Delta \zeta &= \operatorname{div} X && \text{in } P, \\ \zeta &= 0 && \text{on } \partial P. \end{aligned}$$

To have an intuition, the reader can think of the vector field  $X = \frac{e_\theta}{R}$  in  $\mathbb{R}^3 \setminus \{R = 0\}$  in cylindrical coordinates  $(R, \theta, z)$ . This vector field is divergence and curl free but is not in the image of a gradient.

We recall Maxwell's equations for a the static magnetic field in vacuum:

$$\operatorname{curl} B = 0 \quad \text{in } P, \quad (\text{II.23})$$

$$\operatorname{div} B = 0 \quad \text{in } P. \quad (\text{II.24})$$

Equation (II.23) implies that there exist a scalar potential  $\xi \in \mathcal{C}^\infty(P)$  and  $\alpha \in \mathbb{R}$  such that

$$B = \nabla \xi + \alpha X.$$

Using Stoke's theorem, the line integral of  $B$  along  $\Gamma$  is given by the total flux  $I_p$  of electric currents across any surface enclosed by  $\Gamma$ . In particular the contribution of the term  $\nabla \xi$  to  $I_p$  is zero, yielding

$$I_p = \oint_{\Gamma} \langle B, \frac{\Gamma'}{|\Gamma'|} \rangle d\mu_{\Gamma} = \oint_{\Gamma} \langle (\nabla \xi + \alpha X), \frac{\Gamma'}{|\Gamma'|} \rangle d\mu_{\Gamma} = \alpha \oint_{\Gamma} \langle X, \frac{\Gamma'}{|\Gamma'|} \rangle d\mu_{\Gamma}.$$

The quantity  $\oint_{\Gamma} \langle X, \frac{\Gamma'}{|\Gamma'|} \rangle d\mu_{\Gamma}$  is nonzero, since otherwise, by the De Rham isomorphism,  $X$  would be in  $\operatorname{Im} \nabla$ . Thus,  $\alpha$  is uniquely determined by  $I_p$ , since  $X$  does not depend on  $B$ .

Equation (II.24) together with the normal component of  $B$  on  $\partial P$  give

$$\begin{aligned} \Delta \xi &= 0 \quad \text{in } P, \\ \partial_n \xi &= g - \alpha \langle X, n \rangle \quad \text{on } \partial P. \end{aligned} \quad (\text{II.25})$$

Thus  $\xi$  is determined by  $g$  and  $\alpha$  as the unique solution of a Laplace equation with Neumann boundary conditions.

Finally, let  $B = \nabla \xi + \alpha X$  and  $\tilde{B} = \nabla \tilde{\xi} + \alpha X$  with  $\xi$  and  $\tilde{\xi}$  the solutions of equation (II.25) corresponding to  $g$  and  $\tilde{g}$ , respectively. The difference  $\delta = \xi - \tilde{\xi}$  is solution of

$$\begin{aligned} \Delta \delta &= 0 \quad \text{in } P, \\ \partial_n \delta &= g - \tilde{g} \quad \text{on } \partial P. \end{aligned}$$

By well-posedness of the Laplace equation with Neumann boundary conditions, there exists a constant  $C(\partial P)$  such that  $|\nabla \delta|_{H^{1/2}} \leq C(\partial P) |g - \tilde{g}|_{L^2(\partial P)}$ . Thus, there exists  $C > 0$  independent of  $g$  and  $\tilde{g}$  such that

$$|B - \tilde{B}|_{L^2(P, \mathbb{R}^3)} \leq C |g - \tilde{g}|_{L^2(\partial P)},$$

concluding the proof.  $\square$

## II.A.2 Reach constraint and sets of positive reach

In this section, we gather some reminders about the notion of reach. We refer to [DZ11, Chapter 6, Section 6] for more exhaustive explanations around this notion.

Recall first that, if  $V$  is a nonempty subset of  $\mathbb{R}^n$ , its *skeleton*, denoted by  $\operatorname{Sk}(V)$ , is the set of all points in  $\mathbb{R}^n$  whose projection onto  $V$  is not unique. The set  $V$  is said to have a *positive reach* whenever there exists  $h > 0$  such that

$$\text{every point } v \text{ of the tubular neighborhood } U_h(V) \text{ has a unique projection point on } V. \quad (\text{II.26})$$

Recall that the definition of  $U_h(V)$  is provided in Section II.1.3. One thus defines the reach of  $V$  as

$$\operatorname{Reach}(V) = \sup\{h > 0 \mid (\text{II.26}) \text{ is satisfied}\}.$$

An equivalent definition of the reach writes

$$\text{Reach}(V) = \inf\{\text{Reach}(V, v) \mid v \in V\},$$

where

$$\text{Reach}(V, v) = \begin{cases} 0 & \text{if } v \in \partial\bar{V} \cap \overline{\text{Sk}(V)} \\ \sup\{h > 0 \mid \text{Sk}(V) \cap B_h(v) = \emptyset\} & \text{otherwise,} \end{cases}$$

where  $B_h(v)$  denotes the Euclidean open ball centered at  $v$  with radius  $h$ .

The notion of reach is actually closely related to the so-called uniform ball condition. The next result makes this relationship precise.

**Theorem II.23** (Theorems 2.6 and 2.7 in [Dal18]). *Let  $\Omega$  be an open subset of  $\mathbb{R}^n$  with a nonempty boundary.*

— *If there exists  $h > 0$  such that  $\Omega$  satisfies a uniform ball condition, namely*

$$\forall x \in \partial\Omega, \exists d_x \in \mathbb{R}^n \mid \|d_x\|_{\mathbb{R}^n} = 1, B_h(x - hd_x) \subset \Omega \text{ and } B_h(x + hd_x) \subset \mathbb{R}^n \setminus \Omega, \quad (\text{II.27})$$

*then  $\partial\Omega$  has a positive reach which is larger than  $h$  and the Lebesgue measure of  $\partial\Omega$  in  $\mathbb{R}^n$  is equal to 0. Furthermore,  $\partial\Omega$  is a  $\mathcal{C}^{1,1}$  hypersurface of  $\mathbb{R}^n$ .*

— *If  $\partial\Omega$  is a nonempty compact  $\mathcal{C}^{1,1}$ -hypersurface of  $\mathbb{R}^n$ , then there exists  $h > 0$  such that  $\Omega$  satisfies (II.27).*

— *If  $\partial\Omega$  has a positive reach and if its Lebesgue measure in  $\mathbb{R}^n$  is equal to 0, then it satisfies the ball condition (II.27) for every  $h \in (0, \text{Reach}(\partial\Omega))$  and in particular,  $\partial\Omega$  is a  $\mathcal{C}^{1,1}$  hypersurface of  $\mathbb{R}^n$ .*

### II.A.3 Jacobian determinant and changes of variables on manifolds

We recall here some basic results about integration on manifolds which can be found in [AMR88] or [Ste13] for example. Let  $M$  and  $N$  be two compact Riemannian  $n$ -dimensional manifolds with volume forms  $\mu_M$  and  $\mu_N$ . Let  $\vartheta : M \rightarrow N$  be an orientation preserving diffeomorphism. Then, for any  $v \in \mathcal{C}^1(N)$ ,

$$\int_N v d\mu_N = \int_M d\vartheta^*(v\mu_N)$$

Besides, there exists a function  $J(\mu_M, \mu_N)\vartheta$  on  $M$ , called the *Jacobian determinant*, such that  $\vartheta^*\mu_N = [J(\mu_M, \mu_N)\vartheta]\mu_M$ . This implies the well-known change of variable formula

$$\int_N v d\mu_N = \int_M (v \circ \vartheta)[J(\mu_M, \mu_N)\vartheta] d\mu_M. \quad (\text{II.28})$$

In the particular, when  $M$  and  $N$  are closed 2-dimensional submanifolds of  $\mathbb{R}^3$  of class  $\mathcal{C}^{1,1}$ , and  $\vartheta$  is of the type  $\vartheta = \text{Id} + \theta$  with  $\theta \in W^{2,\infty}(\mathbb{R}^3, \mathbb{R}^3)$  and  $\|\theta\|_{W^{2,\infty}(\mathbb{R}^3, \mathbb{R}^3)} < 1$  (so that  $\vartheta$  defines a diffeomorphism in  $W^{2,\infty}(\mathbb{R}^3, \mathbb{R}^3)$ ), one has

$$J(\mu_M, \mu_N)\vartheta = \det(\text{Id} + D\theta) |((\text{Id} + D\theta)^\top)^{-1}\nu|$$

with  $\nu$  the outward normal to  $M$ . We refer for instance to [HP18, Section 5.4.5] for a shape optimization oriented proof or [Jos17, Chapter 5] for a more differential geometry oriented presentation.

## Chapter III

# Existence of surfaces optimizing geometric and PDE shape functionals under reach constraint

This chapter is taken from the following submitted article (also referred as [PRS22a]):

Y. Privat, R. Robin, and M. Sigalotti. *Existence of surfaces optimizing geometric and PDE shape functionals under reach constraint*. 2022. arXiv: 2206.04357 [math],

This chapter deals with the existence of hypersurfaces minimizing general shape functionals under certain geometric constraints. We consider as admissible shapes orientable hypersurfaces satisfying a so-called *reach* condition, also known as the uniform ball property, which ensures  $\mathcal{C}^{1,1}$  regularity of the hypersurface. In this chapter, we revisit and generalise the results of [GY13; Dal18; Dal20]. We provide a simpler framework and more concise proofs of the results contained in these references and extend them to a new class of problems involving PDEs. Indeed, by using the signed distance introduced by Delfour and Zolesio (see for instance [DZ11]), we avoid the intensive and technical use of local maps, as was the case in the above references. Our approach, originally developed to solve an existence problem in [PRS22b], can be easily extended to costs involving different mathematical objects associated with the domain, such as solutions of elliptic equations on the hypersurface.

### III.1 Framework and main results

#### III.1.1 Introduction

In this chapter, we are interested in the question of the existence of optimal sets for shape optimization problems involving surfaces. More precisely, we are interested in shape functionals written as

$$J(\Omega) = \int_{\partial\Omega} j(x, \nu_{\partial\Omega}(x), B_{\partial\Omega}(x)) d\mu_{\partial\Omega}(x)$$

where  $\Omega$  denotes a smooth subset of  $\mathbb{R}^d$ , the wording ‘smooth’ being understood at this stage such that all the involved quantities make sense,  $\nu$  denotes the outward pointing normal vector

to  $\partial\Omega$  and  $B_{\partial\Omega}$  is either a purely geometric quantity such as the mean curvature, or the solution of a PDE on  $\partial\Omega$  or on  $\Omega$ .

We are then interested in the existence of solutions for the optimization problem

$$\inf_{\Omega \text{ admissible}} J(\Omega).$$

This kind of problem is very generic. What matters here is that the standard techniques, exposed and developed for example in [DZ11; HP18], do not apply to  $d - 1$  objects and it is necessary to adopt a particular approach. The first question to ask is the choice of the set  $\mathcal{O}_{\text{ad}}$  of all admissible domains. Since the shape functionals we consider involve geometric quantities of the type “outward normal vector to the boundary” or “mean curvature”, it is necessary that the manipulated surfaces are not too irregular. For this reason, we choose to impose a constraint that guarantees a uniform regularity, say  $\mathcal{C}^{1,1}$ , of the manipulated sets. This uniform regularity constraint is imposed by using the notion of “reach”. Thus, the set  $\mathcal{O}_{\text{ad}}$  represents the set of surfaces having a reach uniformly bounded by below. The precise definition of this notion will be given in Section III.1.3.

This kind of problem has been the subject of recent studies and results [GY13; Dal18; Dal20], which have provided positive answers to the existence issues. In their approach, the authors used an efficient, but nevertheless laborious, approach based on the parametrization of the manipulated surfaces, seen as regular manifolds, using local charts.

The objective of this chapter is to promote a different approach, based on the extension of the functions defined on the manipulated surfaces to volume neighborhoods, the introduction of an extruded surface and the rewriting of the surface integrals as volume integrals using ad-hoc variable changes. This is a methodological chapter, in which a proof method is presented that may work in many cases. The results contained in the article illustrate this point. We discuss possible generalizations of these results in a concluding section.

This method allows to gain conciseness and provides much shorter and direct existence proofs than in the above references. The method also allows to extend the field of investigation to new families of problems, involving the solution of a PDE defined on a hypersurface. Nevertheless, some arguments used by the authors of [GY13; Dal18; Dal20] cannot be shortened by using our approach. We have therefore chosen to expose our method in a short article, in which we detail all the parts of the proof that can be condensed and we make the necessary reminders concerning the results that cannot be condensed.

The chapter is organized as follows: we introduce the definition of the reach of a surface as well as the class of admissible sets we will deal with in Section III.1.3. The main results of this chapter, regarding several existence results for shape optimization problems involving surfaces, are provided in Section III.1.4. The whole section III.2 is devoted to the proofs of the main results. In these proofs, we detail the arguments based on our approach and leading to simplified proofs of the results in [GY13; Dal18; Dal20]. In order to illustrate the potential of our approach, we also provide an existence result involving a general functional depending on the solution of a PDE on the sought manifold.

### III.1.2 Notations

Let us recall some classical notations used throughout this chapter:

- For the sake of notational simplicity, we will sometime use the notation  $\Gamma$  (resp.  $\Gamma_n$ ) to denote the hypersurfaces  $\partial\Omega$  (resp.  $\partial\Omega_n$ ).

- The Euclidean inner product (resp. norm) will be denoted  $\langle \cdot, \cdot \rangle$  (resp.  $\| \cdot \|$  or sometimes  $| \cdot |$  when no confusion with other notations is possible).
- Given two positive integers  $k \leq d$  and  $\Omega \subset \mathbb{R}^d$ ,  $\mathcal{H}^k(\Omega)$  denotes the  $k$ -dimensional Hausdorff measure of  $\Omega$ .
- Given  $\Omega \subset \mathbb{R}^d$ , the distance (resp., signed distance) to  $\Omega$  is defined for all  $x \in \mathbb{R}^d$  by

$$d_\Omega(x) = \inf_{y \in \Omega} \|x - y\| \quad (\text{resp., } b_\Omega(x) = d_\Omega(x) - d_{\mathbb{R}^d \setminus \Omega}(x)).$$

- Given  $\Omega \subset \mathbb{R}^d$  and  $h > 0$ , the tubular neighborhood  $U_h(\Omega)$  is defined as

$$U_h(\Omega) = \{x \in \mathbb{R}^d \mid d_\Omega(x) \leq h\}.$$

- Given  $\Omega \subset \mathbb{R}^d$ , the reach of  $\Omega$  is defined as

$$\text{Reach}(\Omega) = \sup\{h > 0 \mid d_\Omega \text{ is differentiable in } U_h(\Omega) \setminus \Omega\}.$$

Recall that, if  $\partial\Omega$  is a nonempty compact  $\mathcal{C}^{1,1}$ -hypersurface of  $\mathbb{R}^d$ , then there exists  $h > 0$  such that  $\Omega$  satisfies a uniform ball condition, namely

$$\forall x \in \partial\Omega, \exists d_x \in \mathbb{R}^n \mid \|d_x\|_{\mathbb{R}^d} = 1, B_h(x - hd_x) \subset \Omega \text{ and } B_h(x + hd_x) \subset \mathbb{R}^n \setminus \Omega, \quad (\mathcal{B}_h)$$

where  $B_h(x)$  stands for the open ball of radius  $h$  centered in  $x$ . Furthermore, assuming  $\mathcal{H}^d(\partial\Omega) = 0$ , we have the simpler characterization

$$\text{Reach}(\partial\Omega) = \sup\{h \mid \Omega \text{ satisfies } (\mathcal{B}_h)\}.$$

Conversely, if  $\partial\Omega$  is nonempty and satisfies Condition  $(\mathcal{B}_h)$ , then its reach is larger than  $h$  and the Lebesgue measure of  $\partial\Omega$  in  $\mathbb{R}^n$  is equal to 0. Furthermore,  $\partial\Omega$  is a  $\mathcal{C}^{1,1}$  hypersurface of  $\mathbb{R}^n$ . We refer for instance to Theorems 2.6 and 2.7 in [Dal18].

- For a given oriented  $\mathcal{C}^{1,1}$ -hypersurface  $\partial\Omega$ , we denote by  $\nabla_{\partial\Omega}$  or  $\nabla_\Gamma$  the tangential gradient and by  $\nabla$  the full gradient in  $\mathbb{R}^d$ . When needed, each gradient will be assimilated to a line vector in  $\mathbb{R}^d$ .
- $\overline{\mathbb{N}}$  denotes  $\mathbb{N} \cup \{+\infty\}$ .
- $\mathcal{S}^{d-1}$  denotes the unit sphere of  $\mathbb{R}^d$ .
- $M_d(\mathbb{R})$  denotes the linear space of  $d \times d$  matrices with real entries, endowed with the Euclidean operator norm  $\| \cdot \|$ . Id denotes the identity matrix in  $\mathbb{R}^d$ .
- For a given  $\mathcal{C}^{1,1}$  hypersurface  $\partial\Omega$ , we denote by  $H_{\partial\Omega} : \partial\Omega \rightarrow \mathbb{R}$ , its mean curvature. We refer to Appendix III.A.1 for proper definitions.

### III.1.3 Preliminaries on sets of uniformly positive reach

Given  $r_0 > 0$  and a nonempty compact set  $D \subset \mathbb{R}^d$ , let us introduce the set  $\mathcal{O}_{r_0}$  of admissible shapes whose reach is bounded by  $r_0$ , namely

$$\mathcal{O}_{r_0} = \{\Omega \subset D \mid \Omega \text{ is closed, } \text{Reach}(\partial\Omega) \geq r_0, \Omega \neq \emptyset, \text{ and } \mathcal{H}^d(\partial\Omega) = 0\}. \quad (\text{III.1})$$

The elements of  $\mathcal{O}_{r_0}$  are known to satisfy the following properties.

**Lemma III.1.** *Let  $\Omega \in \mathcal{O}_{r_0}$ . Then*

1.  $\partial\Omega$  is a  $\mathcal{C}^{1,1}$   $(d-1)$ -submanifold. Conversely,

$$\mathcal{O}_{r_0} = \{\Omega \subset D \mid \Omega \text{ is closed, } \text{Reach}(\partial\Omega) \geq r_0, \partial\Omega \text{ is a } (d-1)\text{-submanifold}\}.$$

2. For  $x \in \partial\Omega$ ,  $\nabla b_\Omega(x)$  is the unit outward normal vector.

3. For  $h < r_0$ ,  $\nabla b_\Omega$  is  $\frac{2}{r_0-h}$ -Lipschitz continuous on the tubular neighborhood  $U_h(\partial\Omega)$ .

4. The restriction of  $\nabla b_\Omega$  to  $\partial\Omega$  is  $\frac{1}{r_0}$ -Lipschitz continuous.

5. There exists a constant  $C$  depending only on  $d$ ,  $r_0$ , and  $D$  such that  $\mathcal{H}^{d-1}(\partial\Omega) \leq C$ .

Points 1 and 2 are proved in [DZ11, Theorem 8.2, Chapter 7]. Points 3 and 4 are proved in [Dal18, Theorems 2.7 and 2.8]. The proof of Point 5 is given in Section III.2.1.1.

We will endow the set  $\mathcal{O}_{r_0}$  with a ‘sequential’ topology, by introducing a notion of convergence in this set.

**Definition III.2** ( $R$ -convergence in  $\mathcal{O}_{r_0}$ ). Given  $(\Omega_n)_{n \in \mathbb{N}} \in \mathcal{O}_{r_0}^{\mathbb{N}}$ , we say that  $(\Omega_n)_{n \in \mathbb{N}}$   $R$ -converges to  $\Omega_\infty \in \mathcal{O}_{r_0}$  and we write  $\Omega_n \xrightarrow{R} \Omega_\infty$  if

$$b_{\Omega_n} \rightarrow b_{\Omega_\infty} \begin{cases} \text{in } \mathcal{C}(\overline{D}), \\ \text{in } \mathcal{C}^{1,\alpha}(U_r(\partial\Omega_\infty)), \forall r < r_0, \forall \alpha \in [0, 1), \\ \text{weakly-star in } W^{2,\infty}(U_r(\partial\Omega_\infty)), \forall r < r_0. \end{cases} \quad (\text{III.2})$$

The next result justifies the interest of the class  $\mathcal{O}_{r_0}$  endowed with the  $R$ -convergence for existence issues.

**Proposition III.3.**  $\mathcal{O}_{r_0}$  is sequentially compact for the  $R$ -convergence.

The proof of this proposition can be found in Section III.A.2. Let us end this section by providing several additional properties of the  $R$ -convergence.

**Lemma III.4.** If  $\Omega_n \xrightarrow{R} \Omega_\infty$  then

1.  $\mathcal{H}^{d-1}(\partial\Omega_n)$  converges toward  $\mathcal{H}^{d-1}(\partial\Omega_\infty)$  as  $n \rightarrow +\infty$ .

2.  $\mathcal{H}^d(\Omega_n)$  converges toward  $\mathcal{H}^d(\Omega_\infty)$  as  $n \rightarrow +\infty$ .

3. If all the  $\partial\Omega_n$  belong to the same isotopic class, then  $\partial\Omega_\infty$  also belongs such a class.

The proof of this lemma is given in Section III.2.2.

**Remark III.5.** According to Lemma III.4, we obtain for example that for a given  $\Omega_0 \in \mathcal{O}_{r_0}$ , and  $a \leq b$ ,

$$\{\Omega \in \mathcal{O}_{r_0} \mid a \leq \mathcal{H}^{d-1}(\partial\Omega) \leq b, \partial\Omega \text{ is isotopic to } \partial\Omega_0\}$$

is a sequentially compact set.

### III.1.4 Main results

Let us introduce the general shape functional

$$F_1(\Omega) = \int_{\partial\Omega} j_1(x, \nu(x), H_{\partial\Omega}(x)) d\mu_{\partial\Omega}(x), \quad (\text{III.3})$$

where  $j_1$  is continuous from  $\mathbb{R}^d \times \mathcal{S}^{d-1} \times \mathbb{R}$  to  $\mathbb{R}$  and convex with respect to its last variable. We recall that  $\nu$  and  $H_{\partial\Omega}$  denote respectively the outward pointing normal vector and the mean curvature.

According to Theorem III.3, the set  $\mathcal{O}_{r_0}$  is sequentially compact for the  $R$ -convergence. Therefore, in order to infer the existence of an optimal surface minimizing  $F_1$  over  $\mathcal{O}_{r_0}$  it is enough to prove the lower semicontinuity of functional  $F_1$  (under suitable assumptions on the function  $j_1$ ). This is the main purpose of the following result.

**Theorem III.6** ([Dal18], Theorem 1.3). *Let us assume that  $j_1$  is continuous with respect to all variables and convex with respect to its last one. Then,  $F_1$  is a lower semi-continuous shape functional for the  $R$ -convergence, i.e., for every sequence  $(\Omega_n)_{n \in \mathbb{N}} \in \mathcal{O}_{r_0}^{\mathbb{N}}$  that  $R$ -converges toward  $\Omega_\infty$ , one has*

$$\liminf_{n \rightarrow +\infty} F_1(\Omega_n) \geq F_1(\Omega_\infty). \quad (\text{III.4})$$

As a consequence, the shape optimization problem

$$\inf_{\Omega \in \mathcal{O}_{r_0}} F_1(\Omega)$$

has a solution.

It is notable that, by applying Theorem III.6 both to  $j_1$  and  $-j_1$ , we get the following corollary.

**Corollary III.7.** *If  $j_1$  is continuous and linear in the last variable, then  $F_1$  is a continuous shape functional for the  $R$ -convergence.*

**Remark III.8.** *In the case where  $d = 3$ , it is proved in [Dal18, Theorem 1.3] that Theorem III.6 holds if we replace the mean curvature by the Gaussian one in the definition of  $F_2$ . We do not provide a proof here since most of the difficulties are related to the convergence of a product of weak-star converging sequences and our approach does not change the proof in a significant way.*

Let us now consider two classes of shape optimization problems involving either an elliptic PDE inside  $\Omega$  or an elliptic PDE on the  $\mathcal{C}^{1,1}$  hypersurface  $\partial\Omega$ .

**Problems involving an elliptic PDE on a  $\mathcal{C}^{1,1}$ -hypersurface of  $\mathbb{R}^d$ .** Given  $f \in \mathcal{C}^0(D)$ , we consider the problem of minimizing a shape functional depending on the solution  $v_{\partial\Omega}$  of the equation

$$\Delta_\Gamma v_{\partial\Omega}(x) = f(x) \quad \text{in } \partial\Omega, \quad (\text{III.5})$$

where  $\Delta_{\partial\Omega}$  denotes the Laplace–Beltrami operator on  $\partial\Omega$ . Since we are not considering  $\mathcal{C}^\infty$  manifolds but rather  $\mathcal{C}^{1,1}$  ones, we need to explain how the PDE must be understood. We use here an energy formulation, defining, for a closed and nonempty hypersurface  $\partial\Omega$ , the functional

$$\mathcal{E}_{\partial\Omega} : H_*^1(\partial\Omega) \ni u \mapsto \frac{1}{2} \int_{\partial\Omega} |\nabla_\Gamma u(x)|^2 d\mu_{\partial\Omega} - \int_{\partial\Omega} f(x)u(x) d\mu_{\partial\Omega} \quad (\text{III.6})$$

where  $H_*^1(\partial\Omega)$  denotes the Sobolev space of functions in  $H^1(\partial\Omega)$  with zero mean on  $\partial\Omega$ . We hence define  $v_{\partial\Omega}$  as the unique solution of the minimization problem

$$\min_{u \in H_*^1(\partial\Omega)} \mathcal{E}_{\partial\Omega}(u). \quad (\text{III.7})$$



**Lemma III.9.** *Let  $\Omega \in \mathcal{O}_{r_0}$ . Problem (III.7) has a unique solution  $v_{\partial\Omega}$ . Furthermore, if  $\partial\Omega$  is  $\mathcal{C}^2$  and if  $f \in \mathcal{C}^0(D)$ , then  $v_{\partial\Omega}$  satisfies (III.5) almost everywhere in  $\partial\Omega$ .*

The proof of this result is postponed to Section III.A.3.

Let us introduce the shape functional

$$F_2(\Omega) = \int_{\partial\Omega} j_2(x, \nu(x), v_{\partial\Omega}(x), \nabla_{\Gamma} v_{\partial\Omega}(x)) d\mu_{\partial\Omega}(x),$$

where  $j_2 : \mathbb{R}^d \times \mathcal{S}^{d-1} \times \mathbb{R} \times \mathbb{R}^d \rightarrow \mathbb{R}$  is assumed to be continuous.

**Theorem III.10.** *The shape functional  $F_2$  is lower semi-continuous for the  $R$ -convergence, i.e., for every sequence  $(\Omega_n)_{n \in \mathbb{N}} \in \mathcal{O}_{r_0}^{\mathbb{N}}$  that  $R$ -converges toward  $\Omega_{\infty}$ , one has*

$$\liminf_{n \rightarrow +\infty} F_2(\Omega_n) \geq F_2(\Omega_{\infty}). \quad (\text{III.8})$$

As a consequence, the shape optimization problem

$$\inf_{\Omega \in \mathcal{O}_{r_0}} F_2(\Omega)$$

has a solution.

**Problems involving an elliptic PDE in a domain of  $\mathbb{R}^d$ .** Finally, let us investigate the case of a shape criterion involving the solution of a PDE on a domain of  $\mathbb{R}^d$ . We consider hereafter a Poisson equation with non-homogeneous boundary condition, but we claim that all conclusions can be easily extended to a larger class of elliptic PDEs.

Let  $h \in L^2(D)$ ,  $g \in H^2(D)$ , and define  $u_{\Omega}$  as the solution of

$$\begin{cases} \Delta u_{\Omega} = h & \text{in } \Omega, \\ u_{\Omega} = g & \text{in } \partial\Omega. \end{cases} \quad (\text{III.9})$$

Let us introduce the shape functional  $F_3$  given by

$$F_3(\Omega) = \int_{\partial\Omega} j_3(x, \nu(x), u_{\Omega}(x), \nabla u_{\Omega}(x)) d\mu_{\partial\Omega}(x),$$

where  $j_3 : \mathbb{R}^d \times \mathcal{S}^{d-1} \times \mathbb{R} \times \mathbb{R}^d \rightarrow \mathbb{R}$  is continuous.

**Theorem III.11** ([Dal20], Theorem 2.1). *The shape functional  $F_3$  is lower semi-continuous for the  $R$ -convergence.*

We mention this theorem demonstrated in [Dal20]. Nevertheless, it is notable that by adapting the proof of Theorem III.10, it is possible to obtain a much shorter proof of this theorem. In order not to make this article unnecessarily heavy, we only give the main steps of the proof in Section III.2.5. This example is mentioned both for the sake of completeness, in order to review the existing literature, and also to underline the potential of the approach introduced here, which allows to find more direct proofs of all the known results and to extend them.

In addition, it is interesting to notice that our approach allows to deal with problems involving PDEs both using weak formulations as in (III.9) and also whose solutions are obtained using a minimization principle, as is the case in (III.5). The approach thus seems robust and we believe that it can be easily adapted to general families of problems (for example to a general non-degenerate elliptic PDE).

## III.2 Proofs

### III.2.1 The extruded surface approach

One of the key ideas to prove sequential continuity of functionals involving an integral on the boundary is to approximate such an integral by an integral on a small tubular neighborhood (as done e.g. in [Del00]).

Let us first illustrate the method by proving Point 5 of Theorem III.1.

#### III.2.1.1 Proof of Theorem III.1, Point 5

For  $0 < h < r_0$ , consider

$$\begin{aligned} T : (-h, h) \times \partial\Omega &\rightarrow U_h(\partial\Omega) \\ (t, x) &\mapsto x + t\nabla b_\Omega(x). \end{aligned} \quad (\text{III.10})$$

Since  $T$  is Lipschitz continuous, it is differentiable at almost every  $(t_0, x_0)$ , with

$$d_{(t_0, x_0)}T(s, y) = y + s\nabla b_\Omega(x_0) + t_0 d_{x_0} \nabla b_\Omega(y), \quad \forall (s, y) \in \mathbb{R} \times T_{x_0} \partial\Omega. \quad (\text{III.11})$$

**Remark III.12.** Note that as  $\nabla b_\Omega(x_0)$  is a normal unit vector to  $\partial\Omega$  at  $x_0$ , we can identify the tangent hyperplane  $T_{x_0} \partial\Omega$  with  $\mathbb{R}^{d-1}$  endowed with an Euclidean structure inherited from that of  $\mathbb{R}^d$ . We will use this identification several times in this chapter.

As a result, we can identify  $\mathbb{R} \times T_{x_0} \partial\Omega \ni (s, y) \mapsto y + s\nabla b_\Omega(x_0)$  with an orthogonal matrix. Moreover, up to the choice of a different orientation on  $T_{x_0} \partial\Omega$ , such a matrix belongs to the special orthogonal group  $\text{SO}(n)$ . We use the same coordinate representation to identify  $\mathbb{R} \times T_{x_0} \partial\Omega \ni (s, y) \mapsto d_{x_0} \nabla b_\Omega(y)$  with a  $n \times n$  matrix. By uniform continuity of the determinant around  $\text{SO}(d)$ , there exists  $C_0 > 0$  such that, for every  $M \in \text{SO}(d)$  and every  $l \in M_d(\mathbb{R})$  such that  $\|l\| \leq C_0$ ,

$$\frac{1}{2} \leq \det(M + l) \leq \frac{3}{2}.$$

As  $\nabla b_\Omega$  is  $\frac{2}{r_0}$ -Lipschitz continuous on  $\partial\Omega$ , we have that for almost every  $x_0 \in \partial\Omega$  and every  $t_0 \in \mathbb{R}$ ,  $\|t_0 d_{x_0} \nabla b_\Omega\| \leq \frac{2|t_0|}{r_0}$ .

Let us fix  $h < \min(r_0, r_0 C_0/2)$  (independent of  $\Omega$ ), so that  $\|t_0 d_{x_0} \nabla b_\Omega\| \leq C_0$  for almost every  $x_0 \in \partial\Omega$  and every  $t_0 \in (-h, h)$ . By the change of variable formula we then have

$$\mathcal{H}^{d-1}(\partial\Omega) = \int_{\partial\Omega} d\mu_{\partial\Omega} = \frac{1}{2h} \int_{U_h(\partial\Omega)} \det(d_{T^{-1}(y)}T) dy \leq \frac{3}{4h} \mathcal{H}^d(U_h(D)),$$

whence the conclusion.

#### III.2.1.2 Extruded surface and $R$ -convergence

Let us now illustrate the power of this approach in the case of a  $R$ -converging sequence.

Let  $\Omega_n \xrightarrow{R} \Omega_\infty$ . From now on, we will use the notation  $\Gamma_n := \partial\Omega_n$  for the hypersurfaces.

For  $h < r_0$  and  $n \in \mathbb{N}$ , let us define a parametrization of a neighborhood of  $\Gamma_n$  by

$$\begin{aligned} T_n : (-h, h) \times \Gamma_n &\rightarrow U_h(\Gamma_n) \\ (t, x) &\mapsto x + t\nabla b_{\Omega_n}(x). \end{aligned} \quad (\text{III.12})$$

**Lemma III.13.** *For every  $\varepsilon > 0$ , there exists  $h > 0$  such that for all  $n \in \bar{\mathbb{N}}$ ,*

$$1 - \varepsilon \leq \det(d_{(t_0, x_0)} T_n) \leq 1 + \varepsilon, \quad \text{for a.e. } (t_0, x_0) \in (-h, h) \times \Gamma_n.$$

*Proof.* We follow the same argument as in Section III.2.1.1. Namely, for a given  $\varepsilon > 0$ , there exists  $C_0 > 0$  such that for every  $M \in \text{SO}(d)$  and every  $l \in M_d(\mathbb{R})$  such that  $\|l\| \leq C_0$ ,

$$1 - \varepsilon \leq \det(M + l) \leq 1 + \varepsilon.$$

Let us fix  $h < \min(r_0, r_0 C_0/2)$  (independent of  $n$ ). As  $\nabla b_{\Omega_n}$  is  $\frac{2}{r_0}$ -Lipschitz continuous on  $\Gamma_n$ , we get  $\|t_0 d_{x_0} \nabla b_{\Omega_n}\| \leq C_0$  for almost every  $x_0 \in \Gamma$  and every  $t_0 \in (-h, h)$ . Whence, using Equation (III.11) with the previous estimate, we conclude the proof.  $\square$

**Remark III.14.** *In what follows, we will use the Bachmann–Landau notation  $\mathfrak{o}_{h \rightarrow 0}(1)$  for a function converging to 0 in  $L^\infty$  as  $h$  goes to 0 and for a given  $n$ , large enough. For example, Theorem III.13 implies that*

$$\det(dT_n) = 1 + \mathfrak{o}_{h \rightarrow 0}(1), \quad \text{on } (-h, h) \times \Gamma_n,$$

which means  $\forall \varepsilon > 0, \exists N_0 \in \mathbb{N}, \exists h > 0, \forall n \in \bar{\mathbb{N}}, n \geq N_0$  implies

$$|\det(d_{(t_0, x_0)} T_n) - 1| \leq \varepsilon, \quad \text{for a.e. } (t, x) \in (-h, h) \times \Gamma_n.$$

Let us now introduce the orthogonal projection  $p_n$  onto  $\Gamma_n$ , defined on  $U_h(\Gamma_n)$  for every  $h \in (0, r_0)$ .

**Lemma III.15.** *The following properties hold:*

1.  $p_n$  coincides with the second component of  $T_n^{-1} : U_h(\Gamma_n) \rightarrow (-h, h) \times \Gamma_n$ .
2. For all  $x \in U_h(\Gamma_n)$ ,  $p_n(x) = x - b_{\Omega_n}(x) \nabla b_{\Omega_n}(x)$ .
3.  $p_n$  converges toward  $p_\infty$  in  $L^\infty(U_h(\Gamma_\infty))$ .

*Proof.* Items 1 and 2 are obviously equivalent and are proved in [DZ11, Theorem 7.2, Chapter 7]. Item 3 follows from the  $\mathcal{C}^1$  convergence of  $b_{\Omega_n}$  toward  $b_{\Omega_\infty}$ .  $\square$

We can now state the key equality to relate surface and volume integrals. Apply Lemma III.13 with  $\varepsilon \in (0, 1)$  to select  $h > 0$  such that  $T_n : (-h, h) \times \Gamma_n \rightarrow U_h(\Gamma_n)$  is invertible for every  $n \in \bar{\mathbb{N}}^1$ .

**Lemma III.16.** *For all  $n \in \bar{\mathbb{N}}$ ,  $f \in L^1(\Gamma_n)$ , and  $t \in (0, h)$  we have*

$$\int_{\Gamma_n} f(x) d\mu_{\Gamma_n}(x) = \frac{1}{2t} \int_{U_t(\Gamma_n)} f \circ p_n(y) \det(d_{T_n^{-1}(y)} T_n) dy.$$

*Proof.* Using the change of variable formula (also known as area formula for Lipschitz continuous functions), one gets

$$\int_{-t}^T \int_{\Gamma_n} f(x) d\mu_{\Gamma_n}(x) dt = \int_{U_t(\Gamma_n)} f \circ p_n(y) \det(d_{T_n^{-1}(y)} T_n) dy.$$

$\square$

From now on, we will the  $T_n^{-1}(y)$  inside the determinant to improve the readability.

---

1. It is actually well-known that the domain of invertibility of  $T_n$  contains  $U_{r_0}(\Gamma_n)$ .

**Lemma III.17.** *For every  $h \leq r_0/2$  and  $0 < t < h$ , there exists  $N_0$  such that*

$$\forall n \geq N_0, \quad U_{h-t}(\Gamma_\infty) \subset U_h(\Gamma_n) \subset U_{h+t}(\Gamma_\infty).$$

*Proof.* By uniform convergence of  $b_{\Omega_n}$  toward  $b_{\Omega_\infty}$ , we have that for  $n$  large enough

$$b_{\Omega_\infty}^{-1}((t-h, h-t)) \subset b_{\Omega_n}^{-1}((-h, h)) \subset b_{\Omega_\infty}^{-1}((-h-t, h+t)).$$

□

In order to perform changes of variable in surface integrals, it is convenient to use directly  $p_n$  as a way to map  $\Gamma_\infty$  onto  $\Gamma_n$ . To this aim, we define

$$\begin{aligned} \tau_n : \Gamma_\infty &\rightarrow \Gamma_n \\ x &\mapsto p_n(x). \end{aligned} \quad (\text{III.13})$$

Note that for  $n$  large enough, Theorem III.17 ensure that  $\tau_n$  is well-defined. We also introduce  $\text{Jac}(\tau_n)$  to denote the Jacobian of  $\tau_n$ . Then we have the following lemma.

**Lemma III.18.** *For  $n$  large enough,  $\tau_n : \Gamma_\infty \rightarrow \Gamma_n$  is a diffeomorphism. Besides*

$$\sup_{x \in \Gamma_\infty} |\text{Jac}(\tau_n)(x) - 1| \xrightarrow{n \rightarrow \infty} 0. \quad (\text{III.14})$$

*Proof.* Let  $x \in \Gamma_\infty$ . We take  $v \in T_x \Gamma_\infty$ , and identify it with an element of the tangent hyperplane (see Theorem III.12). As  $v$  is tangent to  $\Gamma_\infty$  at  $x$ , we get

$$\langle v, \nabla b_{\Omega_\infty}(x) \rangle = 0.$$

Using Lemma III.15, Item 2, we get,

$$d_x p_n(v) = v - \langle \nabla b_{\Omega_n}(x), v \rangle \nabla b_{\Omega_n}(x) - b_{\Omega_n}(x) \nabla^2 b_{\Omega_n}(x) v.$$

Let us now fix  $h < \frac{r_0}{3}$ . For  $n$  large enough, thanks to Theorem III.17, we have  $\Gamma_n \subset U_h(\Gamma_\infty)$ . Thus,

$$\begin{aligned} \|d_x p_n(v) - v\| &\leq \|\nabla b_{\Omega_n}(x) - \nabla b_{\Omega_\infty}(x)\| \|v\| + \|b_{\Omega_n}\|_{L^\infty(\Gamma_\infty)} \|\nabla^2 b_{\Omega_n}(x)\| \|v\| \\ &\leq \|v\| \left( \|\nabla b_{\Omega_n} - \nabla b_{\Omega_\infty}\|_{L^\infty(U_{\frac{r_0}{3}}(\Gamma_\infty))} + \|b_{\Omega_n}\|_{L^\infty(\Gamma_\infty)} \|\nabla^2 b_{\Omega_n}(x)\|_{L^\infty(U_{\frac{r_0}{3}}(\Gamma_\infty))} \right). \end{aligned}$$

We recall that both  $\|\nabla b_{\Omega_n} - \nabla b_{\Omega_\infty}\|_{L^\infty(U_{\frac{r_0}{3}}(\Gamma_\infty))}$  and  $\|b_{\Omega_n}\|_{L^\infty(\Gamma_\infty)}$  converge toward zero. Besides, the quantity  $\|\nabla^2 b_{\Omega_n}(x)\|_{L^\infty(U_{\frac{r_0}{3}}(\Gamma_\infty))}$  is uniformly bounded. As a consequence,

$$\sup_{x \in \Gamma_\infty} \sup_{\substack{v \in T_x \Gamma_n \\ \|v\|=1}} \|d_x p_n(v) - v\| \xrightarrow{n \rightarrow \infty} 0. \quad (\text{III.15})$$

Using a similar argument to the one used in Theorem III.13, we take the determinant and obtain Eq. (III.14).

As a result we know that  $\tau_n$  is a local diffeomorphism. It remains to prove that  $\tau_n$  is injective. To this aim, we suppose that  $n$  is large enough to ensure that

$$\|\nabla b_{\Omega_n} - \nabla b_{\Omega_\infty}\|_{L^\infty(U_{\frac{r_0}{3}}(\Gamma_\infty))} < \frac{1}{2}.$$

Let  $x, y \in \Gamma_\infty$  such that  $p_n(x) = p_n(y)$ . If  $x \neq y$ , it implies that there exists  $t \in (-\frac{2r_0}{3}, \frac{2r_0}{3}) \setminus \{0\}$  such that

$$x = y + t \nabla b_{\Omega_n}(p_n(y)) = y + t \nabla b_{\Omega_n}(y)$$

As  $\Omega_\infty \in \mathcal{O}_{r_0}$ , it satisfies the  $r_0$  uniform ball property (see  $(\mathcal{B}_h)$ ). Thus, one has

$$B_{r_0}(y + r_0 \operatorname{sign} t \nabla b_{\Omega_\infty}(y)) \cap \Gamma_\infty = \emptyset.$$

But we have

$$\begin{aligned} |x - y - r_0 \operatorname{sign} t \nabla b_{\Omega_\infty}(y)| &= |t \nabla b_{\Omega_n}(y) - r_0 \operatorname{sign} t \nabla b_{\Omega_\infty}(y)| \\ &\leq \frac{t}{2} + |t - r_0 \operatorname{sign} t| < r_0. \end{aligned}$$

This is a contradiction, hence  $\tau_n$  is injective which implies that it is a diffeomorphism from  $\Gamma_\infty$  to  $\Gamma_n$ .  $\square$

## III.2.2 Proof of Lemma III.4

Suppose that  $(\Omega_n)_{n \in \mathbb{N}} \in \mathcal{O}_{r_0}^{\mathbb{N}}$   $R$ -converges toward  $\Omega_\infty \in \mathcal{O}_{r_0}$ .

### III.2.2.1 Proof of Point 1

For  $h < r_0$  and using Lemma III.16, we have

$$\mathcal{H}^{d-1}(\Gamma_n) = \int_{\Gamma_n} d\mu_{\Gamma_n}(x) = \frac{1}{2h} \int_{U_h(\Gamma_n)} \det(dT_n) dy.$$

By Lemma III.17, moreover,

$$\mathcal{H}^{d-1}(\Gamma_n) = \frac{1}{2h} \int_{U_{h-t}(\Gamma_\infty)} \det(dT_n) dy + \frac{1}{2h} \int_{U_h(\Gamma_n) \setminus U_{h-t}(\Gamma_\infty)} \det(dT_n) dy$$

for  $t \in (0, h)$  and  $n$  large enough. Let us compare the first term in the right-hand side with

$$\mathcal{H}^{d-1}(\Gamma_\infty) = \frac{1}{2(h-t)} \int_{U_{h-t}(\Gamma_\infty)} \det(dT_\infty) dy.$$

Using Lemma III.13,  $\det(dT_\infty) = \det(dT_n) + o_{h \rightarrow 0}(1)$  on  $(-h, h) \times \Gamma_\infty$ . Besides,  $\frac{1}{2h} - \frac{1}{2(h-t)} = O\left(\frac{t}{h}\right)$ . Hence,

$$\frac{1}{2h} \int_{U_{h-t}(\Gamma_\infty)} \det(dT_n) dy = \mathcal{H}^{d-1}(\Gamma_\infty) + o_{h \rightarrow 0}(1) + O\left(\frac{t}{h}\right).$$

On the other hand, using again the relation  $\det(dT_\infty) = \det(dT_n) + o_{h \rightarrow 0}(1)$ ,

$$\begin{aligned} \frac{1}{2h} \int_{U_h(\Gamma_n) \setminus U_{h-t}(\Gamma_\infty)} \det(dT_n) dy &\leq \frac{1}{2h} \int_{U_{h+t}(\Gamma_\infty) \setminus U_{h-t}(\Gamma_\infty)} \det(dT_n) dy \\ &= \frac{1}{2h} \left( \int_{U_{h+t}(\Gamma_\infty)} \det(dT_n) dy - \int_{U_{h-t}(\Gamma_\infty)} \det(dT_n) dy \right) \\ &= \frac{1}{2h} (2(h+t)\mathcal{H}^{d-1}(\Gamma_\infty) - 2(h-t)\mathcal{H}^{d-1}(\Gamma_\infty) + o_{h \rightarrow 0}(h)) \\ &= \frac{2t}{h} \mathcal{H}^{d-1}(\Gamma_\infty) + o_{h \rightarrow 0}(1) = O\left(\frac{t}{h}\right) + o_{h \rightarrow 0}(1). \end{aligned}$$

By taking  $h$  arbitrary small while  $t = h^2$ , we prove that  $\mathcal{H}^{d-1}(\Gamma_n) \rightarrow \mathcal{H}^{d-1}(\Gamma_\infty)$ .

### III.2.2.2 Proof of Point 2

Using the uniform convergence of  $b_{\Omega_n}$  to  $b_{\Omega_\infty}$ , we deduce that for every  $\varepsilon > 0$  there exists  $N_0 \in \mathbb{N}$  such that

$$b_{\Omega_\infty}^{-1}((-\infty, -\varepsilon]) \subset b_{\Omega_n}^{-1}((-\infty, 0)) \subset b_{\Omega_\infty}^{-1}((-\infty, \varepsilon)), \quad \forall n \geq N_0.$$

Hence, we get

$$\mathcal{H}^d(b_{\Omega_\infty} \leq -\varepsilon) \leq \mathcal{H}^d(\Omega_n) \leq \mathcal{H}^d(b_{\Omega_\infty} < \varepsilon).$$

By inner regularity of  $\mathcal{H}^d$ ,  $\mathcal{H}^d(b_{\Omega_\infty} \leq -\varepsilon) \xrightarrow{\varepsilon \rightarrow 0} \mathcal{H}^d(b_{\Omega_\infty} < 0) = \mathcal{H}^d(\Omega_\infty)$ . Similarly, by outer regularity  $\mathcal{H}^d(b_{\Omega_\infty} < \varepsilon) \xrightarrow{\varepsilon \rightarrow 0} \mathcal{H}^d(b_{\Omega_\infty} \leq 0) = \mathcal{H}^d(\Omega_\infty)$ , where we used that  $\Omega_\infty$  belongs to  $\mathcal{O}_{r_0}$ .

### III.2.2.3 Proof of Point 3

We want to prove that  $\Gamma_n$  is isotopic to  $\Gamma_\infty$  for  $n$  large enough. We consider

$$\begin{aligned} \varphi_n(t, x) : [0, 1] \times \Gamma_\infty &\rightarrow \mathbb{R}^3 \\ (t, x) &\mapsto x + t(p_n(x) - x). \end{aligned}$$

According to Theorem III.18,  $\varphi_n(1, \cdot) = \tau_n$  is a diffeomorphism from  $\Gamma_\infty$  onto  $\Gamma_n$ . Besides, following the proof of Theorem III.18, we easily get that for  $t \in (0, 1)$ ,  $\varphi_n(t, \cdot)$  is a diffeomorphism onto its image.

## III.2.3 Proof of Theorem III.6

Suppose that  $(\Omega_n)_{n \in \mathbb{N}} \in \mathcal{O}_{r_0}^{\mathbb{N}}$   $R$ -converges toward  $\Omega_\infty \in \mathcal{O}_{r_0}$ . Let  $0 < t < h$  small enough (to be fixed later) and  $n$  large enough.

We recall that the unit normal vector to  $\Gamma_n$  is given by  $\nabla b_{\Omega_n}$  (see Theorem III.15). Then, according to Theorem III.16,

$$\begin{aligned} F_1(\Omega_n) &= \int_{\Gamma_n} j_1(x, \nabla b_{\Omega_n}(x), H_{\Gamma_n}(p_n(y))) d\mu_{\Gamma_n}(x) \\ &= \frac{1}{2h} \int_{U_h(\Gamma_n)} j_1(p_n(y), \nabla b_{\Omega_n}(p_n(y)), H_{\Gamma_n}(p_n(y))) \det(d_{T_n^{-1}(y)} T_n) dy. \end{aligned}$$

Using Lemma III.17, moreover,

$$F_1(\Omega_n) = \frac{1}{2h} \int_{U_{h-t}(\Gamma_\infty)} j_1(p_n(y), \nabla b_{\Omega_n}(p_n(y)), H_{\Gamma_n}(p_n(y))) \det(dT_n) dy \quad (\text{III.16})$$

$$+ \frac{1}{2h} \int_{U_h(\Gamma_n) \setminus U_{h-t}(\Gamma_\infty)} j_1(p_n(y), \nabla b_{\Omega_n}(p_n(y)), H_{\Gamma_n}(p_n(y))) \det(dT_n) dy. \quad (\text{III.17})$$

The key idea is to prove that all arguments of  $j_1$  in the first term convergence toward their analogues for  $n = \infty$  and to ensure that the second term is small for  $t$  small.

Let us start with comparing the first term in the right-hand side with  $F_1(\Omega_\infty)$ . Notice that

$$\begin{aligned} & \frac{1}{2h} \int_{U_{h-t}(\Gamma_\infty)} j_1(p_n(y), \nabla b_{\Omega_n}(p_n(y)), H_{\Gamma_n}(p_n(y))) \det(dT_n) dy \\ &= \frac{1}{2(h-t)} \int_{U_{h-t}(\Gamma_\infty)} \frac{2(h-t)}{2h} \frac{\det(dT_n)}{\det(dT_\infty)} j_1(p_n(y), \nabla b_{\Omega_n}(p_n(y)), H_{\Gamma_n}(p_n(y))) \det(dT_\infty) dy. \end{aligned}$$

By Lemma III.13, we have

$$\left\| \frac{2(h-t)}{2h} \frac{\det(dT_n)}{\det(dT_\infty)} - 1 \right\|_{L^\infty(U_h(\Gamma_\infty))} = o_{h \rightarrow 0}(1) + O\left(\frac{t}{h}\right). \quad (\text{III.18})$$

Let us now investigate the mean curvature term. Note that this term is slightly technical to handle for two reasons:

- the mean curvature  $H_{\Gamma_n}$  is defined as the trace of the shape operator, which is itself defined as the differential of the restriction to the hypersurface of  $\nabla b_{\Omega_n}$  (see III.A.1);
- the Hessian of  $b_{\Omega_n}$  converges only in a weak sense.

We will use the following lemma, which is obtained thanks to the chain rule.

**Lemma III.19** (Theorem 4.4 of [Del00]). *Let  $h < r_0$  and  $n \in \bar{\mathbb{N}}$ . If  $\nabla^2 b_{\Omega_n}(x)$  exists for  $x \in U_h(\Gamma_\infty)$ , then  $\nabla^2 b_{\Omega_n}(p_n(x))$  exist and*

$$\nabla^2 b_{\Omega_n}(p_n(x)) = \nabla^2 b_{\Omega_n}(x) [\text{Id} - b_{\Omega_n}(x) \nabla^2 b_{\Omega_n}(x)]^{-1}.$$

*Besides, one has that  $\nabla^2 b_{\Omega_n}(\tau_n^{-1}(p_n(x)))$  exists as well.*

Notice that the last part of the statement is not explicitly contained in [Del00] but can be obtained by straightforwardly adapting the proof of its Theorem 4.4.

As  $\nabla^2 b_{\Omega_n}$  is uniformly bounded on a neighborhood of  $\Gamma_\infty$  and that  $b_{\Omega_n}(x) \leq h$  for  $x \in U_h(\Gamma_n)$ , there exists  $C > 0$  such that

$$\text{ess sup}_{x \in U_h(\Gamma_n)} \left\| [\text{Id} - b_{\Omega_n}(x) \nabla^2 b_{\Omega_n}(x)]^{-1} - \text{Id} \right\| \leq Ch,$$

for  $h$  small. As a consequence, using Theorem III.27 (given in the appendix), one has

$$H_{\Gamma_n}(p_n(x)) = \text{Tr} \nabla^2 b_{\Omega_n}(p_n(x)) = \text{Tr} \nabla^2 b_{\Omega_n}(x) + O(h).$$

Note also that  $H_{\Gamma_n} \leq \frac{1}{r_0}$  on  $\Gamma_n$ . We can use the uniform continuity of  $j_1$  on a compact set to

ensure that for  $n$  large enough and  $n = \infty$ ,

$$\begin{aligned} & \frac{1}{2(h-t)} \int_{U_{h-t}(\Gamma_\infty)} j_1(p_n(y), \nabla b_{\Omega_n}(p_n(y)), H_{\Gamma_n}(p_n(y))) \det(dT_\infty) dy \\ &= \frac{1}{2(h-t)} \int_{U_{h-t}(\Gamma_\infty)} j_1(p_n(y), \nabla b_{\Omega_n}(p_n(y)), \text{Tr } \nabla^2 b_{\Omega_n}(y)) \det(dT_\infty) dy + O(h). \end{aligned} \quad (\text{III.19})$$

The next step is to pass to the limit within the integral. Note that, by definition of  $R$ -convergence,

$$\left\{ \begin{array}{ll} p_n \xrightarrow[n \rightarrow \infty]{} p_\infty & \text{strongly in } L^\infty(U_{\frac{r_0}{2}}(\Gamma_\infty)), \\ \nabla b_{\Omega_n} \circ p_n \xrightarrow[n \rightarrow \infty]{} \nabla b_{\Omega_\infty} \circ p_\infty & \text{strongly in } L^\infty(U_{\frac{r_0}{2}}(\Gamma_\infty)), \\ \text{Tr } \nabla^2 b_{\Omega_n} \xrightarrow[n \rightarrow \infty]{} \text{Tr } \nabla^2 b_{\Omega_\infty} & \text{weak star in } L^\infty(U_{\frac{r_0}{2}}(\Gamma_\infty)). \end{array} \right.$$

Thus, using for example [Ber74], we have

$$\begin{aligned} & \liminf_{n \rightarrow \infty} \frac{1}{2(h-t)} \int_{U_{h-t}(\Gamma_\infty)} j_1(p_n(y), \nabla b_{\Omega_n}(p_n(y)), \text{Tr } \nabla^2 b_{\Omega_n}(y)) \det(dT_\infty) dy \\ & \geq \frac{1}{2(h-t)} \int_{U_{h-t}(\Gamma_\infty)} j_1(p_\infty(y), \nabla b_{\Omega_\infty}(p_\infty(y)), \text{Tr } \nabla^2 b_{\Omega_\infty}(y)) \det(dT_\infty) dy \\ &= \frac{1}{2(h-t)} \int_{U_{h-t}(\Gamma_\infty)} j_1(p_\infty(y), \nabla b_{\Omega_\infty}(p_\infty(y)), \text{Tr } \nabla^2 b_{\Omega_\infty}(p_\infty(y))) \det(dT_\infty) dy + O(h) \\ &= F_1(\Omega_\infty) + O(h). \end{aligned} \quad (\text{III.20})$$

In order to conclude, let us check that the term in line (III.17) is small. Since  $j_1$  is continuous on a compact set, it admits a minimum  $m_0 \in \mathbb{R}$ . Let  $m_1 = \min(0, m_0) \leq 0$ . Then,

$$\begin{aligned} & \frac{1}{2h} \int_{U_h(\Gamma_n) \setminus U_{h-t}(\Gamma_\infty)} j_1(p_n(y), \nabla b_{\Omega_n}(p_n(y)), H_{\Gamma_n}(p_n(y))) \det(dT_n) dy \\ & \geq \frac{1}{2h} \int_{U_h(\Gamma_n) \setminus U_{h-t}(\Gamma_\infty)} m_1 \det(dT_n) dy \\ & \geq \frac{1}{2h} \int_{U_{h+t}(\Gamma_\infty) \setminus U_{h-t}(\Gamma_\infty)} m_1 \frac{\det(dT_n)}{\det(dT_\infty)} \det(dT_\infty) dy. \end{aligned}$$

Using

$$\left\| \frac{\det(dT_n)}{\det(dT_\infty)} - 1 \right\|_{L^\infty(U_{2h}(\Gamma_\infty))} = o_{h \rightarrow 0}(1)$$

and

$$\int_{U_{h \pm t}(\Gamma_\infty)} m_1 \det(dT_\infty) dy = 2(h \pm t) m_1 \mathcal{H}^{d-1}(\Gamma_\infty),$$

we get

$$\begin{aligned} & \frac{1}{2h} \int_{U_h(\Gamma_n) \setminus U_{h-t}(\Gamma_\infty)} j_1(p_n(y), \nabla b_{\Omega_n}(p_n(y)), H_{\Gamma_n}(p_n(y))) \det(dT_n) dy \\ & \geq m_1 \mathcal{H}^{d-1}(\Gamma_\infty) \left( o_{h \rightarrow 0}(1) + O\left(\frac{t}{h}\right) \right). \end{aligned} \quad (\text{III.21})$$



Finally, combining Equations (III.19)–(III.21), we obtain

$$\liminf_{n \rightarrow +\infty} F_1(\Omega_n) \geq (F_1(\Omega_\infty) + O(h)) \left(1 + O\left(\frac{t}{h}\right)\right) + m_1 \mathcal{H}^{d-1}(\Gamma_\infty) \left(o_{h \rightarrow 0}(1) + O\left(\frac{t}{h}\right)\right).$$

Hence, taking  $h \rightarrow 0$  while ensuring  $t = o(h)$  gives

$$\liminf_{n \rightarrow +\infty} F_1(\Omega_n) \geq F_1(\Omega_\infty),$$

and finishes the proof.

### III.2.4 Proof of Theorem III.10

Let  $(\Omega_n)_{n \in \mathbb{N}}$  denote a sequence that  $R$ -converges to  $\Omega_\infty$ , and let  $v_n$  denote the unique solution  $v_{\Gamma_n}$  to Problem (III.7) for  $\Omega = \Omega_n$ . The difficult part here is that  $v_{\Gamma_n}$  is not defined on  $\Gamma_\infty$ . Our main tool will be  $\tau_n$ , the restriction to  $\Gamma_\infty$  of the orthogonal projection  $p_n$  on  $\Gamma_n$ . Those objects were introduced in Section III.2.1.2 and we proved that  $\tau_n$  is a diffeomorphism between  $\Gamma_\infty$  and  $\Gamma_n$  in Theorem III.18.

We also have to be careful when we transport the tangential gradient of a function. In order to relate the tangential gradient and the ambient gradient, we establish the following pointwise estimate.

**Lemma III.20.** *Let  $n \in \mathbb{N}$  and  $f_n \in H^1(\Gamma_n)$ . Then  $f_n \circ \tau_n \in H^1(\Gamma_\infty)$  and, for almost every  $x \in \Gamma_\infty$ ,*

$$\nabla_{\Gamma_\infty}(f_n \circ \tau_n)(x) = \nabla_{\Gamma_n} f_n(\tau_n(x))(\text{Id} + C_n(x)), \quad (\text{III.22})$$

where tangential gradients are understood as  $d$ -dimensional line vectors and

$$\begin{aligned} C_n(x) &= (\nabla b_{\Omega_n}(x)^\top \nabla b_{\Omega_n}(x) - \text{Id}) \nabla b_{\Omega_\infty}(x)^\top \nabla b_{\Omega_\infty}(x) \\ &\quad + b_{\Omega_n}(x) \nabla^2 b_{\Omega_n}(x) (\nabla b_{\Omega_\infty}(x)^\top \nabla b_{\Omega_\infty}(x) - \text{Id}). \end{aligned}$$

Besides,  $C_n$  converges toward zero in the  $L^\infty$  norm:

$$\text{ess sup}_{x \in \Gamma_\infty} \|C_n(x)\| \xrightarrow{n \rightarrow \infty} 0. \quad (\text{III.23})$$

*Proof.* First notice that

$$\nabla_{\Gamma_\infty}(f_n \circ p_n)(x) = \nabla(f_n \circ p_n \circ p_\infty)(x)$$

for almost every  $x \in \partial\Omega_\infty$ , since the directional derivative of  $f_n \circ p_n \circ p_\infty$  at the point  $x$  in the direction  $\nabla b_{\Omega_\infty}(x)$  is zero. By Lemma III.19,  $\nabla^2 b_{\Omega_n}(x)$  is well-defined for almost every  $x$  in  $\Gamma_\infty$ . By Lemma III.15 and the chain rule we obtain, almost everywhere on  $\Gamma_\infty$ ,

$$\begin{aligned} \nabla(f_n \circ p_n \circ p_\infty)(x) &= ((\nabla f_n) \circ p_n)(\text{Id} - \nabla b_{\Omega_n}^\top \nabla b_{\Omega_n} - b_{\Omega_n} \nabla^2 b_{\Omega_n})(\text{Id} - \nabla b_{\Omega_\infty}^\top \nabla b_{\Omega_\infty} - b_{\Omega_\infty} \nabla^2 b_{\Omega_\infty}) \\ &= ((\nabla_{\Gamma_n} f_n) \circ \tau_n)(\text{Id} - (\text{Id} - \nabla b_{\Omega_n}^\top \nabla b_{\Omega_n}) \nabla b_{\Omega_\infty}^\top \nabla b_{\Omega_\infty} - b_{\Omega_n} \nabla^2 b_{\Omega_n} (\text{Id} - \nabla b_{\Omega_\infty}^\top \nabla b_{\Omega_\infty})), \end{aligned}$$

where we used that  $\nabla f_n = \nabla_{\Gamma_n} f_n$ ,  $p_n = \tau_n$ , and  $b_{\Omega_\infty} = 0$  on  $\Gamma_\infty$ . This shows Eq. (III.22).

Let us now bound the  $L^\infty$  norm of  $C_n$ . There exists  $C > 0$  such that, for every  $n$  satisfying  $\Gamma_\infty \subset U_{\frac{r_0}{2}}(\Gamma_n)$ ,

$$\text{ess sup}_{x \in \Gamma_\infty} \|\nabla^2 b_{\Omega_n}(x) (\nabla b_{\Omega_\infty}^\top(x) \nabla b_{\Omega_\infty}(x) - \text{Id})\| \leq C.$$

Besides,  $\|b_{\Omega_n}\|_{L^\infty(\Gamma_\infty)}$  converges to zero. Finally, using the uniform convergence of  $\nabla b_{\Omega_n}$  toward  $\nabla b_{\Omega_\infty}$ , we get

$$\nabla b_{\Omega_n}^\top \nabla b_{\Omega_n} \nabla b_{\Omega_\infty}^\top \nabla b_{\Omega_\infty} \xrightarrow[n \rightarrow \infty]{L^\infty(\Gamma_\infty)} \nabla b_{\Omega_\infty}^\top (\nabla b_{\Omega_\infty} \nabla b_{\Omega_\infty}^\top) \nabla b_{\Omega_\infty} = \nabla b_{\Omega_\infty}^\top \nabla b_{\Omega_\infty}.$$

This concludes the proof of Eq. (III.23).  $\square$

From the solution  $v_n$  in  $H_*^1(\Gamma_n)$ , we introduce the function  $w_n$  defined on  $\Gamma_\infty$  by

$$w_n = v_n \circ \tau_n - \frac{1}{\mathcal{H}^{d-1}(\Gamma_\infty)} \int_{\Gamma_\infty} v_n \circ \tau_n d\mu_{\Gamma_\infty}. \quad (\text{III.24})$$

Note that, defined as such,  $w_n$  belongs to  $H_*^1(\Gamma_\infty)$ .

**Step 1: convergence of  $(w_n)_{n \in \mathbb{N}}$ .** Let us start by considering the sequence of energies  $(\mathcal{E}_{\Gamma_n}(v_n))_{n \in \mathbb{N}}$ . This sequence is upper bounded by 0, since  $\mathcal{E}_{\Gamma_n}(v_n) \leq \mathcal{E}_{\Gamma_n}(0) = 0$  for every  $n$ . By using the uniform Poincaré inequality stated in Theorem III.29 combined with the Cauchy–Schwarz inequality, we get that  $(\int_{\Gamma_n} |v_n|^2 d\mu_{\Gamma_n}(x))_{n \in \mathbb{N}}$  is bounded. We now compute

$$\begin{aligned} \frac{1}{\mathcal{H}^{d-1}(\Gamma_\infty)} \int_{\Gamma_\infty} v_n \circ \tau_n d\mu_{\Gamma_\infty} &= \frac{1}{\mathcal{H}^{d-1}(\Gamma_\infty)} \int_{\Gamma_n} v_n \text{Jac}(\tau_n) d\mu_{\Gamma_n} \\ &= \left( \frac{\sqrt{\mathcal{H}^{d-1}(\Gamma_n)}}{\mathcal{H}^{d-1}(\Gamma_\infty)} \|v_n\|_{L^2(\Gamma_n)} \right) o_{n \rightarrow \infty}(1), \end{aligned}$$

where we used Lemma III.18, the Cauchy–Schwarz inequality, and the fact that  $v_n$  has zero average on  $\Gamma_n$ . Hence, we infer that  $w_n = v_n \circ \tau_n + o_{n \rightarrow \infty}(1)$ . Besides, by performing a change of variable and by using Theorems III.18 and III.20, we get

$$\begin{aligned} &\int_{\Gamma_n} \left( \frac{1}{2} |\nabla_{\Gamma_n} v_n(y)|^2 - f(y)v_n(y) \right) d\mu_{\Gamma_n}(y) \\ &= \int_{\Gamma_\infty} \left( \frac{1}{2} |\nabla_{\Gamma_n} v_n(\tau_n(y))|^2 - f(\tau_n(y))(v_n \circ \tau_n)(y) \right) \text{Jac}(\tau_n)^{-1} d\mu_{\Gamma_\infty}(y) \\ &= \int_{\Gamma_\infty} \left( \frac{1}{2} |\nabla_{\Gamma_\infty} w_n(y)|^2 - f(\tau_n(y))w_n(y) \right) d\mu_{\Gamma_\infty}(y) + o_{n \rightarrow \infty}(1), \end{aligned}$$

where we used that  $\nabla_{\Gamma_\infty} w_n = \nabla_{\Gamma_\infty} (v_n \circ \tau_n)$  by definition of  $w_n$ . Using Theorem III.29 and again the Cauchy–Schwarz inequality, we successively infer that the sequences  $(\int_{\Gamma_\infty} |w_n|^2 d\mu_{\Gamma_\infty}(x))_{n \in \mathbb{N}}$  and  $(\int_{\Gamma_\infty} |\nabla_{\Gamma_\infty} w_n|^2 d\mu_{\Gamma_\infty}(x))_{n \in \mathbb{N}}$  are bounded. By using Theorem III.28, the sequence  $(w_n)_{n \in \mathbb{N}}$  converges up to a subsequence toward  $w_\infty \in H_*^1(\Gamma_\infty)$ , weakly in  $H^1(\Gamma_\infty)$  and strongly in  $L^2(\Gamma_\infty)$ . Up to extracting a subsequence, we get

$$\begin{aligned} \int_{\Gamma_\infty} |\nabla_{\Gamma_\infty} w_\infty(x)|^2 d\mu_{\Gamma_\infty} &\leq \liminf_{n \rightarrow +\infty} \int_{\Gamma_\infty} |\nabla_{\Gamma_\infty} w_n(x)|^2 d\mu_{\Gamma_\infty}, \\ \lim_{n \rightarrow \infty} \int_{\Gamma_\infty} w_n(x) f(\tau_n(x)) d\mu_{\Gamma_\infty} &= \int_{\Gamma_\infty} w_\infty(x) f(x) d\mu_{\Gamma_\infty}. \end{aligned}$$

As a consequence,

$$\mathcal{E}_{\Gamma_\infty}(w_\infty) \leq \liminf_{n \rightarrow +\infty} \mathcal{E}_{\Gamma_n}(v_n).$$

**Step 2: Minimality of  $w_\infty$ .** Let  $u \in H_*^1(\Gamma_\infty)$  be given and define  $z_n$  in  $H_*^1(\Gamma_n)$  by

$$z_n = u \circ \tau_n^{-1} - \frac{1}{\mathcal{H}^{d-1}(\Gamma_n)} \int_{\Gamma_n} u \circ \tau_n^{-1} d\mu_{\Gamma_n}. \quad (\text{III.25})$$

Let  $n \in \mathbb{N}$ . By minimality, one has

$$\mathcal{E}_{\Gamma_n}(v_n) \leq \mathcal{E}_{\Gamma_n}(z_n).$$

By mimicking the arguments and computations of the first step, we easily get that

$$\mathcal{E}_{\Gamma_n}(z_n) = \mathcal{E}_{\Gamma_\infty}(u) + o_{n \rightarrow \infty}(1), \quad (\text{III.26})$$

yielding at the end  $\mathcal{E}_{\Gamma_\infty}(w_\infty) \leq \mathcal{E}_{\Gamma_\infty}(u)$ . We infer that  $w_\infty$  is the unique solution to the variational problem (III.7). Since the reasoning above holds for any closure point of  $(w_n)_{n \in \mathbb{N}}$ , it follows that the whole sequence  $(w_n)_{n \in \mathbb{N}}$  converges toward  $w_\infty$ , weakly in  $H^1(\Gamma_\infty)$  and strongly in  $L^2(\Gamma_\infty)$ . Finally, using  $u = w_\infty$  in (III.26), we obtain that

$$\mathcal{E}_{\Gamma_\infty}(w_\infty) = \liminf_{n \rightarrow +\infty} \mathcal{E}_{\Gamma_n}(v_n).$$

In particular  $(\|w_n\|_{H^1(\Gamma_\infty)}^2)_{n \in \mathbb{N}}$  converges toward  $\|w_\infty\|_{H^1(\Gamma_\infty)}^2$  which implies the strong convergence of  $w_n$  in  $H^1(\Gamma_\infty)$ .

**Step 3: lower semi-continuity of  $F_2$ .** Let us use the same notations as previously. Using a change of variable, we get

$$\begin{aligned} F_2(\Omega_n) &= \int_{\Gamma_n} j_2(x, \nabla b_{\Omega_n}(x), v_n(x), \nabla_{\Gamma_n} v_n(x)) d\mu_{\Gamma_n}(x) \\ &= \int_{\Gamma_\infty} j_2(\tau_n(x), \nabla b_{\Omega_n}(\tau_n(x)), v_n(\tau_n(x)), \nabla_{\Gamma_n} v_n \circ \tau_n(x)) \text{Jac}(\tau_n)^{-1} d\mu_{\Gamma_\infty}(x). \end{aligned}$$

Besides, according to the results above and Theorem III.15, the following convergences hold

$$\left\{ \begin{array}{lll} \text{Jac}(\tau_n)^{-1} & \xrightarrow[n \rightarrow \infty]{} 1 & \text{strongly in } L^\infty(\Gamma_\infty) \\ \tau_n & \xrightarrow[n \rightarrow \infty]{} \text{Id}|_{\Gamma_\infty} & \text{strongly in } L^\infty(\Gamma_\infty) \\ \nabla b_{\Omega_n} \circ \tau_n & \xrightarrow[n \rightarrow \infty]{} \nabla b_{\Omega_\infty} & \text{strongly in } L^\infty(\Gamma_\infty) \\ v_n \circ \tau_n & \xrightarrow[n \rightarrow \infty]{} w_\infty & \text{strongly in } L^2(\Gamma_\infty) \\ \nabla_{\Gamma_n} v_n \circ \tau_n & \xrightarrow[n \rightarrow \infty]{} \nabla_{\Gamma_\infty} w_\infty & \text{strongly in } L^2(\Gamma_\infty), \end{array} \right.$$

where  $w_\infty$  is the unique solution to the variational problem (III.7).

By applying [Ber74, Theorem 1], one has

$$\liminf_{n \rightarrow +\infty} F_2(\Omega_n) \geq F_2(\Omega_\infty).$$

This is the desired conclusion.

### III.2.5 Main steps in the proof of Theorem III.11

First note that  $u_\Omega - g$  solves Eq. (III.9) with source term  $h - \Delta g$  and Dirichlet boundary condition. As a consequence, we can reduce our study to the case of homonegeous Dirichlet condition (i.e.,  $u_\Omega = 0$  on  $\Gamma$ ).

The method relies on a uniform extension property proved by Chenaïs in [Che75] for surfaces satisfying an  $\varepsilon$ -cone condition, which is weaker than the uniform ball condition.

**Lemma III.21** ([Che75, Theorem II.1]). *There exists a positive constant  $C$  (depending only on  $r_0$  and  $D$ ) such that for every  $\Omega \in \mathcal{O}_{r_0}$  there exists an extension operator  $E_\Omega \in \mathcal{L}(H^2(\Omega), H^2(D))$  satisfying*

$$E_\Omega(u)|_\Omega = u, \quad \|E_\Omega\|_{\mathcal{L}(H^2(\Omega), H^2(D))} \leq C. \quad (\text{III.27})$$

We will use this lemma to extend the solution of the PDEs to the whole box  $D$ . The next step is to find a uniform  $H^2$  estimate of the solutions. In our case such an estimate was proved by Dalphin who extended a result for domains with  $\mathcal{C}^2$  boundary obtained by Grisvard in [Gri85].

**Lemma III.22** ([Dal20, Proposition 3.1]). *There exists  $C > 0$  (depending only on  $r_0$  and  $D$ ) such that for every  $\Omega \in \mathcal{O}_{r_0}$  and  $f \in H^2(\Omega) \cap H_0^1(\Omega)$ , we have*

$$\|f\|_{H^2(\Omega)} \leq C \|\Delta f\|_{L^2(\Omega)}. \quad (\text{III.28})$$

As a consequence, we have a uniform  $H^2(D)$  estimate on the extension of the solution  $u_\Omega$ , namely,

$$\|E_\Omega(u_\Omega)\|_{H^2(D)} \leq C \|h\|_{L^2(D)}, \quad \forall \Omega \in \mathcal{O}_{r_0}. \quad (\text{III.29})$$

Let us now consider  $\Omega_n \xrightarrow{R} \Omega_\infty$ . Using Eq. (III.29), we get that  $(E_{\Omega_n}(u_{\Omega_n}))_{n \in \mathbb{N}}$  is uniformly bounded in  $H^2(D)$ . Up to extracting a subsequence, we can assume that

$$E_{\Omega_n}(u_{\Omega_n}) \xrightarrow{n \rightarrow \infty} u^* \begin{cases} \text{weakly in } H^2(D) \\ \text{strongly in } H^1(D). \end{cases} \quad (\text{III.30})$$

The next step is to prove that the restriction to  $\Omega_\infty$  of  $u^*$  is  $u_{\Omega_\infty}$ .

To this aim, let us consider an arbitrary compact set  $K$  contained in the interior of  $\Omega_\infty$  and a  $\mathcal{C}^\infty$  function  $\varphi$  with compact support included in  $K$ . For  $n$  large enough,  $K$  is contained in the interior of  $\Omega_n$  (see Theorem III.17), and, therefore, one has  $\varphi \in H_0^1(\Omega_n)$  for such integers  $n$ . Using the variational formulation of the PDE (III.9), we get

$$\int_D \langle \nabla E_{\Omega_n}(u_{\Omega_n}), \nabla \varphi \rangle - f \varphi = 0. \quad (\text{III.31})$$

Using the density of  $\mathcal{C}^\infty$  functions with compact support in  $H_0^1(\Omega_\infty)$  and passing to the limit yields that  $u^*|_{\Omega_\infty} = u_{\Omega_\infty}$ .

**Remark III.23.** *In order to replace Dirichlet boundary conditions by Neumann's ones, one can follow similar steps as those leading to Equation (III.30). Then, by considering the variational formulation with  $\varphi \in \mathcal{C}^\infty(D)$  and passing to the limit in*

$$\int_{\Gamma_n} g \partial_\nu \varphi \rightarrow \int_{\Gamma_\infty} g \partial_\nu \varphi,$$

(consequence of Theorem III.7 if  $g \in \mathcal{C}^0(D)$ ) one gets that  $u^*|_{\Omega_\infty} = u_{\Omega_\infty}$ .

The last step is to relate  $F_3(\Omega_n)$  and  $F_3(\Omega_\infty)$ . Since the involved functions belong to Sobolev spaces and since one aims at comparing surface integrals with tubular ones, we need a suitable uniform trace result.

**Lemma III.24.** *There exists  $C$  such that for every  $h < \frac{r_0}{2}$ , every  $n \in \bar{\mathbb{N}}$  and every  $f \in H^1(U_{\frac{r_0}{2}}(\Gamma_n))$ ,*

$$\|f - \tilde{f} \circ p_n\|_{L^2(U_h(\Gamma_n))} \leq Ch\|f\|_{H^1(U_{\frac{r_0}{2}}(\Gamma_n))}, \quad (\text{III.32})$$

where  $\tilde{f}$  denotes the trace of  $f$  on  $\Gamma_n$ .

*Proof.* Let  $f$  be a smooth function. According to Theorem III.15, every point  $y \in U_h(\Gamma_n)$  can be written in a unique way as  $y = x + t\nabla b_{\Omega_n}(x)$  with  $x = p_n(y) \in \Gamma_n$  and  $t \in (-h, h)$ . Moreover, one has

$$|f(x + t\nabla b_{\Omega_n}(x)) - f(x)|^2 \leq C^2 \|\partial_{\nabla b_{\Omega_n}(x)} f(x + y\nabla b_{\Omega_n}(x))\|_{L_y^2(-\frac{r_0}{2}, \frac{r_0}{2})}^2 |t|,$$

where  $\partial_{\nabla b_{\Omega_n}}$  stands for the derivative in the direction  $\nabla b_{\Omega_n}(x)$  and  $C$  is the norm of the continuous embedding of  $H^1([-\frac{r_0}{2}, \frac{r_0}{2}])$  into the space  $\mathcal{C}^{\frac{1}{2}}$  of  $\frac{1}{2}$ -Hölder continuous functions. Hence, using Theorem III.13, we get

$$\begin{aligned} \|f - f \circ p_n\|_{L^2(U_h(\Gamma_n))}^2 &= \int_{-h}^h \int_{\Gamma_n} |f(x + t\nabla b_{\Omega_n}(x)) - f(x)|^2 \det(dT_n) \, dx dt \\ &\leq \int_{-h}^h \int_{\Gamma_n} C^2 \|\partial_{\nabla b_{\Omega_n}} f(x + y\nabla b_{\Omega_n}(x))\|_{L_y^2(-\frac{r_0}{2}, \frac{r_0}{2})}^2 |t| \det(dT_n) \, dx dt \\ &\leq C^2 h^2 \int_{\Gamma_n} \|\partial_{\nabla b_{\Omega_n}} f(x + y\nabla b_{\Omega_n}(x))\|_{L_y^2(-\frac{r_0}{2}, \frac{r_0}{2})}^2 (1 + o_{h \rightarrow 0}(1)) \, dx \\ &\leq C^2 h^2 \|f\|_{H^1(U_{\frac{r_0}{2}}(\Gamma_n))}^2 (1 + o_{h \rightarrow 0}(1)). \end{aligned}$$

We conclude thanks to the density of the smooth functions in  $H^1$ .  $\square$

Using that  $u_{\Omega_n}$  is uniformly bounded in  $H^2(D)$ , let us apply Theorem III.24 to  $u_{\Omega_n}$  and  $\nabla u_{\Omega_n}$ . We obtain

$$\|u_{\Omega_n} - u_{\Omega_n} \circ p_n\|_{L^2(U_h(\Gamma_n))}^2 + \|\nabla u_{\Omega_n} - (\nabla u_{\Omega_n}) \circ p_n\|_{L^2(U_h(\Gamma_n))}^2 = o_{h \rightarrow 0}(h).$$

The end of the proof is similar to the one of Theorem III.6 and consists in using the extruded

surface approach to prove

$$\begin{aligned}
& \liminf_{n \rightarrow +\infty} F_3(\Omega_n) \\
& \geq (1 + o_{h \rightarrow 0}(1)) \liminf_{n \rightarrow +\infty} \frac{1}{2h} \int_{U_h(\Gamma_n)} j_3(x, \nabla b_{\Omega_n}(p_n(x)), E_{\Omega_n}(u_{\Omega_n})(x), \nabla E_{\Omega_n}(u_{\Omega_n})(x)) dx \\
& \geq (1 + o_{h \rightarrow 0}(1)) (1 + O\left(\frac{t}{h}\right)) \liminf_{n \rightarrow +\infty} \frac{1}{2(h-t)} \int_{U_{h-t}(\Gamma_\infty)} j_3(x, \nabla b_{\Omega_\infty}(p_\infty(x)), u^*(x), \nabla u^*(x)) dx \\
& \quad + o_{h \rightarrow 0}(1) + O\left(\frac{t}{h}\right) \\
& \geq F_3(\Omega_\infty) + o_{h \rightarrow 0}(1) + O\left(\frac{t}{h}\right),
\end{aligned}$$

which concludes the proof.

### III.3 Conclusion

In this chapter, we have introduced a new method to tackle the existence issue for shape optimization problems under uniform reach constraints on the considered shapes, of the type

$$\inf_{\Omega \in \mathcal{C}_{r_0}} \int_{\partial\Omega} j(x, \nu_{\partial\Omega}(x), B_{\partial\Omega}(x)) d\mu_{\partial\Omega}(x).$$

While several references such as [GY13; Dal18; Dal20] have already addressed similar questions on the same type of problems, we believe that the approaches developed in this chapter are on the one hand simpler, but also sufficiently robust to allow easy extension of the results to more general settings.

For example, we believe that minor adaptations of the developed proof techniques allow one to extend our results to the following cases without much effort:

- Under weaker regularity hypotheses, one could think of replacing the continuity assumption by lower semicontinuity on the integrand  $j_{\{1,2,3\}}$ . Another example would be to assume that  $f$  in equation (III.5) belongs to  $H^{1/2}(D)$  instead of  $\mathcal{C}(D)$ .

- More general PDEs could be considered Theorems III.10 and III.11. Extension to general elliptic equations associated with differential operators of the kind  $\nabla_\Gamma \cdot (\sigma \nabla_\Gamma)$  satisfying a coercivity property should be straightforward. We also believe that our framework allows extensions to nonlinear elliptic PDEs under reasonable assumptions.

- One could consider costs involving the solution of a minimization problem depending on  $\Omega$  but not necessarily related to a PDE. Indeed, in the proof of Theorem III.10, our study of the variational problem does not rely on the underlying PDE. We treated a case involving a convex minimization problem over the set of divergence-free vectors fields on  $\partial\Omega$  in [PRS22b].

All those generalizations do not seem obvious when using other methods.

## III.A Appendix

### III.A.1 Curvatures of a submanifold

Let us quickly review the definition of the mean curvature for an oriented  $(d-1)$ -submanifold of  $\mathbb{R}^d$  with  $\mathcal{C}^{1,1}$  regularity. To stick with our notation, we consider the submanifold to be the

boundary of some  $\Omega \in \mathcal{O}_{r_0}$ .

**Definition III.25.** *The Gauss map is the application which assigns to each  $x \in \Gamma = \partial\Omega$  the direct unit normal vector to  $\Gamma$  at  $x$ . In our setting it can be defined as*

$$\begin{aligned} N: \Gamma &\rightarrow \mathcal{S}^{d-1} \\ x &\mapsto \nabla b_\Omega(x). \end{aligned}$$

We can now define the following objects:

- the shape operator (or Weingarten map) is the differential of the Gauss map. For every  $x \in \Gamma$ , the tangent spaces  $T_x\Gamma$  and  $T_{N(x)}\mathcal{S}^{d-1}$  are equal as linear subspaces of  $\mathbb{R}^d$  and, the shape operator at  $x$  is self-adjoint where it is defined. See e.g. [Jos17, Chapter 5] for a general introduction.
- The trace of the shape operator is called the mean curvature and is denoted  $H^2$ .
- The determinant of the shape operator is called the Gauss curvature.

**Remark III.26.** *The Gauss map is  $\frac{1}{r_0}$ -Lipschitz continuous (see Theorem III.1), where  $r_0$  is the reach of  $\Gamma$ . Thus, the shape operator is in  $L^\infty$  and at almost every  $x \in \Gamma$ , all the eigenvalues  $\kappa_1(x), \dots, \kappa_{d-1}(x)$  of the shape operator are bounded in modulus by  $\frac{1}{r_0}$ .*

We insist on the fact that  $N$  is defined only on  $\Gamma$  and thus the shape operator is not defined on  $\mathbb{R}^d$  or any tubular neighborhood of  $\Gamma$ . Nevertheless, we have the following property.

**Lemma III.27.** *The mean curvature coincides with the trace of  $\nabla^2 b_\Omega$  on  $\Gamma$ .*

*Proof.* Let  $x \in \Gamma$  and let  $\mathcal{B}$  be an orthonormal basis of  $T_x\Gamma$ . Using the identification between  $T_x\Gamma$  and the tangent hyperplane (see Theorem III.12), we obtain that  $\{\nabla b_\Omega(x)\} \cup \mathcal{B}$  is an orthonormal basis of  $\mathbb{R}^d$ .  $\nabla b_\Omega$  is constant along the direction  $\nabla b_\Omega(x)$  (see e.g. [DZ11, Theorem 7.8.5.ii]). As a consequence, the trace of  $\nabla^2 b_\Omega$  and the mean curvature coincide.  $\square$

### III.A.2 $R$ -convergence: proof of Theorem III.3

The compactness property follows from two facts. First, the Arzelà–Ascoli theorem, combined with the fact that every function  $b_\Omega$ , for  $\Omega \in \mathcal{O}_{r_0}$ , is 1-Lipschitz continuous. Second, the reach constraint which imposes a uniform bound on the second derivative of  $b_\Omega$ . These two facts are used in [DZ11] and [Dal18] to get the sequential compactness results used below.

Let  $(\Omega_n)_{n \in \mathbb{N}}$  denote a sequence in  $\mathcal{O}_{r_0}$ . By the compactness property of sets of uniformly positive reach proved in [DZ11, Chapter 6], it follows that, up to a subsequence,  $b_{\Omega_n}$  converges to  $b_{\Omega_\infty}$  for the  $C^0$  topology on  $D$ . In [Dal18] the convergence is shown to hold also for the strong  $\mathcal{C}^{1,\alpha}$  topology (for  $\alpha < 1$ ) and for the weak  $W^{2,\infty}$  topology in a  $r$ -tubular neighborhood of  $\partial\Omega_\infty$ , with  $r < r_0$ .

As a consequence,  $\text{Reach}(\Gamma_\infty) \geq r_0$ . In particular, according to Theorem III.1,  $b_{\Omega_\infty}$  is  $\mathcal{C}^{1,1}$  on  $\overline{U_r(\Gamma_\infty)}$ .

### III.A.3 The Laplace–Beltrami equation on a manifold: proof of Theorem III.9

Let  $(\partial\Omega, \mathbf{g})$  denote a closed compact manifold. We explain hereafter how to understand the equation  $\Delta_{\partial\Omega} v = h$  in  $\partial\Omega$  in a weak sense, whenever  $\Omega \in \mathcal{O}_{r_0}$ . Indeed, under this assumption,

---

2. Note that in differential geometry it is common to define the mean curvature as the trace of the shape operator divided by  $(d-1)$ .

$\partial\Omega$  is a  $\mathcal{C}^{1,1}$  submanifold according to Theorem III.1, not necessarily  $\mathcal{C}^2$ , which justifies why such an equation cannot be understood in a strong sense.

The key ingredient in what follows is the Rellich–Kondrachov lemma, stating the compactness of the embedding  $H_*^1(\partial\Omega) \hookrightarrow L^2(\partial\Omega)$ .

**Theorem III.28** (Rellich–Kondrachov theorem on surfaces). *Let  $\Omega \in \mathcal{O}_{r_0}$ . Let  $(u_n)_{n \in \mathbb{N}}$  denote a sequence in  $H_*^1(\partial\Omega)$  such that  $(\int_{\partial\Omega} |\nabla u_n(x)|^2 d\mu_{\partial\Omega})_{n \in \mathbb{N}}$  is bounded. There exists  $u^* \in H_*^1(\partial\Omega)$  such that, up to a subsequence,  $(u_n)_{n \in \mathbb{N}}$  converges to  $u^*$  weakly in  $H_*^1(\partial\Omega)$  and strongly in  $L^2(\partial\Omega)$ .*

*Proof.* According to [Del00, Th 4.5.ii], since  $\partial\Omega$  is  $\mathcal{C}^{1,1}$ , the  $L^2$  norm  $\|\cdot\|_{L^2(\partial\Omega)}$  on the surface  $\partial\Omega$  and the  $L^2$  norm  $L^2(\partial\Omega) \ni u \mapsto \|u \circ p_\Omega\|_{L^2(U_h(\partial\Omega))}$  on the thickened surface  $U_h(\partial\Omega)$  are equivalent whenever  $h > 0$  is small enough, where  $p_\Omega(x)$  denotes the orthogonal projection of  $x$  onto  $\partial\Omega$ , that is,  $p_\Omega(x) = x - b_\Omega(x)\nabla b_\Omega(x)$ , and  $U_h(\partial\Omega) = \{x \in \mathbb{R}^d \mid |b_\Omega(x)| < h \text{ and } p_\Omega(x) \in \partial\Omega\}$ .

Similarly, according to [Del00, Th 4.7.v], since  $\partial\Omega$  is  $\mathcal{C}^{1,1}$ , the norm  $\|\cdot\|_{H_*^1(\partial\Omega)}$  defined as

$$\|u\|_{H_*^1(\partial\Omega)}^2 = \int_{\partial\Omega} |\nabla_\Gamma u|^2 d\mu_{\partial\Omega},$$

and the norm  $\|\cdot\|_{H_{U_h}^1(\partial\Omega)}$  given by

$$\|u\|_{H_{U_h}^1(\partial\Omega)} = \frac{1}{2h} \int_{U_h(\partial\Omega)} |\nabla_\Gamma u \circ p_\Omega|^2 d\mu_{\partial\Omega}$$

are equivalent whenever  $h > 0$  is small enough. We conclude by using the standard Rellich–Kondrachov theorem (see e.g. [Bre11, sub 9.3]) on the thickened surface  $U_h(\partial\Omega)$ .  $\square$

The following result is a Poincaré type lemma, uniform with respect to the chosen surface in the set  $\mathcal{O}_{r_0}$ .

**Proposition III.29** (Poincaré lemma on a surface). *Let  $r_0 > 0$  and  $\Omega \in \mathcal{O}_{r_0}$ . There exists  $C(r_0, D) > 0$  such that*

$$\forall u \in H_*^1(\Gamma), \quad \int_\Gamma |\nabla_\Gamma u(x)|^2 d\mu_\Gamma \geq C(r_0, D) \int_\Gamma |u(x)|^2 d\mu_\Gamma.$$

*Proof.* Let  $(\Omega_n, v_n)_{n \in \mathbb{N}}$ , with  $v_n \in H_*^1(\Gamma_n)$ , be a minimizing sequence for the problem

$$\inf_{\Omega \in \mathcal{O}_{r_0}} \inf_{u \in H_*^1(\Gamma)} \frac{\int_\Gamma |\nabla_\Gamma u(x)|^2 d\mu_\Gamma}{\int_\Gamma |u(x)|^2 d\mu_\Gamma}.$$

Let us argue by contradiction, assuming that

$$\int_{\Gamma_n} |\nabla_\Gamma v_n(x)|^2 d\mu_{\Gamma_n} \leq \frac{1}{n} \quad \text{and} \quad \int_{\Gamma_n} |v_n(x)|^2 d\mu_{\Gamma_n} = 1$$

by homogeneity of the Rayleigh quotient. According to Theorem III.3, we can assume without loss of generality that  $(\Omega_n)_{n \in \mathbb{N}}$   $R$ -converges toward  $\Omega_\infty \in \mathcal{O}_{r_0}$ .

Let  $p_n$  denote the orthogonal projection on  $\Gamma_n$  and let us introduce the function  $w_n$  defined in  $U_h(\Gamma_n)$  for  $h$  as in Theorem III.13 and  $n$  large enough by  $w_n = v_n \circ p_n$ . We follow exactly the same lines as in the first step of the proof of Theorem III.10. A direct adaptation of the first



step of the proof of Theorem III.10 yields

$$\int_{\Gamma_\infty} |\nabla_{\Gamma_\infty} w_n(y)|^2 d\mu_{\Gamma_\infty}(y) = \int_{\Gamma_n} |\nabla_{\Gamma_n} v_n|^2 d\mu_{\Gamma_n}(x) + o(1). \quad (\text{III.33})$$

We infer that

$$\int_{\Gamma_\infty} |\nabla_{\Gamma_\infty} w_n|^2 d\mu_{\Gamma_\infty}(x) \leq \frac{1}{n} + o(1).$$

By using Theorem III.28, we get that the sequence  $(w_n)_{n \in \mathbb{N}}$  converges up to a subsequence toward  $w^* \in H_*^1(\Gamma_\infty)$  weakly in  $H^1(\Gamma_\infty)$  and strongly in  $L^2(\Gamma_\infty)$ . Up to extracting a subsequence, we get

$$\begin{aligned} \int_{\Gamma_\infty} |\nabla_{\Gamma_\infty} w^*(x)|^2 d\mu_{\Gamma_\infty} &\leq \liminf_{n \rightarrow +\infty} \int_{\Gamma_\infty} |\nabla_{\Gamma_\infty} w_n(x)|^2 d\mu_{\Gamma_\infty} = 0, \\ \int_{\Gamma_\infty} |w^*(x)|^2 d\mu_{\Gamma_\infty} &= 1, \\ \int_{\Gamma_\infty} w_n(x) d\mu_{\Gamma_\infty} &= 0. \end{aligned}$$

By using the first equality, we get that  $w^*$  is constant on  $\Gamma$  and we obtain a contradiction with the two last equalities above.  $\square$

Let us now prove Theorem III.9. Let  $(u_n)_{n \in \mathbb{N}}$  denote a minimizing sequence for Problem (III.7). Since  $(\mathcal{E}_\Gamma(u_n))_{n \in \mathbb{N}}$  is bounded, and since

$$\mathcal{E}_\Gamma(u_n) \geq C(d, r_0) \|u_n\|_{L^2(\Gamma)}^2 - \|h\|_{L^2(\Gamma)} \|u_n\|_{L^2(\Gamma)}$$

according to Theorem III.29, we infer that  $(\|u_n\|_{L^2(\Gamma)})_{n \in \mathbb{N}}$  is bounded. Since

$$\int_{\Gamma} |\nabla_{\Gamma} u_n(x)|^2 d\mu_{\Gamma} = \mathcal{E}_\Gamma(u_n) + \int_{\Gamma} u_n(x) h(x) d\mu_{\Gamma} \leq \|h\|_{L^2(\Gamma)} \|u_n\|_{L^2(\Gamma)},$$

we infer the existence of  $u^* \in H_*^1(\Gamma)$  such that, up to a subsequence,  $(u_n)_{n \in \mathbb{N}}$  converges weakly in  $H_*^1(\Gamma)$  and strongly in  $L^2(\Gamma)$ . Up to extracting a subsequence, we get

$$\inf_{u \in H_*^1(\Gamma)} \mathcal{E}_\Gamma(u) = \liminf_{n \rightarrow +\infty} \mathcal{E}_\Gamma(u_n) \geq \int_{\Gamma} |\nabla_{\Gamma} u^*(x)|^2 d\mu_{\Gamma} - \int_{\Gamma} u^*(x) h(x) d\mu_{\Gamma} = \mathcal{E}_\Gamma(u^*)$$

and the existence follows. The uniqueness is standard and follows from the strong convexity of the functional  $\mathcal{E}_\Gamma$ .

## Chapter IV

# Minimization of magnetic forces on Stellarator coils

This chapter is taken from the following article (also referred as [RV22]):

R. Robin and F. A. Volpe. “Minimization of magnetic forces on stellarator coils”. In: *Nuclear Fusion* 62.8 (2022), p. 086041

The Section IV.B is original and is not part of the published article.

Magnetic confinement devices for nuclear fusion can be large and expensive. Compact stellarators are promising candidates for cost-reduction, but introduce new difficulties: confinement in smaller volumes requires higher magnetic field, which calls for higher coil-currents and ultimately causes higher Laplace forces on the coils - if everything else remains the same. This motivates the inclusion of force reduction in stellarator coil optimization. In the present chapter we consider a coil winding surface, we prove that there is a natural and rigorous way to define the Laplace force (despite the magnetic field discontinuity across the current-sheet), we provide examples of cost associated (peak force, surface-integral of the force squared) and discuss easy generalizations to parallel and normal force-components, as these will be subject to different engineering constraints. Such costs can then be easily added to the figure of merit in any multi-objective stellarator coil optimization code. We demonstrate this for a generalization of the REGCOIL code [Lan17], which we rewrote in python, and provide numerical examples for the NCSX [Zar+01](now QUASAR) design. We present results for various definitions of the cost function, including peak force reductions by up to 40 %, and outline future work for further reduction.

### IV.1 Introduction

Stellarators are non-axisymmetric toroidal devices that magnetically confine fusion plasmas [Hel14]. Thanks to specially shaped coils they do not require a current in the plasma, hence are more stable and steady-state than tokamaks. However, they exhibit comparable confinement, hence tend to be about as large. Like tokamaks, confinement can be improved (and size reduced) by adopting stronger magnetic fields.

Fields as high as 8-12 T were only tested in two series of tokamak experiments at the MIT and ENEA, culminated respectively in Alcator C-mod [Mar+15] and FTU [Puc+15]. For comparison,

ITER has a field of 5.3 T on axis [02]. Other high-field tokamaks were designed but not built [Cop+13; Mea+02], although the new high-field SPARC tokamak has been designed, modeled and its construction is expected to start in 2021 [Cre+20].

For stellarators and heliotrons, there is broad agreement that power-plants will require at least 4-6 T [SIN10], but fields as high as 8-12 T have only been proposed very recently [Que+18]. Two private companies are working toward that goal<sup>1,2</sup>. Generating strong fields requires high currents and of course results in high forces on the coils (unless their design is modified, as we will argue in this chapter). Up to 5 T, the issue can be resolved by adequately reinforced coil-support structures and coil-spacers [SEB13]. However, a further increase to 10 T will result in  $4\times$  higher forces. This calls for including force-reduction in the coil design and optimization process, along with other criteria.

Such need was recognized earlier on for heliotrons, and spurred reduced force (so-called force-free) heliotron designs [Ima+02]. From a mathematical standpoint this is not too surprising, since helical fields in heliotrons resemble the eigenfunctions of the curl operator on a torus [Alo+18]:  $\nabla \times \mathbf{B} = \lambda \mathbf{B}$ . This, combined with Maxwell-Ampere law, implies that  $\mathbf{B}$  and the current density  $\mathbf{j}$  are parallel, and there are no Laplace forces on the coils.

Modular coils for advanced stellarators, on the other hand, are the result of numerical optimization. The most common optimization criterion is to reproduce the target magnetic field to within one part in  $10^4$  or  $10^5$ . Typically this is solved on a 2-D toroidal surface external to the plasma, called Coil Winding Surface (CWS). On that surface, numerical codes compute the current potential (thus, ultimately, the current pattern) that best reproduces the target plasma boundary, in a least-squares sense [Lan17]. This is the principle of the seminal NESCOIL code [Mer86; Mer87]. Further developments included engineering-constrained nonlinear optimizers [Str+04] and the Tikhonov-regularized REGCOIL [Lan17]. The latter includes the squared coil-current density in the objective function, which leads to more “gentle”, easier-to-build coil shapes. All these codes fix the CWS; more recently, a free-CWS 3-D search method was developed [Zhu+18c].

In the present chapter, we generalize REGCOIL to include coil-force reduction. This is obtained by adding to the objective function a term quantifying the Laplace forces on the CWS. Several metrics are possible, for example the surface-integral of the squared Laplace force, or the peak value of the Laplace force. We recall that the Laplace forces are a self-interaction  $\mathbf{L}$  of a surface-current (of density  $\mathbf{j}$ ). To that end, first we derive a rigorous and computationally amenable expression, Eqs. IV.14-IV.17, for the force  $\mathbf{L}$  exerted by a surface-current of density  $\mathbf{j}_1$  on a surface-current of density  $\mathbf{j}_2$ . Section IV.2 and Appendices IV.A.1-IV.A.4 offer an extensive mathematical derivation of such expression. Based on that, possible cost functions are proposed and briefly discussed in Section IV.3. In Section IV.4 we determine that, based on the Hilbert space where the current density is defined, the correct norm to use for its regularization is the  $H^1$  norm. This is new w.r.t. REGCOIL and other codes adopting the  $L^2$  norm, and has numerical implications. Finally Section IV.4.3 illustrates the numerical results obtained with the two main cost-functions for the quasi-axisymmetric stellarator design formerly known as NCSX [Zar+01], then QUASAR, including a reduction of the peak force by up to 40%.

## IV.2 Laplace force on a surface

### IV.2.1 Notations

We start by introducing the following notations:

1. Type one energy. <https://www.typeoneenergy.com>
2. Renaissance fusion. <https://stellarator.energy>

- $S$  is a smooth 2-dimensional Riemannian submanifold of  $\mathbb{R}^3$ , diffeomorphic to the 2-torus standing for the CWS.
- $\mathbf{n}$  is the unit vector field normal to  $S$  and pointing outward.
- $\mathfrak{X}(S)$  is the set of smooth vector fields on  $S$ .
- $\langle X \cdot Y \rangle$  denote the scalar product (in  $\mathbb{R}^3$ ) between the vector fields  $X$  and  $Y$ . When both vector are tangent to  $S$ , we sometime denote  $\langle X \cdot Y \rangle_{T_x S}$  the scalar product at  $x \in S$  (which coincides with the one in  $\mathbb{R}^3$ ).
- $L^p(S)$  and  $H^1(S)$  are the Hilbert spaces defined as the completion of  $\mathcal{C}^\infty(S)$  for the norms

$$|f|_{L^p(S)} = \left( \int_S f^p dS \right)^{1/p}$$

$$|f|_{H^1(S)}^2 = \int_S (f^2 + \langle \nabla f \cdot \nabla f \rangle) dS.$$

- $\mathfrak{X}^p(S)$  and  $\mathfrak{X}^{1,2}(S)$  are the Hilbert spaces defined as the completion of  $\mathfrak{X}(S)$  for the norms

$$|\mathbf{j}|_{\mathfrak{X}^p(S)} = \left| \sqrt{\mathbf{j}_x^2 + \mathbf{j}_y^2 + \mathbf{j}_z^2} \right|_{L^p(S)}$$

$$|\mathbf{j}|_{\mathfrak{X}^{1,2}(S)} = \sqrt{|\mathbf{j}_x|_{H^1(S)}^2 + |\mathbf{j}_y|_{H^1(S)}^2 + |\mathbf{j}_z|_{H^1(S)}^2}$$

where  $\mathbf{j}_x, \mathbf{j}_y$  and  $\mathbf{j}_z$  are the components of  $\mathbf{j}$  in  $\mathbb{R}^3$  for an arbitrary orthogonal basis.

- The spaces  $L^p(S, \mathbb{R}^3)$  and  $H^{1,2}(S, \mathbb{R}^3)$  are related to  $\mathcal{C}^\infty(S, \mathbb{R}^3)$  in the same way as  $\mathfrak{X}^p(S)$  and  $\mathfrak{X}^{1,2}(S)$  are related to  $\mathfrak{X}(S)$ .
- $\pi$  is the projector on the tangent bundle. For any  $\mathbf{Y} \in \mathcal{C}^\infty(S, \mathbb{R}^3)$ , we define

$$\forall \mathbf{x} \in S, (\pi(\mathbf{Y}))_{\mathbf{x}} = \mathbf{Y}_{\mathbf{x}} - \langle \mathbf{Y}_{\mathbf{x}} \cdot \mathbf{n}(\mathbf{x}) \rangle \mathbf{n}(\mathbf{x}) \in T_{\mathbf{x}} S. \quad (\text{IV.1})$$

Since  $\pi(\mathbf{Y})$  is clearly a tangent vector field on  $S$ , it belongs to  $\mathfrak{X}(S)$ .

## IV.2.2 Limit definition of Laplace force exerted by a current-sheet on itself

Let  $\mathbf{j}$  be a vector field on  $S$ , representing the surface-current density, i.e. the current per unit length (not per unit surface, as is usually the case for this notation). The Laplace force is the magnetic component of the Lorentz force; the Laplace force per unit *surface* (not per unit volume) is given by  $\mathbf{F} = \mathbf{j} \times \mathbf{B}$ , although here, quite often, it will simply be called 'force', for brevity.

A surface-current of density  $\mathbf{j}$  causes a discontinuity in the component of the magnetic field  $\mathbf{H}$  tangential to the surface:  $\mathbf{n}_{12} \times (\mathbf{H}_2 - \mathbf{H}_1) = \mathbf{j}$ . The resulting jump in the tangential component of  $\mathbf{B}$  results in a normal force wherever  $\mathbf{j} \neq 0$ . That force, proportional to  $|\mathbf{j}|^2$ , tries to increase the thickness of the CWS. To ensure that that force remains reasonably small, one can easily add a cost  $|\mathbf{j}|_{L^2}$  or  $|\mathbf{j}|_\infty$  to the multi-objective figure-of-merit or optimize under a constraint on  $\mathbf{j}$ .

From now on, though, we will focus on the other contributions to the Laplace force. We define them in a location  $\mathbf{y} \in S$  as follows:

$$\mathbf{F}(\mathbf{y}) = \lim_{\varepsilon \rightarrow 0} \frac{1}{2} \{ \mathbf{j}(\mathbf{y}) \times [\mathbf{B}(\mathbf{y} + \varepsilon \mathbf{n}(\mathbf{y})) + \mathbf{B}(\mathbf{y} - \varepsilon \mathbf{n}(\mathbf{y}))] \}.$$

Let us focus on the case where  $\mathbf{B}$  is only generated by currents on  $S$ ; there are no permanent

magnets nor magnetically susceptible media. In any  $\mathbf{y} \notin S$ , the field is given by the Biot-Savart law in vacuo:

$$\mathbf{B}(\mathbf{y}) = \mathbf{BS}(\mathbf{j})(\mathbf{y}) = \int_S \mathbf{j}(\mathbf{x}) \times \frac{\mathbf{y} - \mathbf{x}}{|\mathbf{y} - \mathbf{x}|^3} dS(\mathbf{x}), \quad (\text{IV.2})$$

where, to reduce the amount of notation, we dropped the  $\frac{\mu_0}{4\pi}$  factor in front of the integral. The notation  $\mathbf{BS}(\mathbf{j})$  refers to the Biot-Savart operator, function of  $\mathbf{j}$ , that maps each  $\mathbf{y} \notin S$  in the local field,  $\mathbf{B}(\mathbf{y})$ .

**Remark IV.1.**  $\mathbf{BS}(\mathbf{j})$  cannot be defined on  $S$ , unless  $\mathbf{j} = 0$ . This is a consequence of  $\frac{1}{|\mathbf{y} - \mathbf{x}|^2}$  not being integrable for  $\mathbf{y} \in S$ . The jump in the tangential component of the induced magnetic field mentioned above is thus caused by the discontinuity of  $\mathbf{BS}(\mathbf{j})(\mathbf{y} + \varepsilon \mathbf{n}(\mathbf{y}))$  at  $\varepsilon = 0$ .

However, since  $\mathbf{BS}(\mathbf{j})$  is well-defined in locations  $\notin S$ , we can define for any  $\mathbf{y} \in S$  and any  $\varepsilon > 0$  the bilinear map

$$\mathbf{L}_\varepsilon(\mathbf{j}_1, \mathbf{j}_2)(\mathbf{y}) = \frac{1}{2} \{ \mathbf{j}_1(\mathbf{y}) \times [\mathbf{BS}(\mathbf{j}_2)(\mathbf{y} + \varepsilon \mathbf{n}(\mathbf{y})) + \mathbf{BS}(\mathbf{j}_2)(\mathbf{y} - \varepsilon \mathbf{n}(\mathbf{y}))] \}. \quad (\text{IV.3})$$

This describes the Laplace force that a surface-current of density  $\mathbf{j}_2$  exerts on another of density  $\mathbf{j}_1$ , per unit surface. Since we are dealing with a stellarator, these currents are constant in time and there is no need to include induced fields and the associated forces.

The ‘average Laplace force’ that a current of density  $\mathbf{j}$  exerts on point  $\mathbf{y} \in S$  is thus  $\mathbf{L}(\mathbf{j})(\mathbf{y}) = \lim_{\varepsilon \rightarrow 0} \mathbf{L}_\varepsilon(\mathbf{j}, \mathbf{j})(\mathbf{y})$ .

This definition, however, raises several questions:

1. Under which assumptions on  $\mathbf{j}$  can we ensure that  $\mathbf{L}(\mathbf{j})$  is well defined (i.e., that the limit is well -defined)?
2. Can we find an explicit expression of  $\mathbf{L}(\mathbf{j})$  (i.e., without a limit on  $\varepsilon$ )?
3. Which functional space does  $\mathbf{L}(\mathbf{j})$  belong to (for  $\mathbf{j}$  in a given functional space)?

The first point is more theoretical, but is necessary to answer the second and third one, which have very practical consequences. Indeed, without an explicit expression for  $\mathbf{L}(\mathbf{j})$ , the numerical computation of the Laplace force may be a complex matter. A typical approach would involve 3 different scales. From the smallest to the largest, these are the discretisation-length of  $S$ ,  $h$ , the infinitesimal displacement  $\varepsilon$ , and the characteristic distance of variation of the magnetic field,  $d_B$ .

An accurate computation of  $\mathbf{B}(\mathbf{y} + \varepsilon \mathbf{n}(\mathbf{y}))$  requires  $S$  to be finely discretized, with the discretisation-length  $h \ll \varepsilon$ . This is because  $\int_S |\mathbf{y} + \varepsilon \mathbf{n}(\mathbf{y}) - \mathbf{x}|^{-2} dS(\mathbf{x})$  blows up when  $\varepsilon \rightarrow 0$ . Indeed when we replace the integral with a discrete sum (with  $|y_{i,j} - y_{i+1,j}| \approx |y_{i,j} - y_{i,j+1}| \approx h$ ) and take the limit for small  $\varepsilon$ ,

$$\tilde{\mathbf{B}}(\mathbf{y}_{i,j} + \varepsilon \mathbf{n}(\mathbf{y}_{i,j})) = \sum_{k,l} \mathbf{j}(\mathbf{x}_{k,l}) \times \frac{\mathbf{y}_{i,j} + \varepsilon \mathbf{n}(\mathbf{y}_{i,j}) - \mathbf{x}_{k,l}}{|\mathbf{y}_{i,j} + \varepsilon \mathbf{n}(\mathbf{y}_{i,j}) - \mathbf{x}_{k,l}|^3} \underset{h \text{ fixed}}{\overset{\varepsilon \rightarrow 0}{\approx}} \frac{\mathbf{j}(\mathbf{y}_{i,j}) \times \mathbf{n}(\mathbf{y}_{i,j})}{\varepsilon^2}$$

which for small  $\varepsilon$  diverges like  $\varepsilon^{-2}$ , as shown in Fig. IV.1 for NCSX.

The semi-sum  $\frac{\mathbf{B}(\mathbf{y} + \varepsilon \mathbf{n}(\mathbf{y})) + \mathbf{B}(\mathbf{y} - \varepsilon \mathbf{n}(\mathbf{y}))}{2}$  is numerically more stable, but we still need  $h \lesssim \varepsilon$  (as it will be shown later in Fig. IV.3). Such fine mesh makes it costly to accurately compute  $\mathbf{L}(\mathbf{j})(\mathbf{y})$  as  $\lim_{\varepsilon \rightarrow 0} \mathbf{L}_\varepsilon(\mathbf{j}, \mathbf{j})(\mathbf{y})$ .

The functional space of  $\mathbf{L}(\mathbf{j})$  is also important to understand what type of penalization can be applied to minimize this force, or a related metric.

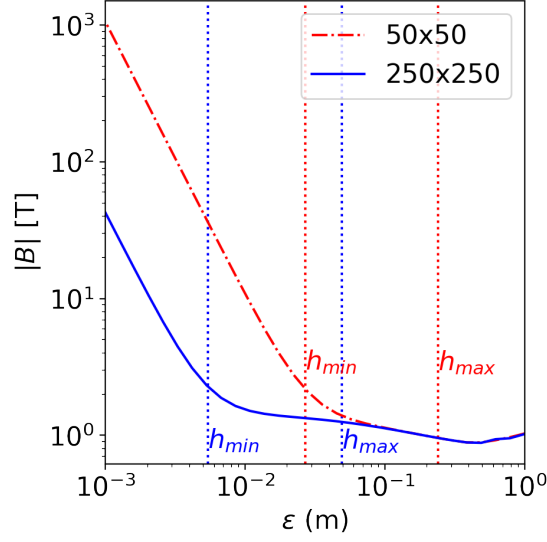


Figure IV.1 – Average norm of  $\mathbf{B}$  as a function of the distance  $\varepsilon$  from the surface  $S$ , for two different grids, more coarse (red) or fine (blue).  $h_{min}$  and  $h_{max}$  refer to the smallest and largest mesh size (the poloidal $\times$ toroidal mesh being non-uniform in real space). The plot guides the selection of  $\varepsilon$ : an excessively small value,  $\varepsilon \lesssim h$ , results in the numerical artifact of a diverging field.

### IV.2.3 Computing the Laplace force exerted by one current-sheet on another

Consider two linear densities of surface-currents  $\mathbf{j}_1, \mathbf{j}_2 \in \mathfrak{X}^{1,2}(S)$  and fix  $\varepsilon > 0$ .

Thanks to the well-known formula  $\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = (\mathbf{A} \cdot \mathbf{C})\mathbf{B} - (\mathbf{A} \cdot \mathbf{B})\mathbf{C}$ , we obtain from Eqs.IV.2 and IV.3 that

$$\begin{aligned} \mathbf{L}_\varepsilon(\mathbf{j}_1, \mathbf{j}_2)(y) &= \int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \left( \frac{\mathbf{y} - \mathbf{x} + \varepsilon \mathbf{n}(\mathbf{y})}{2|\mathbf{y} - \mathbf{x} + \varepsilon \mathbf{n}(\mathbf{y})|^3} + \frac{\mathbf{y} - \mathbf{x} - \varepsilon \mathbf{n}(\mathbf{y})}{2|\mathbf{y} - \mathbf{x} - \varepsilon \mathbf{n}(\mathbf{y})|^3} \right) \rangle \mathbf{j}_2(\mathbf{x}) d\mathbf{x} \\ &\quad - \int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{j}_2(\mathbf{x}) \rangle \left( \frac{\mathbf{y} - \mathbf{x} + \varepsilon \mathbf{n}(\mathbf{y})}{2|\mathbf{y} - \mathbf{x} + \varepsilon \mathbf{n}(\mathbf{y})|^3} + \frac{\mathbf{y} - \mathbf{x} - \varepsilon \mathbf{n}(\mathbf{y})}{2|\mathbf{y} - \mathbf{x} - \varepsilon \mathbf{n}(\mathbf{y})|^3} \right) d\mathbf{x}. \end{aligned} \quad (\text{IV.4})$$

The difficulty here is that  $\frac{1}{|x|^2}$  is not integrable in two dimensions (Remark IV.1). Hence, it does not make sense to take the limit for  $\varepsilon \rightarrow 0$  directly inside the integral. Nevertheless, we can use the following equality:

$$\int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \frac{\mathbf{y} - \mathbf{x} + \varepsilon \mathbf{n}(\mathbf{y})}{|\mathbf{y} - \mathbf{x} + \varepsilon \mathbf{n}(\mathbf{y})|^3} \rangle \mathbf{j}_2(\mathbf{x}) d\mathbf{x} = \int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \nabla_{\mathbf{x}} \frac{1}{|\mathbf{y} - \mathbf{x} + \varepsilon \mathbf{n}(\mathbf{y})|} \rangle \mathbf{j}_2(\mathbf{x}) d\mathbf{x} \quad (\text{IV.5})$$

where  $\nabla_{\mathbf{x}}$  is the gradient in  $\mathbb{R}^3$  with respect to the variable  $\mathbf{x}$ . We would like to integrate by part to take advantage of the integrability of  $\frac{1}{|x|}$ . For this we decompose  $\nabla_{\mathbf{x}}$  into the tangential part of the gradient,  $\nabla_S$ , and normal component,  $\nabla_{\perp}$ . As a consequence, for the first integral in

Eq. IV.4 we have the following equalities:

$$\int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \frac{\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|^3} \rangle \mathbf{j}_2(\mathbf{x}) d\mathbf{x} \quad (\text{IV.6})$$

$$= \int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \nabla_{\mathbf{x}} \frac{1}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|} \rangle \mathbf{j}_2(\mathbf{x}) d\mathbf{x} \quad (\text{IV.7})$$

$$= \int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \nabla_S \frac{1}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|} \rangle \mathbf{j}_2(\mathbf{x}) d\mathbf{x} \quad (\text{IV.8})$$

$$+ \int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \frac{\langle \mathbf{y} - \mathbf{x}, \mathbf{n}(\mathbf{x}) \rangle \pm \varepsilon \langle \mathbf{n}(\mathbf{y}), \mathbf{n}(\mathbf{x}) \rangle}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|^3} \mathbf{n}(\mathbf{x}) \rangle \mathbf{j}_2(\mathbf{x}) d\mathbf{x}. \quad (\text{IV.9})$$

Similarly, for the second integral in Eq. IV.4 we have:

$$\int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{j}_2(\mathbf{x}) \rangle \frac{\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|^3} d\mathbf{x} \quad (\text{IV.10})$$

$$= \int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{j}_2(\mathbf{x}) \rangle \nabla_{\mathbf{x}} \frac{1}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|} d\mathbf{x} \quad (\text{IV.11})$$

$$= \int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{j}_2(\mathbf{x}) \rangle \nabla_S \frac{1}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|} d\mathbf{x} \quad (\text{IV.12})$$

$$+ \int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{j}_2(\mathbf{x}) \rangle \frac{\langle \mathbf{y} - \mathbf{x}, \mathbf{n}(\mathbf{x}) \rangle \pm \varepsilon \langle \mathbf{n}(\mathbf{y}), \mathbf{n}(\mathbf{x}) \rangle}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|^3} \mathbf{n}(\mathbf{x}) d\mathbf{x} \quad (\text{IV.13})$$

Integrals IV.8 and IV.12, with integrands tangential to  $S$  (“tangential terms”) are dealt with in Appendices IV.A.1 and IV.A.2. Integrals IV.9 and IV.13, with integrands normal to  $S$  (“normal terms”) are treated in Appendices IV.A.3 and IV.A.4. Together, those Appendices constitute proof of the following theorem.

**Theorem IV.2.** *Let  $\mathbf{j}_1, \mathbf{j}_2 \in \mathfrak{X}^{1,2}(S)$ . Then  $\mathbf{L}_\varepsilon(\mathbf{j}_1, \mathbf{j}_2)$  has an  $\varepsilon \rightarrow 0$  limit in  $L^p(S, \mathbb{R}^3)$ , for any  $1 \leq p < \infty$ , denoted  $\mathbf{L}(\mathbf{j}_1, \mathbf{j}_2)$ . Furthermore,  $\mathbf{L}$  is a continuous bilinear map  $\mathfrak{X}^{1,2}(S) \times \mathfrak{X}^{1,2}(S) \rightarrow L^p(S, \mathbb{R}^3)$  given by*

$$\mathbf{L}(\mathbf{j}_1, \mathbf{j}_2)(\mathbf{y}) = - \int_S \frac{1}{|\mathbf{y} - \mathbf{x}|} [\operatorname{div}_{\mathbf{x}}(\pi_{\mathbf{x}} \mathbf{j}_1(\mathbf{y})) + \pi_{\mathbf{x}} \mathbf{j}_1(\mathbf{y}) \cdot \nabla_{\mathbf{x}}] \mathbf{j}_2(\mathbf{x}) d\mathbf{x} \quad (\text{IV.14})$$

$$+ \int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{n}(\mathbf{x}) \rangle \frac{\langle \mathbf{y} - \mathbf{x}, \mathbf{n}(\mathbf{x}) \rangle}{|\mathbf{y} - \mathbf{x}|^3} \mathbf{j}_2(\mathbf{x}) d\mathbf{x} \quad (\text{IV.15})$$

$$+ \int_S \frac{1}{|\mathbf{y} - \mathbf{x}|} [\langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{j}_2(\mathbf{x}) \rangle \operatorname{div}_{\mathbf{x}}(\pi_{\mathbf{x}}) + \nabla_{\mathbf{x}} \langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{j}_2(\mathbf{x}) \rangle] d\mathbf{x} \quad (\text{IV.16})$$

$$- \int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{j}_2(\mathbf{x}) \rangle \frac{\langle \mathbf{y} - \mathbf{x}, \mathbf{n}(\mathbf{x}) \rangle}{|\mathbf{y} - \mathbf{x}|^3} \mathbf{n}(\mathbf{x}) d\mathbf{x} \quad (\text{IV.17})$$

**Remark IV.3.** — *The notation  $V \cdot \nabla_{\mathbf{x}} F$  (where  $V \in \mathfrak{X}(S)$  is a 2D vector and  $F \in C^\infty(S, \mathbb{R}^3)$  a 3D one) stands for  $\sum_{\alpha=1}^2 \sum_{i=1}^3 V^\alpha \partial_\alpha F^i \mathbf{e}_i$ . Here  $\alpha$  is the index for the surface coordinates ( $\theta$  and  $\varphi$ , for example), whereas  $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$  is a basis of  $\mathbb{R}^3$ .*  
 —  $\operatorname{div}_{\mathbf{x}}(\pi_{\mathbf{x}})$  stands for the 3D vector  $\sum_{i=1}^3 \operatorname{div}_{\mathbf{x}}(\pi_{\mathbf{x}} \mathbf{e}_i) \mathbf{e}_i$

### IV.2.4 Justification from a 3D current modelisation

Recapitulating, the Laplace force has been initially defined as the  $\varepsilon \rightarrow 0$  limit of the semi-sum of the magnetic field evaluated at a distance  $\varepsilon$  away from the CWS,  $S$ , respectively inward and outward (Eq. IV.3). This was shown to either be numerically costly or subject to numerical errors (Fig. IV.1).

An expression (Eqs. IV.14-IV.17) has then been derived in Theorem IV.2 for the Laplace force exerted by one current-sheet on another, per unit length. The special case  $\mathbf{j}_1 = \mathbf{j}_2$  describes the self-interaction of a current-sheet.

Both treatments relied on an intrinsically 2D model for the currents on the CWS. A third approach is to treat the CWS as a 3D layer of infinitesimal thickness  $\varepsilon$ . For some  $\mathbf{y} \in S$  and if  $\mathbf{j}$  is smooth enough, we could compute the  $\varepsilon \rightarrow 0$  limit of

$$\tilde{\mathbf{L}}_\varepsilon(\mathbf{j}_\varepsilon)(\mathbf{y}) = \int_{-\varepsilon/2}^{\varepsilon/2} [\mathbf{j}_\varepsilon(\mathbf{y} + \varepsilon_1 \mathbf{n}(\mathbf{y})) \times \mathbf{B}(\mathbf{y} + \varepsilon_1 \mathbf{n}(\mathbf{y}))] d\varepsilon_1.$$

Note that  $\mathbf{B}(\mathbf{y} + \varepsilon_1 \mathbf{n}(\mathbf{y}))$  is well-defined as we integrate on a 3D domain, and is given by:

$$\mathbf{B}(\mathbf{y} + \varepsilon_1 \mathbf{n}(\mathbf{y})) = \int_S \int_{-\varepsilon/2}^{\varepsilon/2} [\mathbf{j}_\varepsilon(\mathbf{x} + \varepsilon_2 \mathbf{n}(\mathbf{x})) \times \frac{\mathbf{y} - \mathbf{x} + \varepsilon_1 \mathbf{n}(\mathbf{y}) - \varepsilon_2 \mathbf{n}(\mathbf{x})}{|\mathbf{y} - \mathbf{x} + \varepsilon_1 \mathbf{n}(\mathbf{y}) - \varepsilon_2 \mathbf{n}(\mathbf{x})|^3}] dS(\mathbf{x}) d\varepsilon_2.$$

In order to approximate the 3D volume with a 2D current-sheet, we suppose that  $\forall z \in S$  and  $\forall \varepsilon'$ , it is  $\mathbf{j}_\varepsilon(\mathbf{z} + \varepsilon' \mathbf{n}(\mathbf{z})) = \frac{\mathbf{j}(\mathbf{z})}{\varepsilon}$ . Thus,

$$\tilde{\mathbf{L}}_\varepsilon(\mathbf{j}_\varepsilon) = \frac{1}{\varepsilon^2} \int_{-\varepsilon/2}^{\varepsilon/2} \int_{-\varepsilon/2}^{\varepsilon/2} \left\{ \int_S \mathbf{j}(\mathbf{y}) \times [\mathbf{j}(\mathbf{x}) \times \frac{\mathbf{y} - \mathbf{x} + \varepsilon_1 \mathbf{n}(\mathbf{y}) - \varepsilon_2 \mathbf{n}(\mathbf{x})}{|\mathbf{y} - \mathbf{x} + \varepsilon_1 \mathbf{n}(\mathbf{y}) - \varepsilon_2 \mathbf{n}(\mathbf{x})|^3}] dS(\mathbf{x}) \right\} d\varepsilon_2 d\varepsilon_1$$

The quantity inside the brackets is very close to the one we got in Theorem IV.2, starting with Eqs. IV.2 and IV.3, except that we also have a contribution from  $\varepsilon_2 \mathbf{n}(\mathbf{x})$ . It is possible to prove, using an argument similar to Lemma IV.9, that replacing  $\mathbf{n}(\mathbf{x})$  with  $\mathbf{n}(\mathbf{y})$  does not change the limit. The intuition is that for  $\mathbf{x}$  close to  $\mathbf{y}$ ,  $\mathbf{n}(\mathbf{x})$  is close to  $\mathbf{n}(\mathbf{y})$ . As a result,  $\tilde{\mathbf{L}}_\varepsilon(\mathbf{j})$  has the same limit as  $\mathbf{L}_\varepsilon(\mathbf{j})$  and the expression we found for the Laplace force (Eqs. IV.14-IV.17) is consistent.

## IV.3 Examples of cost functions

After having rigorously defined the Laplace force-density  $\mathbf{L}(\mathbf{j})(\mathbf{y})$  that a current-sheet of density  $\mathbf{j}$  exerts on itself at location  $y$  (Eqs. IV.14-IV.17 for  $\mathbf{j}_1 = \mathbf{j}_2 = \mathbf{j}$ ), we now introduce some cost-functions to penalize high values of the force.

Two main options are possible, and considered here: (1) penalizing high cumulative (or, equivalently, surface-averaged) forces, or (2) penalizing or even forbidding excessively high local maxima of the force. Further variants are possible for specific force-components (e.g. tangential or normal to the CWS) or a weighted combination of them, with higher weights assigned to the engineeringly more demanding component, depending on the specific stellarator design. Such variants go beyond the scope of the present chapter, and are left for future work.

A natural choice from the functional analysis point of view is to use a penalization of the form

$$|\mathbf{L}(\mathbf{j})|_{L^p(S, \mathbb{R}^3)} = \left( \int_S |\mathbf{L}(\mathbf{j})(\mathbf{x})|^p d\mathbf{x} \right)^{1/p}. \quad (\text{IV.18})$$

The case  $p = 2$  is well-known: it represents the cumulative (or, barring a factor, the surface-averaged) root-mean-square force. Higher values of  $p$  penalize more severely high values of the



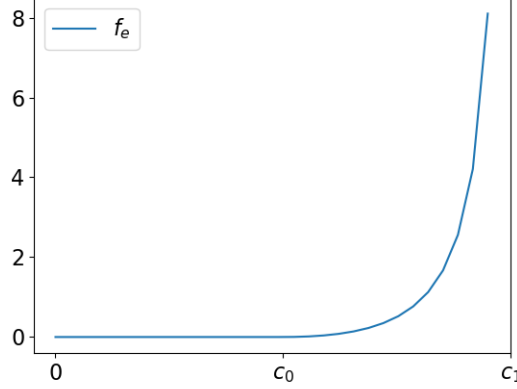


Figure IV.2 – Plot of the local cost  $f_e$  as a function of the local force  $w$ . Note that  $f_e$  diverges at  $c_1$  and vanishes in  $[0, c_0]$ . In other words, the force is non-linearly optimized: small values are permitted, intermediate ones are increasingly, non-linearly penalized, and large ones are forbidden.

Laplace force (i.e., large oscillations around the average norm). By contrast, low values of  $p$  penalize the average norm of the Laplace force.

In principle it is also possible to use a  $L^\infty$  cost,  $\sup_S |\mathbf{L}(\mathbf{j})|$ , but the domain might be smaller than  $\mathfrak{X}^{1,2}(S)$ . However, such cost is not differentiable whenever the maximum is reached at multiple locations.

The second option is to introduce the cost

$$C_e(j) = \int_S f_e(\mathbf{L}(\mathbf{j})(\mathbf{x})) d\mathbf{x} \quad (\text{IV.19})$$

as the surface integral of the local cost

$$f_e(w) = \frac{\max(w - c_0, 0)^2}{1 - \frac{\max(w - c_0, 0)}{c_1 - c_0}}. \quad (\text{IV.20})$$

The domain for this cost is not the entire space  $\mathfrak{X}^{1,2}(S)$ , but this cost captures more effectively the engineering constraints of building a high-field stellarator: the mechanical properties of support-structures and materials are such that forces below a threshold  $c_0$  are negligible, forces higher and higher than  $c_0$  should be penalized more and more, and forces above a second, “rupture” threshold  $c_1$  should be completely forbidden. Indeed, the local cost  $f_e$  evolves with the local force  $w$  as desired, as illustrated in Fig. IV.2.

**Remark IV.4.** *It is unclear whether a minimiser exists in  $\mathfrak{X}^{1,2}(S)$  for the costs discussed. As a consequence, a good practice is to add a regularizing term  $|\mathbf{j}|_{\mathfrak{X}^{1,2}(S)}$ .*

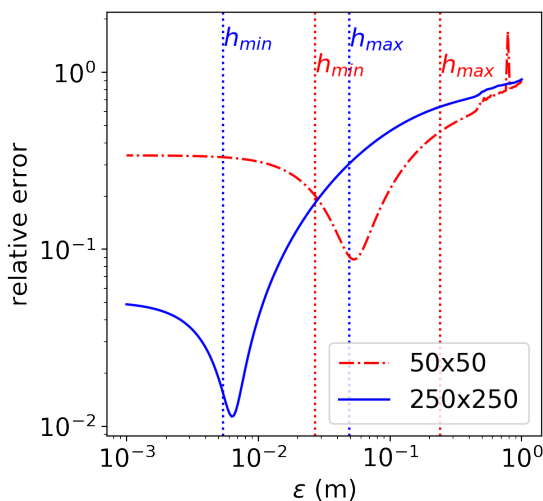


Figure IV.3 – Convergence of  $\mathbf{L}_\varepsilon$  toward  $\mathbf{L}$  for NCSX, for two different grids. Convergence stops when  $\varepsilon \lesssim h$ , due to a numerical error in  $\mathbf{B}$  (Fig. IV.1).

## IV.4 Numerical simulations

### IV.4.1 Setup

To test our force-reduction method, we ran simulations for the NCSX stellarator equilibrium known as LI383 [Zar+01]: since our work is so closely related to REGCOIL, we adopted the same data-set, for ease of comparison with the original REGCOIL paper [Lan17].

As mentioned in the Introduction, the costs defined in Sec. IV.3 are easily added to the cost-function in any stellarator coil optimization code. In our case such code was a new incarnation of REGCOIL, which we rewrote in python instead of fortran, and compiled with the Just In Time compiler Numba [LPS15]. For the most part the new code is conceptually identical to REGCOIL, except that it uses Eq. A5 of Ref. [Lan17] in lieu of its normal, single-valued component (Eq. A8 from the same paper). Eq. A5 would be numerically unstable if derivatives were taken by finite differences, but can be used here because we compute the derivatives explicitly. We compared results from the new code for LI383 and found them to agree with publicly available results from the original REGCOIL for the same case [Lan17] to within 7 significant digits.

The surface-current  $\mathbf{j}$  is divergence-free and thus taken in the form

$$G \frac{\partial \mathbf{r}'}{\partial \theta} - I \frac{\partial \mathbf{r}'}{\partial \zeta} + \frac{\partial \Phi'}{\partial \zeta'} \frac{\partial \mathbf{r}'}{\partial \theta'} - \frac{\partial \mathbf{r}'}{\partial \zeta'} \frac{\partial \Phi'}{\partial \theta'}. \quad (\text{IV.21})$$

Here  $\theta$  and  $\zeta$  are the poloidal and toroidal angle,  $G$  and  $I$  are optimization inputs (net poloidal and toroidal currents) and the current potential  $\Phi$  is decomposed in a 2D Fourier basis.

Fig. IV.3 illustrates how  $\mathbf{L}_\varepsilon$  converges to  $\mathbf{L}$  (or, equivalently, the relative error vanishes) as  $\varepsilon \rightarrow 0$ . We recall that the numerical evaluation of  $\mathbf{L}_\varepsilon$  involves three characteristic distances  $h$  (discretisation length of the mesh),  $\varepsilon$ , and  $d_B$  (characteristic distance of variation of the magnetic field). For reference we computed  $\mathbf{L}$  on the same mesh (that is, for the same  $h$ ), and obviously  $d_B$  was also the same.

We observe that the error decreases with  $\varepsilon$ , as expected, but when  $\varepsilon \lesssim h$  the convergence

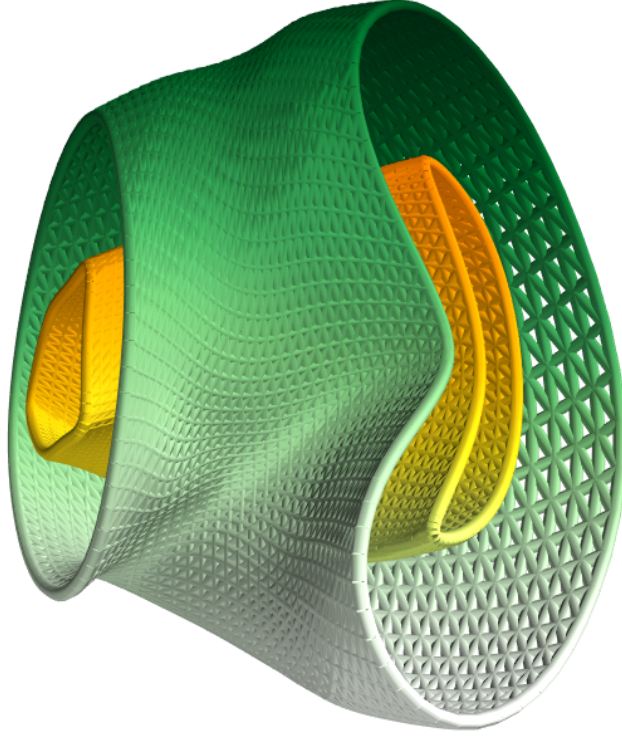


Figure IV.4 – NCSX LI383 plasma surface in orange-yellow and CWS in green-white, for a half period. The triangular mesh is only used for rendering; the actual calculations were carried out on a  $64 \times 64$ , poloidal $\times$ toroidal mesh.

stops and the error grows again. This is a consequence of the calculation of  $\mathbf{B}$  not being accurate anymore, for  $\varepsilon \gtrsim h$  (see Fig. IV.1). Note that the reference value itself is an approximation. As we do not have an analytic expression, we also used a discretisation. The relative error is computed in  $L^2$  norm.

In all simulations presented here, a period of both the CWS and the plasma surface were discretized as  $64 \times 64$  meshes in the poloidal $\times$ toroidal direction. Half-periods of the two surfaces are rendered in 3D in Fig. IV.4.

The single-valued current potential  $\Phi$  from which  $\mathbf{j}$  descends is represented by 8 or 12 harmonics in each direction. As we do not impose stellarator symmetry, we use as a basis the functions

$$\sin(k\theta + l\zeta), \quad \cos(k\theta + l\zeta)$$

with  $0 \leq k \leq N$  and  $-N \leq l \leq N$ . Since for  $k = 0$  we can restrict to  $0 < l$ , the total number of Degrees Of Freedom (DOF) is  $2[(2N + 1)N + N]$ .

Thus  $N = 8$  harmonics in each direction correspond to 288 DOF, and  $N = 12$  yields 624 DOF. Better results can be achieved with more harmonics, as shown in Fig. IV.5. However, a finer

mesh is required, making the problem computationally more expensive.

The optimization is performed by conjugate gradient. With our implementation, a single evaluation of the gradient lasts approximately 2 minutes on a small cluster of 64 cores. The full optimization can last a few days.

### IV.4.2 Adding force minimization and improving regularization in REGCOIL

We propose to integrate the costs introduced above in the same optimization scheme as NESCOIL [Mer87] and REGCOIL [Lan17].

As a reminder, NESCOIL seeks the current of density  $\mathbf{j}$ , on a fixed  $S$ , that maximizes magnetic field accuracy on the plasma boundary  $S_P$  (hence, indirectly, in the plasma). It does so by minimizing the “plasma-shape objective” or “field accuracy objective”

$$\chi_B^2 = \int_{S_P} \langle \mathbf{B}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) \rangle^2 dS(\mathbf{x}). \quad (\text{IV.22})$$

REGCOIL, instead, compromises between field accuracy and coil simplicity by minimizing  $\chi_B^2 + \lambda \chi_j^2$ , where  $\lambda$  is a weight and the “current-density objective” or “regularizing term”  $\chi_j^2$  is a penalty on high values of  $\mathbf{j}$ , in the sense of the  $L^2$  norm:

$$\chi_j^2 = \int_S |\mathbf{j}|^2 dS. \quad (\text{IV.23})$$

Heavier weighting makes  $\Phi$  (hence  $\mathbf{j}$ , hence the coils) more regular, but at the expense of reduced field accuracy. Such cost is identical to  $\chi_K^2$  of Ref. [Lan17], but is renamed  $\chi_j^2$  for consistency of notation with another regularizing term that we need to introduce:

$$\chi_{\nabla \mathbf{j}}^2 = \int_S (|\nabla \mathbf{j}_x|^2 + |\nabla \mathbf{j}_y|^2 + |\nabla \mathbf{j}_z|^2) dS. \quad (\text{IV.24})$$

This new term is motivated by Theorem IV.2: as the Laplace force can only be defined for  $\mathbf{j} \in \mathfrak{X}^{1,2}(S)$ , it is natural to add a penalization on the gradient of  $\mathbf{j}$  and not just on  $\mathbf{j}$ . Basically we are replacing the  $L^2$  norm of  $\mathbf{j}$  with the  $H^1$  norm of  $\mathbf{j}$ .

Here we propose to further generalize the REGCOIL cost function to

$$\chi^2 = \chi_B^2 + \lambda_1 \chi_j^2 + \lambda_2 \chi_{\nabla \mathbf{j}}^2 + \gamma \chi_F^2, \quad (\text{IV.25})$$

where  $\chi_F^2$  is a “force objective” that penalizes strong forces on the current-sheet, i.e. among the coils. Per the discussion in Sec. IV.3, possible definitions include:

$$\chi_F^2 = |\mathbf{L}(\mathbf{j})|_{L^2(S, \mathbb{R}^3)}^2 = \int_S |\mathbf{L}(\mathbf{j})|^2 dS \quad (\text{IV.26})$$

$$\chi_F^2 = C_e = \int_S f_e(|\mathbf{L}(\mathbf{j})|) dS \quad (\text{IV.27})$$

with  $f_e$  defined as in Eq. IV.20 and plotted in Fig. IV.2. As stress limits, here we set  $c_0 = 5 \cdot 10^6$  Pa and  $c_1 = 10^7$  Pa.

### IV.4.3 Numerical results

There is obviously a trade-off between conflicting objectives in Eq. IV.25, or special cases of that equation. Special cases include the REGCOIL-like minimization of  $\chi^2 = \chi_B^2 + \lambda_1 \chi_j^2$  and force-minimization without regularization ( $\chi^2 = \chi_B^2 + \gamma \chi_F^2$ ).

In the REGCOIL-like case (*curves* in Fig. IV.5) we fixed  $\lambda_2 = \gamma = 0$  and minimized  $\chi_B^2 + \lambda_1 \chi_j^2$  for various choices of  $\lambda_1$ . By this scan we re-obtained the well-known trade-off between  $\chi_B^2$  and  $\chi_j^2$  (or, equivalently, field-accuracy and coil-simplicity) [Lan17], but do not plot it for brevity. Interestingly, we also found a trade-off between  $\chi_B^2$  and  $\chi_F^2$ , even though  $\chi_F^2$  was not part of the  $\chi_B^2 + \lambda_1 \chi_j^2$  minimization. In other words, more accurate fields come at the expense of higher forces, even when forces are not accounted in the minimization. The trade-off between these global quantities is plotted in Fig. IV.5a, and a trade-off between related, local quantities is plotted in Fig. IV.5b. This can be explained as follows. Accumulation of currents (high  $\chi_j^2$ ), e.g. due to complicated patterns, typically leads to accumulation of forces (high  $\chi_F^2$ ) because closer current-filaments exert stronger forces onto each other. This correlation between  $\chi_F^2$  and  $\chi_j^2$ , combined with the well-known anti-correlation between  $\chi_j^2$  and  $\chi_B^2$  [Lan17] implies that  $\chi_F^2$  anti-correlates with  $\chi_B^2$ .

In the force-minimization case, instead, we fixed  $\lambda_1 = \lambda_2 = 0$  and minimized  $\chi_B^2 + \gamma \chi_F^2$  for various choices of  $\gamma$ . Not surprisingly, we found a trade-off between  $\chi_B^2$  and  $\chi_F^2$  (*symbols* in Fig. IV.5). Interestingly,  $\chi_B^2$  also exhibits a trade-off with  $\chi_j^2$ , in spite of the latter not being part of the minimization. This suggests that  $\chi_F$  has a regularizing effect on  $\mathbf{j}$ , as it will become apparent in Fig. IV.7 and IV.8.

Finally, Fig. IV.5 confirms that a higher number of Fourier harmonics and hence of DOF reproduces the magnetic field more accurately. This is why for the remainder of the chapter we adopt the higher number of DOF, 624.

Also, we no longer scan the weights, but fix them to yield reasonable compromises between field accuracy, current regularization and/or force minimization. In particular, calculations were performed for the following four choices of weights and  $\chi_F$  in Eq. IV.25:

Case	$\lambda_1$ (T <sup>2</sup> m <sup>2</sup> /A <sup>2</sup> )	$\lambda_2$ (T <sup>2</sup> m <sup>4</sup> /A <sup>2</sup> )	$\gamma$ (T <sup>2</sup> /Pa <sup>2</sup> )	$\chi_F$	
1	$1.5 \cdot 10^{-16}$	0	0	0	(IV.28)
2	0	0	$10^{-17}$	$ \mathbf{L}(\mathbf{j}) _{L^2(S, \mathbb{R}^3)}^2$	
3	0	0	$10^{-16}$	$C_e$	
4	$10^{-19}$	$10^{-19}$	$10^{-16}$	$C_e$	

Case 1 is basically REGCOIL, whereas case 2 and 3 are effectively NESCOIL but with minimized forces, according to two different force metrics. Finally, case 4 explicitly combines force minimization with regularization, but in a broader sense compared to REGCOIL, as discussed in connection with Eq. IV.24. The force metric for case 2 penalizes high root mean squared surface-averaged forces (Eq. (IV.26)), whereas the metric for cases 3 and 4 non-linearly penalizes high local forces (Eq. (IV.27)).

The results for these four cases are plotted in Fig. IV.6 (circles) and compared with REGCOIL results (curve). In particular Fig. IV.6a refers to surface-integrated, “global” objectives, and Fig. IV.6 to “local” maxima. Note the logarithmic plots. As expected, case 1 agrees with REGCOIL. Case 2 (defined in terms of the “global”  $|\mathbf{L}(\mathbf{j})|_{L^2}^2$ ) overperforms in the “global” Fig. IV.6a, as expected. Actually, it performs better than REGCOIL even in terms of local metrics (Fig. IV.6b). Compared to REGCOIL, peak-forces are reduced in cases 3 and 4 (Fig. IV.6b), and remain lower

than the chosen  $c_1$ , as is expected from the definition of  $C_e$  and  $f_e$  (Eqs. IV.19-IV.20) However, this happens at the expense of higher cumulative forces (Fig. IV.6a).

Details on the four cases are presented in Fig. IV.7 and IV.8. Columns from left to right refer to cases from 1 to 4. From top to bottom, the rows in Fig. IV.7 present contours of

1. the norm of  $\mathbf{j}$ , related to  $\chi_{\mathbf{j}}^2$  and  $\chi_{\nabla\mathbf{j}}^2$ ,
2. the magnetic field normal to  $S$ , related to  $\chi_B^2$ , and
3. the norm of the Laplace force per unit surface, related to  $\chi_F^2$ .

The two rows in Fig. IV.8 present the force components normal and tangential to  $S$ . All quantities are plotted as functions of the poloidal and toroidal angles.

As anticipated, case 2 is as regular as case 1, in spite of its  $\chi^2$  not containing a regularizing objective. By contrast, case 3 reproduces the field with high accuracy and exhibits reduced peak forces, as expected from the definition of  $C_e$ , but with a complicated current-pattern. That is ameliorated by adding some regularization: case 4 is the best compromise between coil simplicity (first row in Fig. IV.8), field accuracy (second row) and reduced forces (third row).

Incidentally all cases, including REGCOIL (case 1) and the magnetically most accurate case 3, exhibit residual field errors of up to 60 mT. Lower errors can be achieved by adopting a higher number of DOF, as is intuitive and suggested by Fig. IV.5, but this is computationally more intensive and beyond the scope of the present chapter.

From the point of view of the surface-integrated or surface-averaged forces, the best result in Fig. IV.7 is a modest reduction by 5% for case 2, relative to REGCOIL. From the point of view of peak forces, however, the best result in Fig. IV.7 is a reduction by 40% for case 4, relative to REGCOIL. Correspondingly, the peak tangential force is reduced by 50% and the peak normal force by 20% (Fig. IV.8). Note that maxima for different components occur at different toroidal and poloidal locations.

More dramatic reductions were obtained in Fig. IV.5, especially in peak forces. However, they were obtained for low-accuracy cases on the top left of Fig. IV.5b: a stellarator with those characteristics would suffer from very low coil-forces, but it would also be a poor match of the target field.

There is some arbitrariness in how to discretize the continuous current-distributions of Fig. IV.7. Fig IV.10 illustrates possible filamentations for cases 1 and 4. Note the accumulation of current filaments, i.e. coils, in regions of high forces (the color contours in the background). This reflects the fact that, by definition, current filaments tend to “crowd” in regions of high  $\mathbf{j}$ , and this proximity results in high forces. However, for different discretizations they are subject to different forces. This offers additional degrees of freedom for force-minimization, which are left for future work.

## IV.5 Summary, conclusions and future work

To summarize, force-minimization is an important aspect of stellarator coil-optimization, especially for future high-field stellarators. In the present chapter we rigorously proved in Sec. IV.2.3 that the Laplace force exerted by a surface-current onto one another can be written as in Eqs. IV.14-IV.17. From that, one can calculate the auto-interaction  $\mathbf{L}(\mathbf{j})$  of a current-distribution with itself, and distill that information in a single scalar. Possible metrics were discussed in Sec. IV.3, and two of them were used for detailed numerical calculations: two possible “force objectives” (Eqs. IV.26 and IV.27) were added to the cost function of the well-known REGCOIL code [Lan17]. In addition, the  $L^2$  norm of  $\mathbf{j}$  was replaced with the  $H^1$  norm of  $\mathbf{j}$ , for reasons explained in Secs. IV.2.1 and IV.4.2.

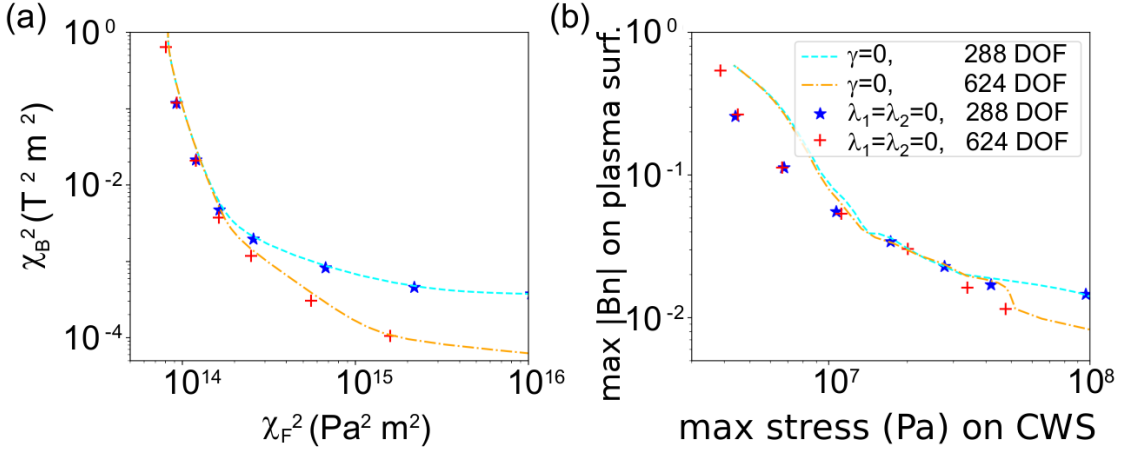


Figure IV.5 – (a) Trade-off between plasma shape accuracy and Laplace force metrics (as defined in Eq. IV.26) for different weightings in Eq. IV.25 and different numbers of harmonics, and thus of Degrees of Freedom (DOF). Such trade-off, expected when optimizing a linear combination of  $\chi_B^2$  and  $\chi_F^2$  (symbols), is also observed in the minimization of  $\chi_B^2$  and  $\chi_j^2$  (curves). (b) Similar trade-off between maximum field and maximum Laplace force (Eq. IV.27).

This approach permitted to simultaneously optimize the coils of the NCSX stellarator for magnetic fidelity, regularity and low forces. For instance, 40% lower peak-forces were obtained compared to REGCOIL, for similar plasma-shape accuracy and current regularity. These results were presented as case 1 and 4 in Figs. IV.6-IV.8. Force reduction is an important criterion in stellarator optimization, and future high-field designs might benefit from our approach.

Unfortunately force minimization made the new approach (case 4) significantly slower than REGCOIL (case 1). This motivated the adoption of a low number of Fourier harmonics and thus of Degrees of Freedom (DOF) in case 4 and, for consistency, in all cases. The resulting field inaccuracies are high, but they are just as high with REGCOIL, under the same circumstances (Fig. IV.7). Fortunately, Ref. [Lan17] and Fig. IV.5 indicate that such inaccuracies rapidly disappear by adopting more DOF, and just as rapidly in our code as for REGCOIL. At the same time, more DOF lead to more coil-simplicity [Lan17] and lower coil-forces (Fig. IV.5). In the future, optimizing the code for speed and/or running it on a super-computer will allow to retain a higher number of DOF.

In the present chapter we introduced (Fig. IV.2) and successfully demonstrated (Fig. IV.5) the non-linear optimization of the coil-forces: we introduced constraints  $c_0$  and  $c_1$  to allow stresses below  $c_0$ , increasingly penalize stresses in the  $[c_0, c_1]$  range, and forbid stresses above  $c_1$ . Future work could impose stricter constraints and tailor them differently for normal and longitudinal forces, as they tend to differ (Fig. IV.8) and they obey to different engineering and material constraints. In addition, we could non-linearly optimize other quantities. For example we could allow field inaccuracies of one part in 10<sup>5</sup>, penalize inaccuracies up to one part in 10<sup>4</sup>, and attribute infinite cost to larger discrepancies.

Finally, in the present work the CWS was fixed. Future shape optimization of the CWS, inspired by Ref. [Pau20], is expected to further reduce the coil-forces.

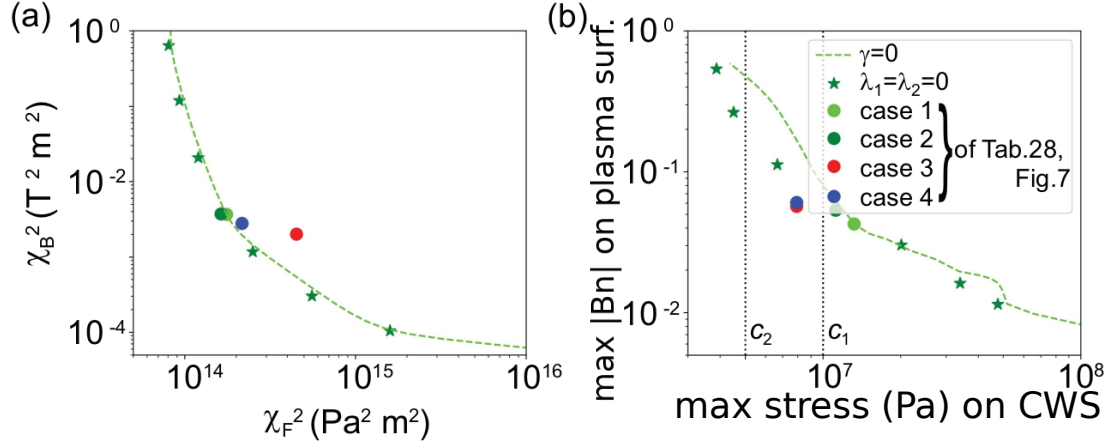


Figure IV.6 – Trade-offs between: (a) plasma shape accuracy and Laplace force metrics (Eq. IV.26) and (b) maximum field and maximum Laplace force (Eq. IV.27). Unlike Fig. IV.5, all simulations here used 624 DOF. Circle symbols correspond to the four cases discussed in Sec. IV.4 and presented in Fig. IV.7. As expected, the  $C_e$  cases (red and blue) fall between the penalized and forbidden forces  $c_0$  and  $c_1$  defined in Fig. IV.2, marked here by vertical dotted lines.

## IV.A Appendix

### IV.A.1 First tangential term

Integration by parts on a compact manifold  $\mathcal{M}$  without boundary is given by the following formula. Let  $f \in C^\infty(\mathcal{M})$  and  $\mathbf{X}$  a smooth vector field on  $\mathcal{M}$ , then

$$\int_{\mathcal{M}} \operatorname{div}(f\mathbf{X}) = 0 = \mathbf{X}f + \int_{\mathcal{M}} f \operatorname{div} \mathbf{X} \quad (\text{IV.29})$$

We also recall that  $\mathbf{X}f = \langle \mathbf{X} \cdot \nabla f \rangle$  in Euclidean coordinates.

Let us start with the first tangential term (Eq. IV.8):

$$\begin{aligned} & \int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \nabla_S \frac{1}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|} \rangle_{\mathbb{R}^3} \mathbf{j}_2(\mathbf{x}) d\mathbf{x} \\ &= \int_S \langle \pi_{\mathbf{x}} \mathbf{j}_1(\mathbf{y}) \cdot \nabla_S \frac{1}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|} \rangle_{T_{\mathbf{x}} S} \mathbf{j}_2(\mathbf{x}) d\mathbf{x} \end{aligned}$$

as  $\mathbf{j}_1(\mathbf{y}) - \pi_{\mathbf{x}} \mathbf{j}_1(\mathbf{y}) \propto \mathbf{n}(\mathbf{x})$ .

Then, let  $j_2^i(\mathbf{x})$  be the  $i$ -th component in  $\mathbb{R}^3$  of  $\mathbf{j}_2$ . Using integration by parts (Eq. IV.29),



the  $i$ -th component of the last integral writes

$$\int_S \langle j_2^i(\mathbf{x}) \pi_{\mathbf{x}} \mathbf{j}_1(\mathbf{y}) \cdot \nabla_S \frac{1}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|} \rangle_{T_x S} d\mathbf{x} \quad (\text{IV.30})$$

$$= - \int_S \frac{1}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|} \operatorname{div}_{\mathbf{x}}(j_2^i(\mathbf{x}) \pi_{\mathbf{x}} \mathbf{j}_1(\mathbf{y})) d\mathbf{x} \quad (\text{IV.31})$$

$$= - \int_S \frac{1}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|} [j_2^i(\mathbf{x}) \operatorname{div}_{\mathbf{x}}(\pi_{\mathbf{x}} \mathbf{j}_1(\mathbf{y})) + \langle \pi_{\mathbf{x}} \mathbf{j}_1(\mathbf{y}) \cdot \nabla j_2^i(\mathbf{x}) \rangle] d\mathbf{x} \quad (\text{IV.32})$$

Thus the term in equation IV.5 is equal to:

$$- \sum_{i=1}^3 \left( \int_S \frac{1}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|} [j_2^i(\mathbf{x}) \operatorname{div}_{\mathbf{x}}(\pi_{\mathbf{x}} \mathbf{j}_1(\mathbf{y})) + \langle \pi_{\mathbf{x}} \mathbf{j}_1(\mathbf{y}) \cdot \nabla j_2^i(\mathbf{x}) \rangle] d\mathbf{x} \right) \mathbf{e}_i$$

that, with the conventions introduced in Remark IV.3, can be rewritten as:

$$- \int_S \frac{1}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|} [\operatorname{div}_{\mathbf{x}}(\pi_{\mathbf{x}} \mathbf{j}_1(\mathbf{y})) + \pi_{\mathbf{x}} \mathbf{j}_1(\mathbf{y}) \cdot \nabla_{\mathbf{x}}] \mathbf{j}_2(\mathbf{x}) d\mathbf{x}$$

Now, we will prove that it is possible to take the limit  $\varepsilon \rightarrow 0$  inside this integral. The first step is to use the following estimate.

**Lemma IV.5.**  $|\operatorname{div}_{\mathbf{x}} \pi_{\mathbf{x}} \mathbf{j}_1(\mathbf{y})| \leq C(S) |\mathbf{j}_1(\mathbf{y})|$  with  $C(S)$  a constant that only depends on the metric of  $S$ .

*Proof.* Indeed, the application

$$\begin{aligned} \operatorname{div}_{\mathbf{x}} \pi_{\mathbf{x}} : \mathbb{R}^3 &\rightarrow \mathcal{C}^\infty(S) \\ \mathbf{v} &\mapsto (\mathbf{x} \mapsto \operatorname{div}_{\mathbf{x}}(\pi_{\mathbf{x}} \mathbf{v})). \end{aligned}$$

is a continuous linear application.  $\square$

We also need a Young-type inequality for 2-dimensional compact manifolds.

**Lemma IV.6.** For all  $1 \leq q < \infty$ , there exists  $C > 0$  such that for all  $f$  in  $L^2(S)$ ,

$$\left| \int_S \frac{1}{|\mathbf{y} - \mathbf{x}|} f(\mathbf{x}) d\mathbf{x} \right|_{L_y^q} \leq C_q |f|_{L^2}.$$

*Proof.* Let  $d_g$  denote the Riemannian distance on  $S$ . By a Hardy-Littlewood-Sobolev inequality  $\int_S \frac{1}{d_g(\mathbf{y}, \mathbf{x})} f(x) d\mathbf{x}$  is in  $L^q(S)$  for all  $1 \leq q < \infty$ . This result can be found, for example, in [HZ16] or can be proved directly with the arguments of the proof of the classical Young inequality. As the Euclidean distance and the Riemannian distance are equivalent, the lemma is proved.  $\square$

Thus, for all  $1 \leq q < \infty$ ,  $\int_S \frac{1}{|\mathbf{y} - \mathbf{x}|} \partial_i \mathbf{j}_2(\mathbf{x}) d\mathbf{x} \in L^q(S, \mathbb{R}^3)$ . Besides, by Sobolev embedding [Heb00], there is a continuous injection  $\mathfrak{X}^{1,2}(S) \hookrightarrow \mathfrak{X}^p(S)$  for all  $1 \leq p < \infty$ . As a consequence  $j_1^i(\mathbf{y}) \int_S \frac{1}{|\mathbf{y} - \mathbf{x}|} \partial_i \mathbf{j}_2(\mathbf{x}) d\mathbf{x} \in \mathfrak{X}^p(S)$  for  $1 \leq p < \infty$ .

With these estimates we can conclude, using dominated convergence, that:

$$\int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \nabla_S \frac{1}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|} \rangle \mathbf{j}_2(\mathbf{x}) d\mathbf{x} \quad (\text{IV.33})$$

$$\xrightarrow{\mathfrak{X}^p(S)} - \int_S \frac{1}{|\mathbf{y} - \mathbf{x}|} \operatorname{div}_{\mathbf{x}}(\pi_{\mathbf{x}} \mathbf{j}_1(\mathbf{y})) \mathbf{j}_2(\mathbf{x}) d\mathbf{x} \quad (\text{IV.34})$$

$$- \int_S \frac{1}{|\mathbf{y} - \mathbf{x}|} (\pi_{\mathbf{x}} \mathbf{j}_1(\mathbf{y}) \cdot \nabla_{\mathbf{x}}) \mathbf{j}_2(\mathbf{x}) d\mathbf{x} \quad (\text{IV.35})$$

Note that the two integrals (respectively with the sign + or - at denominator) converge to the same limit. Their sum yields the integral on the right-hand-side of Eq. IV.14.

### IV.A.2 Second tangential term

Now, let us tackle the term in Eq. IV.12. We start by computing the  $i$ -th component of that integral, i.e. its projection on  $\mathbf{e}_i$ , then follow a derivation similar to Eqs. IV.30-IV.32:

$$\begin{aligned} & \int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{j}_2(\mathbf{x}) \rangle \langle \mathbf{e}_i \cdot \nabla_S \frac{1}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|} \rangle d\mathbf{x} \\ &= \int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{j}_2(\mathbf{x}) \rangle \langle \pi_{\mathbf{x}} \mathbf{e}_i \cdot \nabla_S \frac{1}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|} \rangle d\mathbf{x} \\ &= - \int_S \frac{1}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|} \operatorname{div}_{\mathbf{x}}(\langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{j}_2(\mathbf{x}) \rangle \pi_{\mathbf{x}} \mathbf{e}_i) d\mathbf{x} \\ &= - \int_S \frac{1}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|} [\langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{j}_2(\mathbf{x}) \rangle \operatorname{div}_{\mathbf{x}}(\pi_{\mathbf{x}} \mathbf{e}_i) + \langle \pi_{\mathbf{x}} \mathbf{e}_i \cdot \nabla_{\mathbf{x}} \langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{j}_2(\mathbf{x}) \rangle \rangle] d\mathbf{x}. \end{aligned}$$

Using the notation of Remark IV.3, we find the vector form of the integral:

$$- \int_S \frac{1}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|} (\langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{j}_2(\mathbf{x}) \rangle \operatorname{div}_{\mathbf{x}}(\pi_{\mathbf{x}}) + \nabla_{\mathbf{x}} \langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{j}_2(\mathbf{x}) \rangle) d\mathbf{x}.$$

Due to the same arguments invoked in Eqs. IV.33-IV.35, both integrals, with the sign + and - at denominator, converge in  $\mathfrak{X}^p(S)$  to the same limit,

$$- \int_S \frac{1}{|\mathbf{y} - \mathbf{x}|} [\langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{j}_2(\mathbf{x}) \rangle \operatorname{div}_{\mathbf{x}}(\pi_{\mathbf{x}}) + \nabla_{\mathbf{x}} \langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{j}_2(\mathbf{x}) \rangle] d\mathbf{x}.$$

Their sum yields integral IV.16.

### IV.A.3 First normal term

Let us now focus on the normal component of IV.4, namely Eq. IV.9. This is in effect the sum of two integrals, which we will discuss separately:

$$\int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{n}(\mathbf{x}) \rangle \frac{\langle \mathbf{y} - \mathbf{x}, \mathbf{n}(\mathbf{x}) \rangle}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|^3} \mathbf{j}_2(\mathbf{x}) d\mathbf{x} \quad (\text{IV.36})$$

$$\int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{n}(\mathbf{x}) \rangle \frac{\pm \varepsilon \langle \mathbf{n}(\mathbf{y}), \mathbf{n}(\mathbf{x}) \rangle}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|^3} \mathbf{j}_2(\mathbf{x}) d\mathbf{x} \quad (\text{IV.37})$$

First notice that we have the following estimate:

**Lemma IV.7.**  $\exists C > 0, \forall \mathbf{x} \neq \mathbf{y} \in S, \frac{|\langle \mathbf{y} - \mathbf{x}, \mathbf{n}(\mathbf{x}) \rangle|}{|\mathbf{y} - \mathbf{x}|^2} \leq C.$

*Proof of Lemma IV.7.* Let us suppose there exist two sequences  $(\mathbf{x}_n), (\mathbf{y}_n)$  in  $S$  such that  $\mathbf{x}_n \neq \mathbf{y}_n$  and  $\frac{|\langle \mathbf{y}_n - \mathbf{x}_n, \mathbf{n}(\mathbf{x}_n) \rangle|}{|\mathbf{y}_n - \mathbf{x}_n|^2} \rightarrow \infty.$  Up to an extraction, we can suppose that  $\mathbf{x}_n \rightarrow \mathbf{x}_0 \in S.$  If  $\mathbf{y}_n$  does not converges to  $\mathbf{x}_0,$  we can extract a subsequence such that  $\frac{|\langle \mathbf{y}_n - \mathbf{x}_n, \mathbf{n}(\mathbf{x}_n) \rangle|}{|\mathbf{y}_n - \mathbf{x}_n|^2}$  does not diverge. This is a contradiction, hence both  $\mathbf{x}_n$  and  $\mathbf{y}_n$  converge to  $\mathbf{x}_0 \in S.$

Let  $\Gamma(\mathbf{x}, \mathbf{y}) = \langle \mathbf{y} - \mathbf{x}, \mathbf{n}(\mathbf{x}) \rangle.$  As  $S$  is smooth, so is  $\Gamma.$  Its partial differentials are

$$\begin{aligned} \forall \mathbf{h} \in T_x S, \quad d_x \Gamma_{\mathbf{x}, \mathbf{y}}(\mathbf{h}) &= -\langle \mathbf{h} \cdot \mathbf{n}(\mathbf{x}) \rangle + \langle \mathbf{y} - \mathbf{x} \cdot d\mathbf{n}_x(\mathbf{h}) \rangle \\ \forall \mathbf{h} \in T_y S, \quad d_y \Gamma_{\mathbf{x}, \mathbf{y}}(\mathbf{h}) &= \langle \mathbf{h} \cdot \mathbf{n}(\mathbf{x}) \rangle \end{aligned}$$

Thus, at the point  $(\mathbf{x}_0, \mathbf{x}_0),$  both first derivatives vanish. As a consequence, for  $n$  big enough there exists  $C > 0$  such that  $\Gamma(\mathbf{x}_n, \mathbf{y}_n) \leq C|\mathbf{x}_n - \mathbf{y}_n|^2,$  contradiction.  $\square$

Now, we need to find a minoration of  $|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|.$

**Lemma IV.8.** *For  $\varepsilon$  small enough, for all  $\mu > 0,$*

$$|\mathbf{x} - \mathbf{y} \pm \varepsilon \mathbf{n}(\mathbf{y})|^\mu \geq \max\left(\left(\frac{1}{\sqrt{2}}|\mathbf{x} - \mathbf{y}|\right)^\mu, \varepsilon^\mu\right).$$

*Proof of Lemma IV.8.*

$$\begin{aligned} |\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|^2 &= |\mathbf{y} - \mathbf{x}|^2 \pm \varepsilon \langle \mathbf{y} - \mathbf{x}, \mathbf{n}(\mathbf{y}) \rangle + \varepsilon^2 \\ &\geq |\mathbf{y} - \mathbf{x}|^2 - C\varepsilon |\mathbf{y} - \mathbf{x}|^2 + \varepsilon^2 \quad \text{by lemma IV.7.} \end{aligned}$$

Thus for  $\varepsilon \leq 1/(2C),$  we have,  $\forall \mu > 0,$

$$|\mathbf{x} - \mathbf{y} \pm \varepsilon \mathbf{n}(\mathbf{y})|^\mu \geq \max\left(\left(\frac{1}{\sqrt{2}}|\mathbf{x} - \mathbf{y}|\right)^\mu, \varepsilon^\mu\right).$$

$\square$

Using Lemmas IV.7 and IV.8, for some constant  $C,$   $\left|\frac{\langle \mathbf{y} - \mathbf{x}, \mathbf{n}(\mathbf{x}) \rangle}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|^3}\right|$  is dominated by  $C\frac{1}{|\mathbf{y} - \mathbf{x}|},$  which is integrable. By dominated convergence

$$\int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{n}(\mathbf{x}) \rangle \frac{\langle \mathbf{y} - \mathbf{x}, \mathbf{n}(\mathbf{x}) \rangle}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|^3} \mathbf{j}_2(\mathbf{x}) d\mathbf{x} \xrightarrow{\varepsilon \rightarrow 0} \int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{n}(\mathbf{x}) \rangle \frac{\langle \mathbf{y} - \mathbf{x}, \mathbf{n}(\mathbf{x}) \rangle}{|\mathbf{y} - \mathbf{x}|^3} \mathbf{j}_2(\mathbf{x}) d\mathbf{x}, \quad (\text{IV.38})$$

i.e. we obtained integral IV.15.

Now we have to deal with  $\frac{\varepsilon \langle \mathbf{n}(\mathbf{y}), \mathbf{n}(\mathbf{x}) \rangle}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{x})|^3},$  but we will show their net contribution to converge to zero.

To begin with, we could use the smallness of the term  $\langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{n}(\mathbf{x}) \rangle$  to ensure integrability. Instead, we will prove the following lemma which will also be useful later. Let  $\Delta = \{(\mathbf{z}, \mathbf{z}) \mid \mathbf{z} \in S\} \subset S^2.$

**Lemma IV.9.** *Let  $f_\varepsilon : S^2 \setminus \Delta \ni (\mathbf{x}, \mathbf{y}) \mapsto \frac{1}{|\mathbf{y} - \mathbf{x} + \varepsilon \mathbf{n}(\mathbf{y})|^3} - \frac{1}{|\mathbf{y} - \mathbf{x} - \varepsilon \mathbf{n}(\mathbf{y})|^3} d\mathbf{x}.$  Then  $\exists \eta > 0, \exists M > 0,$   $\forall \alpha \in (-0.5, 3.5), \forall \varepsilon < \eta, \forall (\mathbf{x}, \mathbf{y}), |\varepsilon^\alpha f_\varepsilon(\mathbf{x}, \mathbf{y})| \leq M \frac{1}{|\mathbf{x} - \mathbf{y}|^{5/2 - \alpha}}.$*

*Proof of Lemma IV.9.*

$$\begin{aligned} f_\varepsilon(\mathbf{x}, \mathbf{y}) &= \frac{|\mathbf{y} - \mathbf{x} - \varepsilon \mathbf{n}(\mathbf{y})|^3 - |\mathbf{y} - \mathbf{x} + \varepsilon \mathbf{n}(\mathbf{y})|^3}{|\mathbf{y} - \mathbf{x} + \varepsilon \mathbf{n}(\mathbf{y})|^3 |\mathbf{y} - \mathbf{x} - \varepsilon \mathbf{n}(\mathbf{y})|^3} \\ &= (|\mathbf{y} - \mathbf{x} - \varepsilon \mathbf{n}(\mathbf{y})| - |\mathbf{y} - \mathbf{x} + \varepsilon \mathbf{n}(\mathbf{y})|) \times \\ &\quad \frac{(|\mathbf{y} - \mathbf{x} + \varepsilon \mathbf{n}(\mathbf{y})|^2 + |\mathbf{y} - \mathbf{x} + \varepsilon \mathbf{n}(\mathbf{y})| |\mathbf{y} - \mathbf{x} - \varepsilon \mathbf{n}(\mathbf{y})| + |\mathbf{y} - \mathbf{x} - \varepsilon \mathbf{n}(\mathbf{y})|^2)}{|\mathbf{y} - \mathbf{x} + \varepsilon \mathbf{n}(\mathbf{y})|^3 |\mathbf{y} - \mathbf{x} - \varepsilon \mathbf{n}(\mathbf{y})|^3} \end{aligned}$$

Using the fact that square root is  $1/2$ -Hölder continuous ( $a \geq b \geq 0, \sqrt{a} - \sqrt{b} \leq \sqrt{a-b}$ ) and Lemma IV.7, there exists  $C > 0$  such that

$$||\mathbf{y} - \mathbf{x} + \varepsilon \mathbf{n}(\mathbf{y})| - |\mathbf{y} - \mathbf{x} - \varepsilon \mathbf{n}(\mathbf{y})|| \leq 2\sqrt{\varepsilon} |\langle \mathbf{x} - \mathbf{y}, \mathbf{n}(\mathbf{y}) \rangle| \leq C\sqrt{\varepsilon} |\mathbf{x} - \mathbf{y}|$$

Now we use the minoration of the denominator from Lemma IV.8. Up to a global multiplicative constant  $M$ , we get, for any  $0 \leq \nu \leq 4$ ,

$$\begin{aligned} |f_\varepsilon(\mathbf{x}, \mathbf{y})| &\leq 4C \frac{\sqrt{\varepsilon} |\mathbf{x} - \mathbf{y}|}{|\mathbf{x} - \mathbf{y}|^{4-\nu} \varepsilon^\nu} \\ &\leq M \frac{1}{\varepsilon^\alpha |\mathbf{x} - \mathbf{y}|^{5/2-\alpha}} \end{aligned}$$

for any  $-0.5 \leq \alpha \leq 3.5$  □

Thanks to Lemma IV.9 with any  $\alpha \in (1/2, 1)$ , there exists  $C > 0$  such that

$$\left| \int_S \left( \frac{\varepsilon \langle \mathbf{n}(\mathbf{y}), \mathbf{n}(\mathbf{x}) \rangle}{|\mathbf{y} - \mathbf{x} + \varepsilon \mathbf{n}(\mathbf{y})|^3} - \frac{\varepsilon \langle \mathbf{n}(\mathbf{y}), \mathbf{n}(\mathbf{x}) \rangle}{|\mathbf{y} - \mathbf{x} - \varepsilon \mathbf{n}(\mathbf{y})|^3} \right) \mathbf{j}_2(\mathbf{x}) d\mathbf{x} \right|_{\mathbb{R}^3} \leq C \int_S \frac{\varepsilon}{\varepsilon^\alpha |x - y|^{5/2-\alpha}} |\mathbf{j}_2|.$$

Using an Hardy-Littlewood-Sobolev inequality (e.g. [HZ16]) for  $1 \leq p < \infty$ , there exists  $C_\alpha > 0$  such that

$$\left| \int_S \left( \frac{\varepsilon \langle \mathbf{n}(\mathbf{y}), \mathbf{n}(\mathbf{x}) \rangle}{|\mathbf{y} - \mathbf{x} + \varepsilon \mathbf{n}(\mathbf{y})|^3} - \frac{\varepsilon \langle \mathbf{n}(\mathbf{y}), \mathbf{n}(\mathbf{x}) \rangle}{|\mathbf{y} - \mathbf{x} - \varepsilon \mathbf{n}(\mathbf{y})|^3} \right) \mathbf{j}_2(\mathbf{x}) d\mathbf{x} \right|_{\mathbb{R}^p(S)} \leq C_\alpha \varepsilon^{1-\alpha} |\mathbf{j}_2|_{\mathbb{R}^{1,2}(S)}.$$

Thus

$$\int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{n}(\mathbf{x}) \rangle \left( \frac{\varepsilon \langle \mathbf{n}(\mathbf{y}), \mathbf{n}(\mathbf{x}) \rangle}{|\mathbf{y} - \mathbf{x} + \varepsilon \mathbf{n}(\mathbf{y})|^3} - \frac{\varepsilon \langle \mathbf{n}(\mathbf{y}), \mathbf{n}(\mathbf{x}) \rangle}{|\mathbf{y} - \mathbf{x} - \varepsilon \mathbf{n}(\mathbf{y})|^3} \right) \mathbf{j}_2(\mathbf{x}) d\mathbf{x} \xrightarrow{\varepsilon^p(S)} 0 \quad (\text{IV.39})$$

In summary, Eq. IV.9 is the sum of two integrals converging respectively as in Eq. IV.38 and IV.39. Ultimately the “first normal term” converges to Eq. IV.15.

#### IV.A.4 Second normal term

The same reasoning just applied to integral IV.9 also applies to integral IV.13,

$$\int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{j}_2(\mathbf{x}) \rangle \frac{\langle \mathbf{y} - \mathbf{x}, \mathbf{n}(\mathbf{x}) \rangle \pm \varepsilon \langle \mathbf{n}(\mathbf{y}), \mathbf{n}(\mathbf{x}) \rangle}{|\mathbf{y} - \mathbf{x} \pm \varepsilon \mathbf{n}(\mathbf{y})|^3} d\mathbf{x},$$

which converges to

$$\int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{j}_2(\mathbf{x}) \rangle \frac{\langle \mathbf{y} - \mathbf{x}, \mathbf{n}(\mathbf{x}) \rangle}{|\mathbf{y} - \mathbf{x}|^3} d\mathbf{x},$$

i.e. to Eq. IV.17.

This concludes the proof of Theorem IV.2: one by one, in Secs. IV.A.1-IV.A.4, we have obtained all terms in Eqs. IV.14-IV.17.

Note that we do not expect  $\int_S \langle \mathbf{j}_1(\mathbf{y}) \cdot \mathbf{j}_2(\mathbf{x}) \rangle \frac{\varepsilon \langle \mathbf{n}(\mathbf{y}), \mathbf{n}(\mathbf{x}) \rangle}{|\mathbf{y} - \mathbf{x} + \varepsilon \mathbf{n}(\mathbf{y})|^3} d\mathbf{x}$  to go to 0 as that term is responsible for the magnetic field discontinuity. But we are still able to use Lemma IV.9 to control the term  $\frac{\varepsilon}{|\mathbf{y} - \mathbf{x} + \varepsilon \mathbf{n}(\mathbf{y})|^3} - \frac{\varepsilon}{|\mathbf{y} - \mathbf{x} - \varepsilon \mathbf{n}(\mathbf{y})|^3}$ .

**Remark IV.10.** *We do not expect  $\mathbf{L}(\mathbf{j}_1, \mathbf{j}_2)$  to be in  $L^\infty(S, \mathbb{R}^3)$ . Indeed,  $H^1(S)$  is not embedded in  $L^\infty(S)$  for manifolds of dimension 2. For example, there is no constant  $C > 0$  such that  $|\int_S \frac{1}{|\mathbf{y} - \mathbf{x}|} \operatorname{div}_{\mathbf{x}}(\pi_{\mathbf{x}} \mathbf{j}_1(\mathbf{y})) \mathbf{j}_2(\mathbf{x}) d\mathbf{x}|_{L^\infty(S, \mathbb{R}^3)} \leq C |\mathbf{j}_1|_{\mathfrak{X}^{1,2}(S)}$ .*

## IV.B Proof of existence of minimisers

In this section, we prove the existence of at least one minimiser to the following optimization problem:

$$\inf_{\mathbf{j} \in \mathfrak{X}^{1,2}(S)} \chi_B^2(\mathbf{j}) + \lambda |\mathbf{j}|_{\mathfrak{X}^{1,2}(S)}^2 + \gamma |\mathbf{L}(\mathbf{j})|_{L^p(S, \mathbb{R}^3)} \quad (\text{IV.40})$$

where,  $1 \leq p < \infty$  and  $\lambda, \gamma > 0$ . The proof follows the direct method of the calculus of variations. We consider  $(\mathbf{j}_n)_n$  a minimizing sequence of IV.40. As  $\lambda > 0$ ,  $(\mathbf{j}_n)_n$  is a bounded sequence in  $\mathfrak{X}^{1,2}(S)$ . Hence, up to an extraction, we can assume that it converges toward some  $\mathbf{j}_\infty \in \mathfrak{X}^{1,2}(S)$  weakly in  $\mathfrak{X}^{1,2}(S)$  and strongly in  $\mathfrak{X}^q(S)$  for any  $q < \infty$ . To conclude, we have to prove the following lower semi-continuity property:

$$\liminf_n \chi_B^2(\mathbf{j}_n) + \lambda |\mathbf{j}_n|_{\mathfrak{X}^{1,2}(S)}^2 + \gamma |\mathbf{L}(\mathbf{j}_n)|_{L^p(S)} \leq \chi_B^2(\mathbf{j}_\infty) + \lambda |\mathbf{j}_\infty|_{\mathfrak{X}^{1,2}(S)}^2 + \gamma |\mathbf{L}(\mathbf{j}_\infty)|_{L^p(S)}.$$

Using the convexity of  $\chi_B^2(\cdot)$  and  $|\cdot|_{\mathfrak{X}^{1,2}(S)}$ , we get

$$\liminf_n \chi_B^2(\mathbf{j}_n) + \lambda |\mathbf{j}_n|_{\mathfrak{X}^{1,2}(S)}^2 \leq \chi_B^2(\mathbf{j}_\infty) + \lambda |\mathbf{j}_\infty|_{\mathfrak{X}^{1,2}(S)}^2.$$

It remains to study the sequence  $(|\mathbf{L}(\mathbf{j}_n)|_{L^p(S)})_n$ . To this aim, we show that  $\mathbf{L}(\mathbf{j}_n)$  converges strongly toward  $\mathbf{L}(\mathbf{j}_\infty)$  in  $L^p$ . We recall that

$$\mathbf{L}(\mathbf{j}_n)(\mathbf{y}) = - \int_S \frac{1}{|\mathbf{y} - \mathbf{x}|} \operatorname{div}_{\mathbf{x}}(\pi_{\mathbf{x}} \mathbf{j}_n(\mathbf{y})) \mathbf{j}_n(\mathbf{x}) d\mathbf{x} \quad (\text{IV.41})$$

$$- \int_S \frac{1}{|\mathbf{y} - \mathbf{x}|} (\pi_{\mathbf{x}} \mathbf{j}_n(\mathbf{y}) \cdot \nabla_{\mathbf{x}}) \mathbf{j}_n(\mathbf{x}) d\mathbf{x} \quad (\text{IV.42})$$

$$+ \int_S \langle \mathbf{j}_n(\mathbf{y}) \cdot \mathbf{n}(\mathbf{x}) \rangle \frac{\langle \mathbf{y} - \mathbf{x} \cdot \mathbf{n}(\mathbf{x}) \rangle}{|\mathbf{y} - \mathbf{x}|^3} \mathbf{j}_n(\mathbf{x}) d\mathbf{x} \quad (\text{IV.43})$$

$$+ \int_S \frac{1}{|\mathbf{y} - \mathbf{x}|} \langle \mathbf{j}_n(\mathbf{y}) \cdot \mathbf{j}_n(\mathbf{x}) \rangle \operatorname{div}_{\mathbf{x}}(\pi_{\mathbf{x}}) d\mathbf{x} \quad (\text{IV.44})$$

$$+ \int_S \frac{1}{|\mathbf{y} - \mathbf{x}|} \nabla_{\mathbf{x}} \langle \mathbf{j}_n(\mathbf{y}) \cdot \mathbf{j}_n(\mathbf{x}) \rangle d\mathbf{x} \quad (\text{IV.45})$$

$$- \int_S \langle \mathbf{j}_n(\mathbf{y}) \cdot \mathbf{j}_n(\mathbf{x}) \rangle \frac{\langle \mathbf{y} - \mathbf{x} \cdot \mathbf{n}(\mathbf{x}) \rangle}{|\mathbf{y} - \mathbf{x}|^3} \mathbf{n}(\mathbf{x}) d\mathbf{x}. \quad (\text{IV.46})$$

We prove the convergence term by term only for the ones in Lines (IV.41), (IV.42) and (IV.43). The remaining ones can be obtained by similar arguments.

**Line (IV.41)** For every  $\mathbf{x} \in S$ , we denote by  $F_{\mathbf{x}} : \mathbb{R}^3 \rightarrow \mathbb{R}$  the linear form  $\mathbf{v} \mapsto \operatorname{div}_{\mathbf{x}}(\pi_{\mathbf{x}}\mathbf{v})$ . Using the compactness of  $S$ , the family of linear forms  $(F_{\mathbf{x}})_{\mathbf{x} \in S}$  is uniformly bounded.

Representing  $F_{\mathbf{x}}$  as a row vector, Line (IV.41) becomes

$$\int_S \frac{1}{|\mathbf{y} - \mathbf{x}|} \operatorname{div}_{\mathbf{x}}(\pi_{\mathbf{x}}\mathbf{j}_n(\mathbf{y}))\mathbf{j}_n(\mathbf{x})d\mathbf{x} = \mathbf{j}_n(\mathbf{y})^T \int_S \frac{1}{|\mathbf{y} - \mathbf{x}|} F_{\mathbf{x}}^T \mathbf{j}_n(\mathbf{x})d\mathbf{x},$$

where  $F_{\mathbf{x}}^T \mathbf{j}_n(\mathbf{x})$  is the matrix with coefficients  $(F_{\mathbf{x}})_i \mathbf{j}_n(\mathbf{x})_j$ . Using that the application

$$K : L^2(S) \rightarrow L^q(S), \quad \mathbf{X} \mapsto \int_S \frac{1}{|\mathbf{y} - \mathbf{x}|} \mathbf{X}(\mathbf{x})d\mathbf{x}$$

is compact for any  $q < \infty$  (see e.g. [HZ16]), we get

$$\int_S \frac{1}{|\mathbf{y} - \mathbf{x}|} F_{\mathbf{x}}^T \mathbf{j}_n(\mathbf{x})d\mathbf{x} \xrightarrow{L^{2p}(S)} \int_S \frac{1}{|\mathbf{y} - \mathbf{x}|} F_{\mathbf{x}}^T \mathbf{j}(\mathbf{x})d\mathbf{x}.$$

Together with the fact that  $\mathbf{j}_n \xrightarrow{L^{2p}(S)} \mathbf{j}_{\infty}$ , we conclude that

$$\int_S \frac{1}{|\mathbf{y} - \mathbf{x}|} \operatorname{div}_{\mathbf{x}}(\pi_{\mathbf{x}}\mathbf{j}_n(\mathbf{y}))\mathbf{j}_n(\mathbf{x})d\mathbf{x} \xrightarrow{L^p(S)} \int_S \frac{1}{|\mathbf{y} - \mathbf{x}|} \operatorname{div}_{\mathbf{x}}(\pi_{\mathbf{x}}\mathbf{j}(\mathbf{y}))\mathbf{j}(\mathbf{x})d\mathbf{x}.$$

**Line (IV.42)** We follow a similar reasoning. The application  $\pi_{\mathbf{x}}$  is uniformly bounded with respect to  $\mathbf{x}$ . Now, we reexpress the term of Line (IV.42)

$$\int_S \frac{1}{|\mathbf{y} - \mathbf{x}|} (\pi_{\mathbf{x}}\mathbf{j}_n(\mathbf{y}) \cdot \nabla_{\mathbf{x}})\mathbf{j}_n(\mathbf{x})d\mathbf{x} = \mathbf{j}_n(\mathbf{y})^T \int_S \frac{1}{|\mathbf{y} - \mathbf{x}|} (\pi_{\mathbf{x}}^T \nabla_{\mathbf{x}})\mathbf{j}_n(\mathbf{x}).$$

$\mathbf{j}_n$  converges strongly in  $L^{2p}$  and  $(\pi_{\mathbf{x}}^T \nabla_{\mathbf{x}})\mathbf{j}_n(\mathbf{x})$  weakly in  $L^2$ . Using the compactness of  $K$  again,  $K(\pi_{\mathbf{x}}^T \nabla_{\mathbf{x}}\mathbf{j}_n(\mathbf{x}))$  converges strongly in  $L^{2p}$ . As a consequence,

$$\int_S \frac{1}{|\mathbf{y} - \mathbf{x}|} (\pi_{\mathbf{x}}\mathbf{j}_n(\mathbf{y}) \cdot \nabla_{\mathbf{x}})\mathbf{j}_n(\mathbf{x})d\mathbf{x} \xrightarrow{L^p(S)} \int_S \frac{1}{|\mathbf{y} - \mathbf{x}|} (\pi_{\mathbf{x}}\mathbf{j}(\mathbf{y}) \cdot \nabla_{\mathbf{x}})\mathbf{j}(\mathbf{x})d\mathbf{x}$$

**line (IV.43)** Using Lemma IV.7, there exists  $C > 0$  such that for all  $\mathbf{x}, \mathbf{y} \in S$ ,  $\frac{|\langle \mathbf{y} - \mathbf{x}, \mathbf{n}(\mathbf{x}) \rangle|}{|\mathbf{y} - \mathbf{x}|^3} \leq C \frac{1}{|x-y|}$ . Thus, using the same reasoning again

$$\int_S \frac{\langle \mathbf{y} - \mathbf{x}, \mathbf{n}(\mathbf{x}) \rangle}{|\mathbf{y} - \mathbf{x}|^3} \mathbf{j}_n(\mathbf{x})d\mathbf{x} \xrightarrow{L^{2p}(S)} \int_S \frac{\langle \mathbf{y} - \mathbf{x}, \mathbf{n}(\mathbf{x}) \rangle}{|\mathbf{y} - \mathbf{x}|^3} \mathbf{j}(\mathbf{x})d\mathbf{x}.$$

Leading once more to

$$\int_S \langle \mathbf{j}_n(\mathbf{y}) \cdot \mathbf{n}(\mathbf{x}) \rangle \frac{\langle \mathbf{y} - \mathbf{x}, \mathbf{n}(\mathbf{x}) \rangle}{|\mathbf{y} - \mathbf{x}|^3} \mathbf{j}_n(\mathbf{x})d\mathbf{x} \xrightarrow{L^p(S)} \int_S \langle \mathbf{j}(\mathbf{y}) \cdot \mathbf{n}(\mathbf{x}) \rangle \frac{\langle \mathbf{y} - \mathbf{x}, \mathbf{n}(\mathbf{x}) \rangle}{|\mathbf{y} - \mathbf{x}|^3} \mathbf{j}(\mathbf{x})d\mathbf{x}.$$

This concludes the proof of the existence of at least one minimiser to the problem (IV.40).

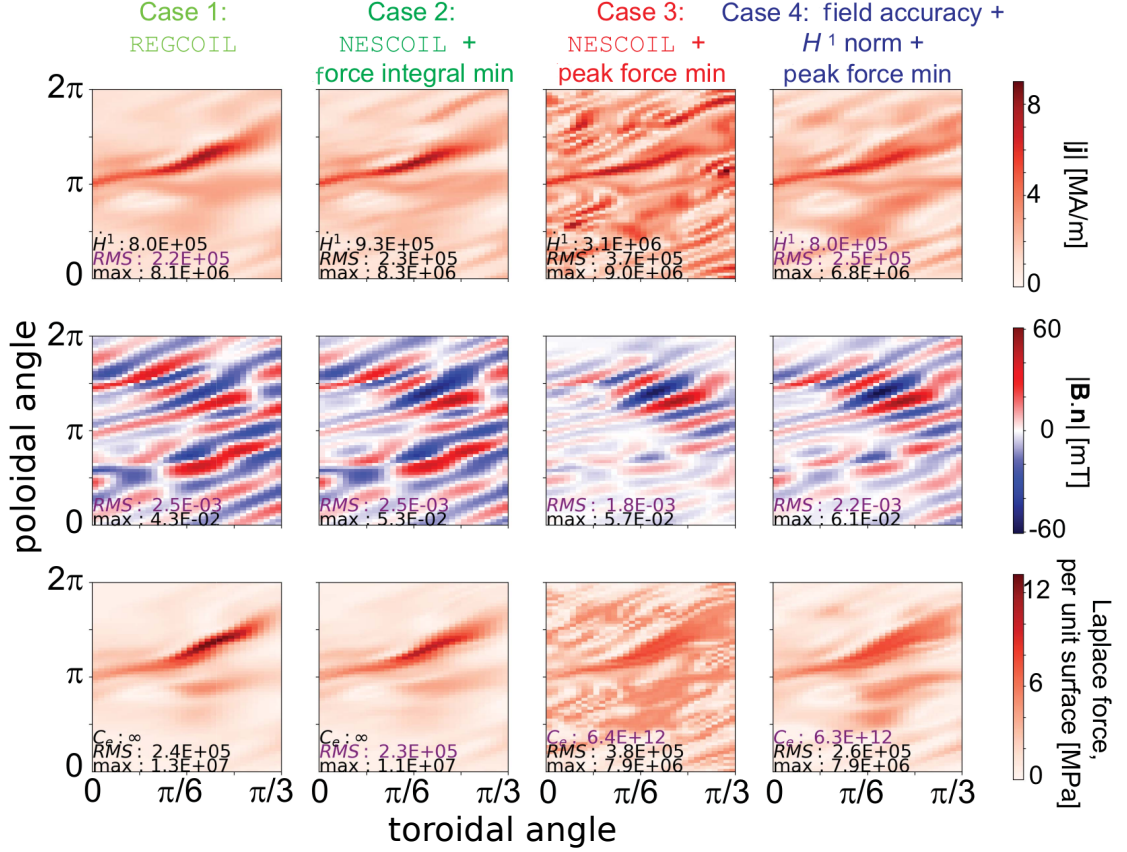


Figure IV.7 – Results of minimizing Eq. IV.25 for NCSX, for four different cases (four different choices of weights in the equation, as summarized by Table IV.28). Each column refers to a different case; its title is color-coded like the corresponding data-point in Fig. IV.6. From top to bottom the three rows refer respectively to the results for simultaneous (1) current regularization (if any), (2) field accuracy and (3) force-minimization (if any). Case 4 (last column) demonstrates that it is possible to simultaneously optimize these three competing objectives without excessively penalizing any of them with respect to established codes. On the contrary, case 4 actually exhibits higher field accuracy and lower peak forces compared to REGCOIL (first column). Shown in the legends are the Root Mean Square (RMS) surface-averages and local maxima of the quantities plotted, as well as the  $H^1$  norm of  $\mathbf{j}$  and  $C_e$  force metric (Eq. IV.27). The quantities actually minimized are marked in purple. 624 DOF are used for  $\mathbf{j}$  in every simulation. It is well-known from [Lan17] and Fig. IV.5 that a higher number of DOF will lower all individual metrics  $\chi_B^2$ ,  $\chi_j^2$ ,  $\chi_{\nabla j}^2$ ,  $\chi_F^2$  and find better compromises among them. Correspondingly, all contours presented here will improve, for all 4 cases, and by the same proportion. However, this will require more computational resources, and is left as future work.

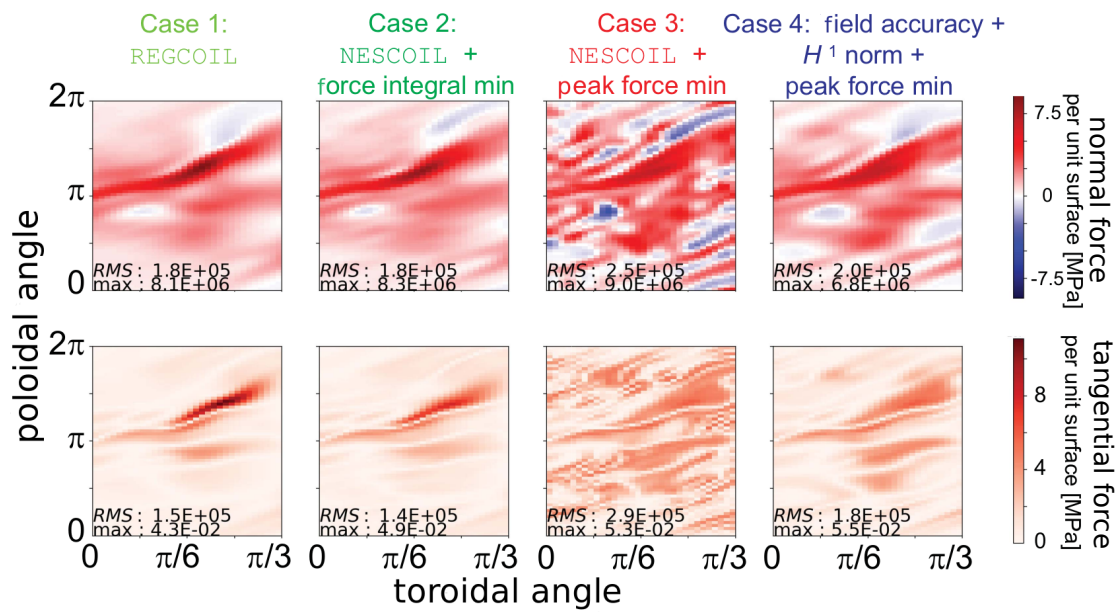


Figure IV.8 – Tangential and normal components of the Laplace forces of the simulations in Fig. IV.7.



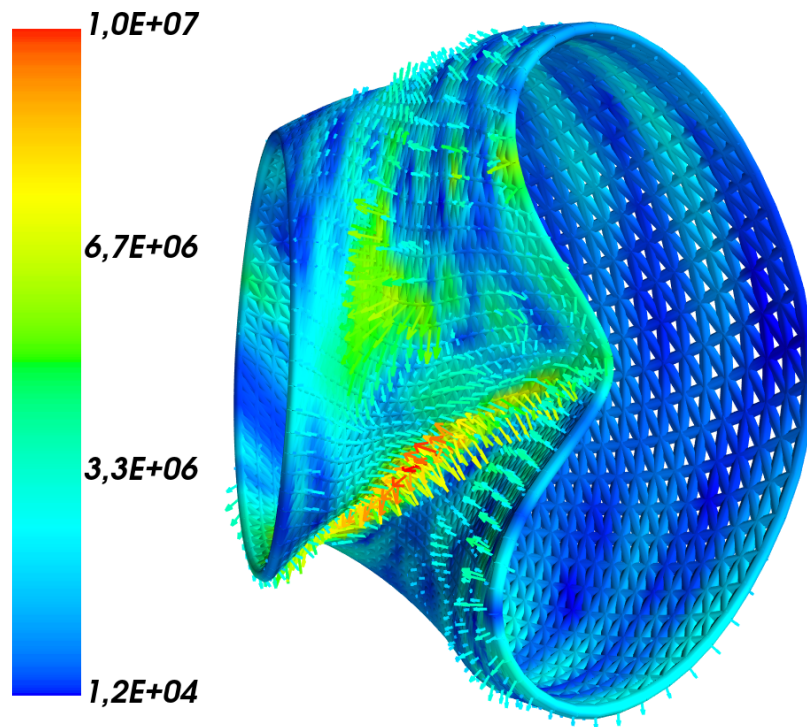


Figure IV.9 – The Laplace forces from the last column of Fig. IV.7. The unit for the pressure is Pascal. The triangular mesh is only used for rendering; the actual calculations were carried out on a  $64 \times 64$ , poloidal  $\times$  toroidal mesh.

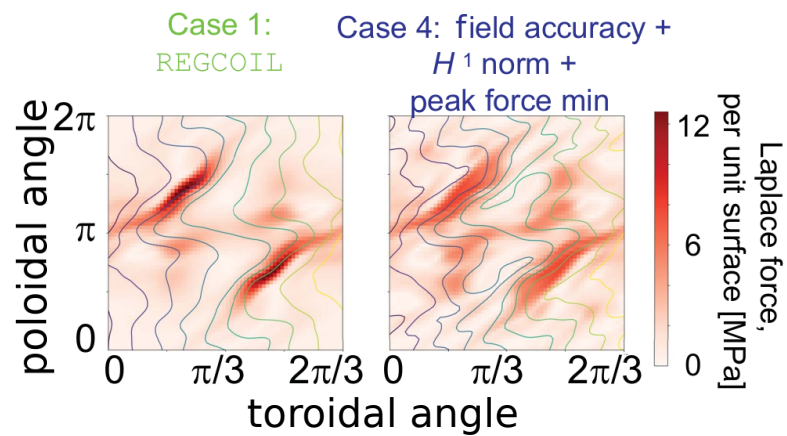


Figure IV.10 – Examples of current filamentation for case 1 and 4.

## Second Part: Quantum control



## Chapter V

# Ensemble qubit controllability with a single control via adiabatic and rotating wave approximations

This chapter is taken from the following article (also referred as [Rob+22b]):

R. Robin, N. Augier, U. Boscain, and M. Sigalotti. “Ensemble qubit controllability with a single control via adiabatic and rotating wave approximations”. In: *Journal of Differential Equations* 318 (2022), pp. 414–442

In the physics literature it is common to use "in cascade" the rotating wave approximation and the adiabatic approximation for chirped pulses of two-level quantum systems driven by one external field, in particular when the resonance frequency is not known precisely. Both approximations need relatively long time and are based on averaging theory of dynamical systems. Unfortunately, the two approximations cannot be done independently since, in a sense, the two time scales interact. We study how the cascade of the two approximations can be justified, while preserving the robustness of the adiabatic strategy. Our first result, based on high-order averaging techniques, gives a precise quantification of the uncertainty interval of the resonance frequency for which the population inversion works. As a by-product, we prove controllability of an ensemble of spin systems by a single real-valued control, providing an extension of a celebrated result with two controls by Khaneja and Li.

### V.1 Introduction

Consider a two-level system described by the Schrödinger equation

$$i\frac{d\psi}{dt} = \begin{pmatrix} E + \alpha & w(t) \\ w(t) & -E - \alpha \end{pmatrix} \psi. \quad (\text{V.1})$$

Here  $w : [0, T] \rightarrow \mathbb{R}$  is a (sufficiently regular) function representing an external field,  $E > 0$ , and  $\alpha \in [\alpha_0, \alpha_1]$  is an unknown parameter representing the fact that the *resonance frequency* of the system  $2(E + \alpha)$  is not known precisely, but lies between  $2(E + \alpha_0)$  and  $2(E + \alpha_1)$ . All along

this chapter we assume the condition

$$[\alpha_0, \alpha_1] \subset (-E, \infty), \quad 0 \in (\alpha_0, \alpha_1),$$

guaranteeing that the eigenvalues of the matrix in equation (V.1) are never zero, independently of the value of  $\alpha$ . The solution of (V.1) (that depends on  $\alpha$  and  $w(\cdot)$ ) with initial condition  $\psi_w^\alpha(0) = (0, 1)$  is the wave function  $\psi_w^\alpha : [0, T] \rightarrow \mathbb{C}^2$ .

One would like to find a function  $w(\cdot)$  (the same for all values of  $\alpha$ ) such that, if at time zero the system is at the ground state  $(0, 1)$  (i.e., it is in the eigenstate corresponding to the eigenvalue  $-E - \alpha$ ), then at time  $T$  the system is close to a state of the form  $(e^{i\theta}, 0)$  for some  $\theta \in \mathbb{R}$ . In mathematical terms this can be rephrased as follows.

**P:** For every  $\varepsilon > 0$ , find a time  $T$  and an external field  $w : [0, T] \rightarrow \mathbb{R}$  such that

$$|\psi_w^\alpha(T) - (e^{i\theta}, 0)| < \varepsilon,$$

for every  $\alpha \in [\alpha_0, \alpha_1]$  and for some  $\theta \in \mathbb{R}$  (possibly depending on  $\varepsilon, E, T, w, \alpha$ ).

In the mathematical literature it has been proved that problem **P** admits a solution when one replaces the real-valued function  $w$  by a complex-valued one, as in equation (V.2) below ([BCR10a; LK06; LK09; Mac+19]). As far as we are aware, the problem is open in the case of real-valued functions. The result proved in this chapter (Theorem V.3) solves problem **P** in a more general framework, in which there is an additional parameter dispersion on the coupling between the control and the system (that is,  $w(t)$  is replaced by  $\delta w(t)$  for  $\delta$  in a compact interval of  $(0, +\infty)$ ).

Solving **P** is a key ingredient to prove ensemble controllability of (V.1) with more general initial and final conditions. This celebrated problem has been solved in the case where  $w$  is replaced by a complex-valued control in [LK09; LK06] and [BCR10a].

The intuitive approach to tackle problem **P**, consists in the following two steps ([Mit13; Sho08; Sho11; Vit+01]):

- use an external field oscillating at the resonance frequency  $2E$  and having a small and slowly varying amplitude and a slowly varying phase, to simulate by rotating wave approximation (RWA, for short) a system driven by a complex-valued function (in a sense, this “duplicates” the number of available external fields);
- use an adiabatic strategy based on chirped pulses (i.e., pulses whose frequency is slowly increasing from a value below  $2(E + \alpha_0)$  to a value above  $2(E + \alpha_1)$ ) to drive the system from an eigenstate to the other one independently of the value of  $\alpha$ . This second step substantially exploits the presence of a complex-valued external field and is called adiabatic approximation (AA, for short) [GD02; MK01; Sha12; Sim+11; Wu+11]. Alternative robust methods are developed, for example, in [Jo+17; TGV11].

However the RWA may affect the precision of the adiabatic strategy, as it has been remarked in [LSR11]. In order to detail in which sense the “cascade” of the two approximations introduced above may break down, let us give some quantitative estimate.

### V.1.1 Rotating wave approximation

Consider a two-level system of the form

$$i \frac{d\psi}{dt} = \begin{pmatrix} E & w(t) \\ w^*(t) & -E \end{pmatrix} \psi. \quad (\text{V.2})$$

Here we assume that the resonance frequency of the system is known precisely, hence we have no  $\alpha$ . The symbol  $w^*$  denotes the complex conjugate of  $w$ , which represents here a complex-valued external field. For every  $\varepsilon > 0$ , consider the external fields

$$w_\varepsilon(t) = 2\varepsilon u(\varepsilon t) \cos(2Et + \Delta(\varepsilon t)), \quad (\text{V.3})$$

$$w_\varepsilon^{\text{R}}(t) = \varepsilon u(\varepsilon t) e^{-i(2Et + \Delta(\varepsilon t))}. \quad (\text{V.4})$$

where  $u(\cdot)$  and  $\Delta(\cdot)$  are two real-valued smooth functions defined on  $[0, T]$ ,  $T > 0$ . We have the following.

**Proposition V.1.** *For  $\varepsilon > 0$  let  $\psi_{w_\varepsilon}$  and  $\psi_{w_\varepsilon^{\text{R}}}$  be the solutions of (V.2) with initial condition  $\psi_0 \in \mathbb{C}^2$  corresponding to the external fields  $w_\varepsilon$  and  $w_\varepsilon^{\text{R}}$ , respectively. Then  $\max_{t \in [0, T/\varepsilon]} |\psi_{w_\varepsilon}(t) - \psi_{w_\varepsilon^{\text{R}}}(t)|$  converges to 0 as  $\varepsilon \rightarrow 0$ .*

The proof of this fact is well known. If one applies the unitary change of variables

$$\psi_{w_\varepsilon}(t) = \begin{pmatrix} e^{-i(Et + \Delta(\varepsilon t)/2)} & 0 \\ 0 & e^{i(Et + \Delta(\varepsilon t)/2)} \end{pmatrix} \hat{\psi}_{w_\varepsilon}(t)$$

then  $\hat{\psi}_{w_\varepsilon}(t)$  satisfies the Schrödinger equation

$$i \frac{d\hat{\psi}_{w_\varepsilon}}{dt} = \varepsilon \left[ \begin{pmatrix} -\Delta'(\varepsilon t)/2 & u(\varepsilon t) \\ u(\varepsilon t) & \Delta'(\varepsilon t)/2 \end{pmatrix} + \begin{pmatrix} 0 & e^{i(4Et + 2\Delta(\varepsilon t))} u(\varepsilon t) \\ e^{-i(4Et + 2\Delta(\varepsilon t))} u(\varepsilon t) & 0 \end{pmatrix} \right] \hat{\psi}_{w_\varepsilon}.$$

Here  $\Delta'$  indicates the derivative of the function  $\Delta : [0, T] \rightarrow \mathbb{R}$ . Now, defining  $s = \varepsilon t$ , varying in the interval  $[0, T]$ , and  $\tilde{\psi}_{w_\varepsilon}(s) = \hat{\psi}_{w_\varepsilon}(t/\varepsilon)$  we obtain

$$i \frac{d\tilde{\psi}_{w_\varepsilon}}{ds} = \left[ \begin{pmatrix} -\Delta'(s)/2 & u(s) \\ u(s) & \Delta'(s)/2 \end{pmatrix} + \underbrace{\begin{pmatrix} 0 & e^{i(4Es/\varepsilon + 2\Delta(s))} u(s) \\ e^{-i(4Es/\varepsilon + 2\Delta(s))} u(s) & 0 \end{pmatrix}}_{=: B(s, \varepsilon)} \right] \tilde{\psi}_{w_\varepsilon}. \quad (\text{V.5})$$

The same change of variables on  $\psi_{w_\varepsilon^{\text{R}}}$  gives rise to

$$i \frac{d\tilde{\psi}_{w_\varepsilon^{\text{R}}}}{ds} = \begin{pmatrix} -\Delta'(s)/2 & u(s) \\ u(s) & \Delta'(s)/2 \end{pmatrix} \tilde{\psi}_{w_\varepsilon^{\text{R}}}. \quad (\text{V.6})$$

Equations (V.5) and (V.6) differ only for the term  $B(s, \varepsilon)$ . Since for every interval  $[s_1, s_2] \subseteq [0, T]$  we have

$$\lim_{\varepsilon \rightarrow 0} \int_{s_1}^{s_2} B(s, \varepsilon) = 0$$

and  $B$  is uniformly bounded, we have that solutions of (V.5) converge uniformly in  $[0, T]$  to solutions of (V.6) with the same initial condition. This is a classical averaging result that can be found, for instance, in [AS04, Chapter 8]. Coming back to the original variables one obtains that  $|\psi_{w_\varepsilon} - \psi_{w_\varepsilon^{\text{R}}}|$  converges uniformly to zero on the interval  $[0, T/\varepsilon]$ .

This simple argument is very useful. We started with a system driven by one scalar function  $w$  and we obtain at the limit a system driven by a complex-valued control or, equivalently, system (V.6) where the controls are the two scalar functions  $u(t)$  and  $v(t) = \Delta'(t)/2$ . A more

detailed quantitative analysis permits to conclude that on  $[0, T/\varepsilon]$  we have

$$|\psi_{w_\varepsilon} - \psi_{w_\varepsilon^R}| = O(\varepsilon).$$

(See, for instance, [ABS20, Appendix A] for a quantitative version of the averaging result mentioned above.) Higher order RWA can be obtained by considering higher-order averaging results.

In recent applications, it is sometimes necessary to use intense external fields. In these cases the RWA may become inaccurate, as pointed out in [Ash+07; Cao+10; Sch+14]. Thus it is crucial to have a precise quantification of the error.

### V.1.2 Adiabatic approximation

We have seen in the previous section how to make the solutions of system (V.2) approximate those of system (V.2). We show here how such a system can be easily driven by adiabatic pulses.

Let us consider the case in which the energy of the system is not known precisely. We are then considering the system

$$i \frac{d\psi}{dt} = \begin{pmatrix} E + \alpha & w(t) \\ w^*(t) & -E - \alpha \end{pmatrix} \psi, \quad \text{where } \alpha \in [\alpha_1, \alpha_2]. \quad (\text{V.7})$$

Let us choose the pulse  $w$  in the form

$$w(t) = u(t)e^{-i(2Et + \Delta(t))}, \quad (\text{V.8})$$

where  $u(\cdot)$  and  $\Delta(\cdot)$  are two real-valued smooth functions. This choice of control corresponds to (V.4) in which  $\varepsilon$  has been set equal to 1. Applying the change of variables

$$\psi(t) = \begin{pmatrix} e^{-i(Et + \Delta(t)/2)} & 0 \\ 0 & e^{i(Et + \Delta(t)/2)} \end{pmatrix} \Psi(t),$$

we obtain

$$i \frac{d\Psi}{ds} = \begin{pmatrix} \alpha - v(s) & u(s) \\ u(s) & -\alpha + v(s) \end{pmatrix} \Psi. \quad (\text{V.9})$$

where  $v(t) := \Delta'(t)/2$ .

Notice that the eigenvalues of the matrix in equation (V.9), seen as functions of the pair  $(u, v)$ , coincide if and only if  $u = 0$  and  $v = \alpha$ , where a conical eigenvalue intersection occurs. Fix now  $v_0 < \alpha_0$  and  $v_1 > \alpha_1$  and consider a smooth path  $t \mapsto (u(t), v(t))$  lying in the half-plane  $u > 0$  except for the initial and final points, where  $u = 0$  (see Figure V.1).

Define

$$u_\varepsilon(t) = u(\varepsilon t), \quad v_\varepsilon(t) = v(\varepsilon t).$$

Since the eigenvalues of the Hamiltonian in Equation (V.9) are

$$\pm \sqrt{(\alpha - v(s))^2 + u(s)^2} \neq 0,$$

the adiabatic theorem (see, e.g., [Teu03]) ensures that, for  $\varepsilon > 0$  small, the trajectory of (V.9) corresponding to  $(u_\varepsilon, v_\varepsilon)$  and starting from  $(0, 1)$  stays close to the eigenvector associated with the negative eigenvalue. More precisely, we have the following estimate.

**Proposition V.2.** *There exists  $C > 0$  such that, for every  $\alpha \in [\alpha_0, \alpha_1]$  and every  $\varepsilon > 0$ , the solution  $\Psi$  of system (V.9) with initial condition  $(0, 1)$  and corresponding to the control  $(u_\varepsilon, v_\varepsilon)$  satisfies  $|\Psi(T/\varepsilon) - (e^{i\theta}, 0)| \leq C\varepsilon$  for some  $\theta \in \mathbb{R}$ .*

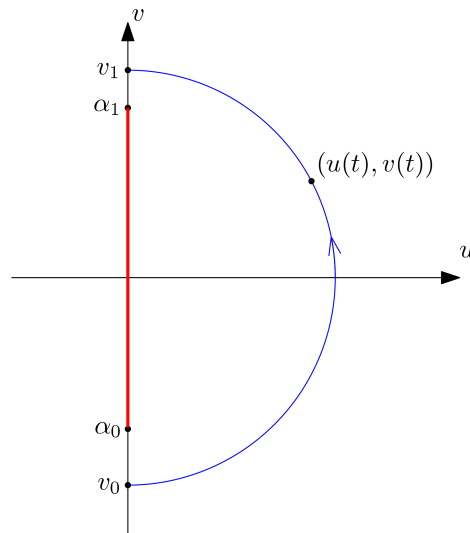


Figure V.1 – An adiabatic path as the one applied in Proposition V.2.

Going back to equation (V.7), the control corresponding to  $(u_\varepsilon, v_\varepsilon)$  is

$$w_\varepsilon(t) = u(\varepsilon t)e^{i(Et + \frac{\Delta(\varepsilon t)}{\varepsilon})}.$$

Such a law is called a (amplitude modulated) *chirped pulse*, since the range of frequency swept by the pulse is  $\{2E + \Delta'(s) \mid s \in [0, T]\}$ , which is independent of  $\varepsilon$ . For more details, see [ABS18].

### V.1.3 Combination of RWA and AA and statement of the population inversion result

What one would like to do is to consider the two approximations in cascade, in order to induce a transition from the state  $(0, 1)$  to  $(1, 0)$  (up to a phase) for an ensemble of systems parameterized by  $\alpha \in [\alpha_0, \alpha_1]$  using a real-valued external field. The cascade of the two approximations is expected to behave well in many experimental setups, such as in NMR, due to the separation of timescales between the RWA and the AA. However, for intense external fields or in presence of large parametric dispersions, the outcome of the cascade is more challenging to predict and quantify precisely. Let us denote by  $\varepsilon_1$  the small parameter that in the RWA was called  $\varepsilon$  and by  $\varepsilon_2$  what in the AA was called  $\varepsilon$ . A formal cascade of the two approaches yields a control law of the form

$$w_{\varepsilon_1, \varepsilon_2}(t) = 2\varepsilon_1 u(\varepsilon_1 \varepsilon_2 t) \cos\left(2Et + \frac{\Delta(\varepsilon_1 \varepsilon_2 t)}{\varepsilon_2}\right),$$

where  $u(\cdot)$  and  $v(\cdot)$  are the same functions as those used in Proposition V.2.

The hope is that the pulse  $w_{\varepsilon_1, \varepsilon_2}$ , for  $\varepsilon_1$  and  $\varepsilon_2$  small, induces approximately a transition from the state  $(0, 1)$  to a state of the form  $(e^{i\theta}, 0)$  in time  $T/(\varepsilon_1 \varepsilon_2)$ . The two approximations are, however, competing: when one decreases  $\varepsilon_2$  (better AA), one needs the RWA to be true for a longer time as the final time is of order  $1/(\varepsilon_1 \varepsilon_2)$ . On the other hand, decreasing  $\varepsilon_1$  deteriorates the performances of the AA:

1. The error on the adiabatic theorem depends of the gap between the eigenvalues, which



goes to zero as  $\varepsilon_1 \rightarrow 0$ ;

2. The range of frequencies swept by the pulse is  $\{2E + \varepsilon_1 \Delta'(s) \mid s \in [0, T]\}$ , that is, the allowed dispersion on the frequency is shrinking as  $\varepsilon_1$  goes to zero.

As a consequence, this method can only work when  $\alpha = 0$ . Under this restriction, and for suitable relations between  $\varepsilon_1$  and  $\varepsilon_2$  as they both go to zero, the cascade of the two approximations can be proved to work (see [ABS19; ABS22]).

Another possibility would be to fix  $\varepsilon_1$  small and to hope that the limit as  $\varepsilon_2 \rightarrow 0$  makes the RWA work as well. Nevertheless, the  $k$ -th order RWA is usually valid up to a time of order  $\frac{1}{\varepsilon_1^k}$ , whereas we would need the time to be of order  $\frac{1}{\varepsilon_1 \varepsilon_2}$ . In fact, without restriction on the allowed frequency, simulations show that convergence does not hold, as illustrated in Figure V.2.

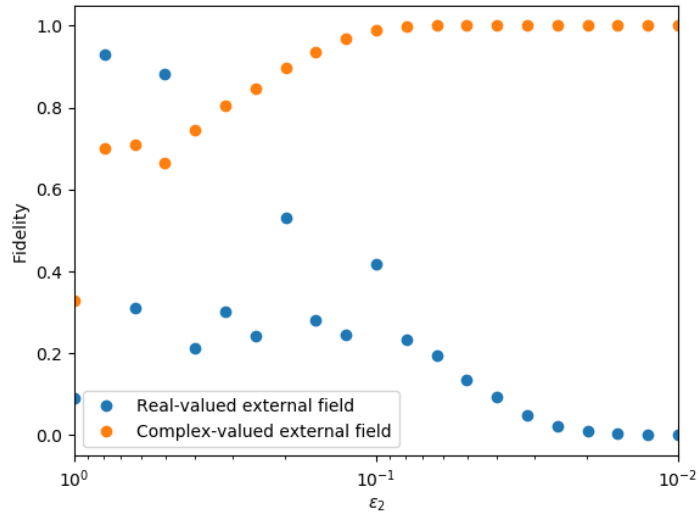


Figure V.2 – Comparison of the real-valued and complex-valued chirp scheme of the first point of Remark V.5 with  $E = 0.75$ ,  $\alpha = 0.25$ ,  $\varepsilon_1 = 1$ ,  $v_0 = -0.5$ ,  $v_1 = 0.5$ . Notice that the assumptions of Theorem V.3 are not satisfied.

An approach to tackle the issue of the shrinking interval of frequencies swept by the pulse is to divide  $\Delta(\varepsilon_1 \varepsilon_2 t)$  by  $\varepsilon_1 \varepsilon_2$  and not just by  $\varepsilon_2$ . We claim that an external field of the type

$$w_{\varepsilon_1, \varepsilon_2}(t) = 2\varepsilon_1 \delta u(\varepsilon_1 \varepsilon_2 t) \cos\left(2Et + \frac{\Delta(\varepsilon_1 \varepsilon_2 t)}{\varepsilon_1 \varepsilon_2}\right), \quad (\text{V.10})$$

where  $\delta$  is a positive constant, can induce a transition for the robust population transfer problem, provided that the relative order between  $\varepsilon_1$  and  $\varepsilon_2$  satisfies some suitable constraint as both parameters go to zero and under some further assumptions on the range  $[\alpha_0, \alpha_1]$ . This is detailed in the following theorem.

**Theorem V.3.** *Assume that  $v_0 < 0 < v_1$  are such that  $3(E + v_0) \geq E + v_1$ . Fix  $T > 0$  and  $u, \Delta : [0, T] \rightarrow \mathbb{R}$  smooth (e.g.,  $u \in C^2$  and  $\Delta \in C^3$ ) such that*

1.  $(u(0), \Delta'(0)) = (0, 2v_0)$  and  $(u(T), \Delta'(T)) = (0, 2v_1)$ ;

2.  $\forall s \in (0, T), u(s) > 0$  and  $\Delta''(s) \geq 0$ .

Denote by  $\psi_{\varepsilon_1, \varepsilon_2}^\alpha$  the solution of (V.1) with initial condition  $\psi_{\varepsilon_1, \varepsilon_2}^\alpha(0) = (0, 1)$  and control  $w_{\varepsilon_1, \varepsilon_2}$  as in (V.10). Then, for every  $N_0 \in \mathbb{N}$ , for every compact interval  $I \subseteq (v_0, v_1)$ , there exist  $C_{N_0} > 0$  and  $\eta > 0$  such that for every  $\alpha \in I$  and every  $(\varepsilon_1, \varepsilon_2) \in (0, \eta)^2$ ,

$$\left| \psi_{\varepsilon_1, \varepsilon_2}^\alpha \left( \frac{T}{\varepsilon_1 \varepsilon_2} \right) - (e^{i\theta}, 0) \right| < C_{N_0} \max \left( \frac{\varepsilon_2}{\varepsilon_1}, \frac{\varepsilon_1^{N_0-1}}{\varepsilon_2} \right)$$

for some  $\theta \in \mathbb{R}$ . Moreover, the constant  $C_{N_0}$  can be taken locally uniform with respect to the parameter  $\delta > 0$  appearing in (V.10).

Roughly speaking,  $\varepsilon_2/\varepsilon_1$  is the AA error and  $\varepsilon_1^{N_0-1}/\varepsilon_2$  the RWA error. We define the fidelity of a pulse as the quantity  $\inf_\theta |\psi_{\varepsilon_1, \varepsilon_2}^\alpha(\frac{T}{\varepsilon_1 \varepsilon_2}) - (e^{i\theta}, 0)|$  (also denoted  $|\langle \psi_{\varepsilon_1, \varepsilon_2}^\alpha(\frac{T}{\varepsilon_1 \varepsilon_2}) | e_1 \rangle|$ ). It is a natural measure of the transition rate induced by a pulse. Thus, by playing on the integer  $N_0$  and on the order of magnitude between  $\varepsilon_1$  and  $\varepsilon_2$ , we can express the fidelity attained by the strategy above in terms of the duration of the pulse.

**Corollary V.4.** Taking  $\varepsilon_1 = \varepsilon_2^{2/N_0}$  ( $N_0 \geq 3$ ) leads to an error of the order  $\mathcal{T}^{\frac{2/N_0-1}{1+2/N_0}}$ , where  $\mathcal{T} = 1/(\varepsilon_1 \varepsilon_2)$  is the duration of the pulse  $w_{\varepsilon_1, \varepsilon_2}$ .

**Remark V.5.** — As an example, one can apply Theorem V.3 with  $T = 1$ ,  $\delta = 1$ ,  $\Delta(s) = \frac{v_0 - v_1}{\pi} \sin(\pi s) + (v_0 + v_1)s$  and  $u(s) = 1 - \cos(2\pi s)$ ,  $s \in [0, 1]$ . More explicitly,

$$w_{\varepsilon_1, \varepsilon_2}(t) = 2\varepsilon_1(1 - \cos(2\pi\varepsilon_1\varepsilon_2t)) \cos \left( 2Et + \frac{(v_0 - v_1) \sin(\pi\varepsilon_1\varepsilon_2t)}{\pi\varepsilon_1\varepsilon_2} + (v_0 + v_1)t \right).$$

All the simulations in this chapter use this pulse scheme and some compare to the complex-valued pulse

$$w_{\varepsilon_1, \varepsilon_2}^R(t) = \varepsilon_1(1 - \cos(2\pi\varepsilon_1\varepsilon_2t)) \exp \left( 2iEt + i \frac{(v_0 - v_1) \sin(\pi\varepsilon_1\varepsilon_2t)}{\pi\varepsilon_1\varepsilon_2} + i(v_0 + v_1)t \right).$$

- By taking  $N_0$  large, one can get, for each  $\eta > 0$ , a fidelity close to one at order  $\mathcal{T}^{-1+\eta}$ , to compare with the standard  $O(\mathcal{T}^{-1})$  of the adiabatic theorem.
- The assumption  $3(E + v_0) \geq E + v_1$  ensures non-overlapping of some characteristic frequencies (cf. Lemma V.25). It could be replaced by the weaker one:  $4E + 3\Delta' - 2\alpha > 0$  for every  $\alpha \in [\alpha_0, \alpha_1]$  and everywhere in  $[0, T]$ . Nevertheless, asking this condition to be valid for every compact subinterval  $[\tilde{\alpha}_0, \tilde{\alpha}_1]$  of  $(v_0, v_1)$  is equivalent to the inequality  $3(E + v_0) \geq E + v_1$ .

Numerical simulations suggest that the inequality  $4E + 3\Delta' - 2\alpha > 0$  is sharp in the following sense: if for a given  $\alpha$ ,  $4E + 3\Delta'(s) - 2\alpha < 0$  for some  $s \in [0, T]$ , an inequality as in Theorem V.3 seem not to hold. As an illustration, in Figure V.3 we observe that for  $\alpha \geq 0$  (condition  $4E + 3\Delta' - 2\alpha > 0$  not satisfied), the accuracy of the RWA is worse than for  $\alpha < 0$  (condition  $4E + 3\Delta' - 2\alpha > 0$  satisfied).

**Remark V.6.** Many questions concerning the combination of the RWA and AA remain open. In particular we do not know if a version of Theorem V.3 holds with  $\varepsilon_1$  fixed, small enough, and  $\varepsilon_2$  going to 0.

Concerning systems with higher number of levels (possibly infinite), we expect the techniques developed in this chapter to work. Nevertheless, such an extension seems not trivial.

We postpone the proof of Theorem V.3 to Section V.3. This proof is technical and is sketched in Sections V.3.1 and V.3.2 (see also Remark V.34).

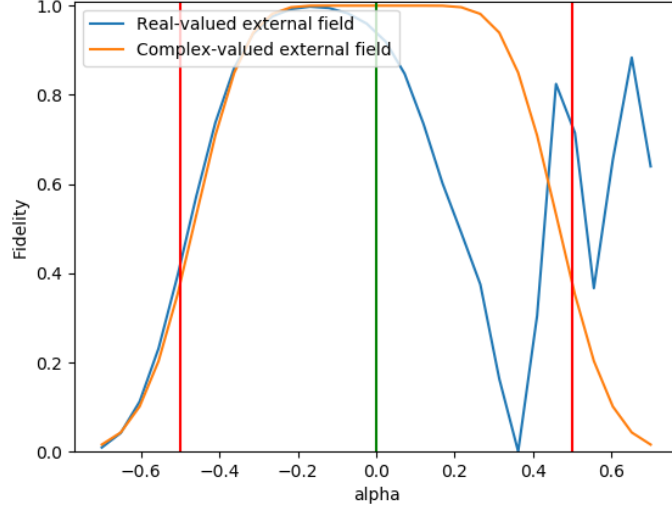


Figure V.3 –  $E = 0.75$ ,  $\alpha = 0.25$ ,  $\varepsilon_1 = 1$ ,  $\varepsilon_2 = 0.1$ ,  $v_0 = -0.5$ ,  $v_1 = 0.5$ . Assumption  $4E + 3\Delta'(s) - 2\alpha < 0$  is satisfied if and only if  $\alpha < 0$ .

## V.2 Application to the ensemble control problem

We denote by  $\sigma_x, \sigma_y, \sigma_z$  the Pauli matrices given by

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad (\text{V.11})$$

and by  $\text{SU}_2$  the special unitary group of degree 2. We recall that its Lie algebra  $\mathfrak{su}_2$  is generated by  $i\sigma_x$ ,  $i\sigma_y$ , and  $i\sigma_z$ .

There is a natural distance on  $\text{SU}_2$  induced by the norm of endomorphism on  $\mathbb{C}^2$ , which we denote  $\|\cdot\|$ . Let  $v_0 < 0 < v_1$  and  $0 < \delta_m \leq \delta_M$ . Let  $\mathcal{D} = [v_0, v_1] \times [\delta_m, \delta_M]$  be the compact set of the dispersion parameters and endow  $\mathcal{F} := C^0(\mathcal{D}, \text{SU}_2)$  with the usual distance  $d_{\mathcal{F}}(f, g) := \max_{d \in \mathcal{D}} \|f(d) - g(d)\|$ .

Li and Khaneja proved in [LK09] the following ensemble operator controllability result.

**Theorem V.7** (Li–Khaneja, 2009). *For any control bound  $K > 0$ , any target distribution  $M_F \in \mathcal{F}$ , and any  $\varepsilon > 0$ , there exist some  $T > 0$  and controls  $u, v \in L^\infty([0, T], [-K, K])$  such that the solution of the equation*

$$i \frac{d}{dt} M(\alpha, \delta, t) = ((E + \alpha)\sigma_z + \delta u(t)\sigma_x + \delta v(t)\sigma_y) M(\alpha, \delta, t), \quad M(\alpha, \delta, 0) = I_2, \quad \forall (\alpha, \delta) \in \mathcal{D} \quad (\text{V.12})$$

satisfies  $d_{\mathcal{F}}(M(\cdot, \cdot, T), M_F(\cdot, \cdot)) < \varepsilon$ .

**Remark V.8.** — *The result was originally stated on  $\text{SO}_3$  for the Bloch sphere, the extension to  $\text{SU}_2$  stated in Theorem V.12 is straightforward.*

— *This is a very strong ensemble controllability result, as it tackles the controllability of the semigroups.*

We extend here this result to the problem of a qubit driven by a single real control, thus replacing Equation (V.12) by

$$i \frac{d}{dt} M(\alpha, \delta, t) = ((E + \alpha)\sigma_z + \delta u(t)\sigma_x)M(\alpha, \delta, t), \quad M(\alpha, \delta, 0) = I_2, \quad \forall (\alpha, \delta) \in \mathcal{D}. \quad (\text{V.13})$$

One of the key ingredients of the proof of Theorem V.7 is the existence of an adiabatic pulse inducing a propagator  $U \in \mathcal{F}$  such that  $\max_{d \in \mathcal{D}} \min_{\theta \in [0, 2\pi]} \|U(d)(0, 1)^T - (e^{i\theta}, 0)^T\|$  is arbitrarily small.

Theorem V.3 ensures the following corollary.

**Corollary V.9.** *Suppose that  $3(E + v_0) > E + v_1$ . Then, for any  $K > 0$  and any  $\varepsilon > 0$ , there exist  $T > 0$  and a control  $u \in L^\infty([0, T], [-K, K])$  such that the solution of Equation (V.13) satisfies  $\max_{(\alpha, \delta) \in \mathcal{D}} \min_{\theta \in [0, 2\pi]} \|M(\alpha, \delta, T)(0, 1)^T - (e^{i\theta}, 0)^T\| < \varepsilon$ .*

Based on Corollary V.9, we will prove the following result, which generalizes Theorem V.7 under the extra assumption on the  $\alpha$ -dispersion.

**Theorem V.10.** *Suppose that  $3(E + v_0) > E + v_1$ . Let  $\epsilon > 0$ ,  $M_F \in \mathcal{F}$ , and  $K > 0$ . Then there exist  $T > 0$  and  $u \in L^\infty([0, T], [-K, K])$  such that the solution of Equation (V.13) satisfies  $d_{\mathcal{F}}(M(\cdot, \cdot, T), M_F(\cdot, \cdot)) < \varepsilon$ .*

The proof, sketched below, is an adaptation of the arguments used in [LK09].

Let  $\mathcal{R} = \{M(\cdot, \cdot, T) \mid T > 0, M \text{ is a solution of (V.13) for some } u \in L^\infty([0, T], [-K, K])\}$ . It is clear that  $\mathcal{R}$  and its closure  $\bar{\mathcal{R}}$  are semigroups of  $\mathcal{F}$ . We have to prove that  $\bar{\mathcal{R}} = \mathcal{F}$ .

**Lemma V.11.** *For all  $t$  in  $\mathbb{R}$ ,  $(\alpha, \delta) \mapsto e^{-t(E+\alpha)i\sigma_z}$  is in  $\bar{\mathcal{R}}$ .*

*Proof.* Using a null control in (V.13) during a time  $t \geq 0$ , we get  $(\alpha, \delta) \mapsto e^{-t(E+\alpha)i\sigma_z}$  belongs to  $\mathcal{R}$ .

Let us prove that the result also holds for  $t < 0$ . Set an arbitrary  $\varepsilon > 0$ . By Corollary V.9, there exists  $U^\varepsilon \in \mathcal{R}$  such that

$$\max_{d \in \mathcal{D}} \min_{\theta \in [0, 2\pi]} \|U^\varepsilon(d)(0, 1)^T - (e^{i\theta}, 0)^T\| < \varepsilon.$$

Using Euler angle decomposition, there exist three functions  $a_\varepsilon, b_\varepsilon, c_\varepsilon$  from  $\mathcal{D}$  to  $[0, 2\pi]$  (not necessarily continuous) such that  $U^\varepsilon(d) = e^{a_\varepsilon(d)i\sigma_z} e^{b_\varepsilon(d)i\sigma_x} e^{c_\varepsilon(d)i\sigma_z}$  for every  $d \in \mathcal{D}$ . In particular,  $\max_{d \in \mathcal{D}} |b_\varepsilon(d) - \pi|$  is of order  $\varepsilon$ , so that  $\sup_{d \in \mathcal{D}} \|U^\varepsilon(d) - \tilde{U}^\varepsilon(d)\|$  is also of order  $\varepsilon$ , where  $\tilde{U}^\varepsilon(d) := e^{a_\varepsilon(d)i\sigma_z} e^{\pi i\sigma_x} e^{c_\varepsilon(d)i\sigma_z}$ . For all  $t > 0$ , we have that  $(\alpha, \delta) \mapsto e^{-t(E+\alpha)i\sigma_z}$  is in  $\bar{\mathcal{R}}$ , by using the control  $u \equiv 0$ . Using the relation  $e^{-\pi i\sigma_x} e^{r i\sigma_z} e^{\pi i\sigma_x} = e^{-r i\sigma_z}$ ,  $r \in \mathbb{R}$ , we deduce that

$$\begin{aligned} \tilde{U}^\varepsilon(d) e^{-t(E+\alpha)i\sigma_z} \tilde{U}^\varepsilon(d) &= e^{a_\varepsilon(d)i\sigma_z} e^{-\pi i\sigma_x} e^{c_\varepsilon(d)i\sigma_z} e^{-t(E+\alpha)i\sigma_z} e^{a_\varepsilon(d)i\sigma_z} e^{\pi i\sigma_x} e^{c_\varepsilon(d)i\sigma_z} \\ &= e^{t(E+\alpha)i\sigma_z}, \end{aligned}$$

for every  $d = (\alpha, \delta)$  in  $\mathcal{D}$ . This shows that  $(\alpha, \delta) \mapsto e^{t(E+\alpha)i\sigma_z}$  is at distance of order  $\varepsilon$  from an element of  $\mathcal{R}$ , concluding the proof.  $\square$

**Lemma V.12.** *Let  $u \in \mathbb{R}$ . Then  $(\alpha, \delta) \mapsto e^{u\delta i\sigma_x}$  is in  $\bar{\mathcal{R}}$ .*

*Proof.* Consider first the case  $|u| \leq K$ . Setting  $V_n(\alpha, \delta) = e^{(-(E+\alpha)i\sigma_z + u\delta i\sigma_x)/n}$ , one can easily check that the sequence  $((\alpha, \delta) \mapsto (V_n(\alpha, \delta) e^{t(E+\alpha)i\sigma_z/n})^n)_{n \in \mathbb{N}}$  is in  $\bar{\mathcal{R}}$  and converges to  $(\alpha, \delta) \mapsto e^{u\delta i\sigma_x}$  in  $\mathcal{F}$ . This concludes the case  $|u| \leq K$ . We deduce the general case using the fact that  $\bar{\mathcal{R}}$  is a semigroup.  $\square$

Let

$$\mathfrak{g} = \{X \in \mathcal{C}^0(\mathcal{D}, \mathfrak{su}_2) \mid \forall t \in \mathbb{R}, e^{tX} \in \bar{\mathcal{R}}\}. \quad (\text{V.14})$$

Thus  $(\alpha, \delta) \mapsto \delta i\sigma_x$  and  $(\alpha, \delta) \mapsto (E + \alpha)i\sigma_z$  belong to  $\mathfrak{g}$ . The space  $\mathcal{C}^0(\mathcal{D}, \mathfrak{su}_2)$  has a natural addition, product, and Lie bracket. Moreover, it has the structure of Banach algebra using as norm the sup norm, denoted by  $|\cdot|_\infty$ . Before concluding the proof of Theorem V.10, let us to show that  $\mathfrak{g}$  is a Lie algebra by proving that it is stable by addition and Lie bracket.

**Lemma V.13.** *The set  $\mathfrak{g}$  defined in (V.14) is stable under addition and Lie brackets:*

$$[\mathfrak{g}, \mathfrak{g}] \subset \mathfrak{g}, \quad \mathfrak{g} + \mathfrak{g} \subset \mathfrak{g}.$$

*Proof.* Pick  $X, Y \in \mathfrak{g}$ . Let us first prove that  $e^{t[X, Y]} \in \bar{\mathcal{R}}$  for every  $t \in \mathbb{R}$ . To this purpose, consider  $U(s) = e^{sX} e^{sY} e^{-sX} e^{-sY}$ ,  $s \in [0, 1]$ . Then there exists a constant  $C > 0$  depending only on  $|X|_\infty$  and  $|Y|_\infty$  such that  $d_{\mathcal{F}}(U(\sqrt{s}), e^{s[X, Y]}) \leq Cs^{3/2}$  for every  $s \in [0, 1]$ . Using the fact that the application  $x \mapsto x^n$  is  $(n-1)$ -Lipschitz on the unit ball of any Banach algebra, we get

$$d_{\mathcal{F}}(U(\sqrt{s/n})^n, e^{s[X, Y]}) \leq C(s/n)^{3/2}(n-1) \leq Cs^{3/2}n^{-1/2}, \quad s \in [0, 1].$$

As a consequence,  $e^{s[X, Y]} \in \bar{\mathcal{R}}$  for every  $s \in [0, 1]$ . Applying the same reasoning to  $-X$  instead of  $X$ , we get that  $e^{s[X, Y]} \in \bar{\mathcal{R}}$  for every  $s \in [-1, 1]$ . We conclude the proof of the stability under Lie bracket by using the semigroup structure of  $\bar{\mathcal{R}}$ .

Concerning the stability under addition, set  $V(s) = e^{sX} e^{sY}$  and notice that  $V(s) \in \bar{\mathcal{R}}$  for every  $s \in \mathbb{R}$ . Noticing that  $V(t/n)^n \xrightarrow[n \rightarrow \infty]{d_{\mathcal{F}}} e^{t(X+Y)}$ , we deduce that  $e^{t(X+Y)} \in \bar{\mathcal{R}}$  for every  $t \in \mathbb{R}$ .  $\square$

Denote by  $\text{ad}_X(Y) = [X, Y]$  the adjoint representation both in  $\mathfrak{su}_2$  and in  $\mathcal{C}^0(\mathcal{D}, \mathfrak{su}_2)$ . We recall the Pauli matrices commutation laws

$$[i\sigma_x, i\sigma_y] = -i\sigma_z, \quad [i\sigma_y, i\sigma_z] = -i\sigma_x, \quad [i\sigma_z, i\sigma_x] = -i\sigma_y.$$

After some straightforward computations, one gets

$$\begin{aligned} \text{ad}_{\delta i\sigma_x}^{2l}(\text{ad}_{(E+\alpha)i\sigma_z}^{2k+1}(\delta i\sigma_x)) &= (-1)^{l+k}(E+\alpha)^{2k+1}\delta^{2l+1}i\sigma_y, \\ \text{ad}_{\delta i\sigma_x}^{2l+1}(\text{ad}_{(E+\alpha)i\sigma_z}^{2k+1}(\delta i\sigma_x)) &= (-1)^{l+k}(E+\alpha)^{2k+1}\delta^{2l+2}i\sigma_z, \\ \text{ad}_{(E+\alpha)i\sigma_z}^{2l}\text{ad}_{\delta i\sigma_x}^{2l}(\text{ad}_{(E+\alpha)i\sigma_z}^{2k+1}(\delta i\sigma_x)) &= (-1)^{l+k+1}(E+\alpha)^{2k+2}\delta^{2l+1}i\sigma_x. \end{aligned}$$

Thus for any  $n, m \in \mathbb{N}$ , and any sequence  $(b_{k,l})_{k,l}$  of real numbers, we have

$$\begin{aligned} \sum_{k=0}^m \sum_{l=0}^n b_{k,l} \delta^{2k+2} (E+\alpha)^{2l+1} i\sigma_x &\in \mathfrak{g}, \\ \sum_{k=0}^m \sum_{l=0}^n c_{k,l} \delta^{2k+1} (E+\alpha)^{2l+1} i\sigma_y &\in \mathfrak{g}, \\ \sum_{k=0}^m \sum_{l=0}^n d_{k,l} \delta^{2k+1} (E+\alpha)^{2l+2} i\sigma_z &\in \mathfrak{g}. \end{aligned}$$

By the Stone–Weierstrass theorem, for any continuous function  $f \in C(\mathcal{D}, \mathbb{R})$  we can approximate  $\frac{f(d)}{(E+\alpha)\delta^2}$  uniformly on  $\mathcal{D}$  by polynomials of the form  $\sum_{k=0}^m \sum_{l=0}^n b_{k,l} \delta^{2k} (E+\alpha)^{2l}$ . This proves

that  $f(d)i\sigma_x \in \mathfrak{g}$ . With a similar argument, we get  $f(d)i\sigma_\star \in \mathfrak{g}$  for  $\star = x, y, z$ .

Finally, let  $\rho > 0$  be such that  $(a_1, a_2, a_3) \mapsto e^{a_1 i\sigma_x + a_2 i\sigma_y + a_3 i\sigma_z}$  is a diffeomorphism between a neighborhood of 0 in  $\mathbb{R}^3$  and the ball of radius  $\rho$  centered at  $I_2$  in  $SU_2$ . Then for every  $M_F \in \mathcal{F}$  such that  $d_{\mathcal{F}}(M_F, I_2) < \rho$  there exist  $f_1, f_2, f_3 \in \mathcal{C}^0(\mathcal{D}, \mathbb{R})$  such that  $M_F(d) = e^{f_1(d)i\sigma_x + f_2(d)i\sigma_y + f_3(d)i\sigma_z}$ . Thus  $M_F \in \bar{\mathcal{R}}$ . Since  $\bar{\mathcal{R}}$  is a semigroup, we deduce that  $\bar{\mathcal{R}}$  is both open and closed in  $\mathcal{F}$ , yielding that  $\bar{\mathcal{R}} = \mathcal{F}$ . This concludes the proof of Theorem V.10.

## V.3 Proof of Theorem V.3

### V.3.1 A first change of variables

Let  $w_{\varepsilon_1, \varepsilon_2}$  be as in (V.10). In order to recast the equation

$$i \frac{d}{dt} \psi = H\psi = ((E + \alpha)\sigma_z + w_{\varepsilon_1, \varepsilon_2} \sigma_x) \psi$$

in the interaction frame, set

$$\psi_{\text{I}}(t) = e^{i(E+\alpha)\sigma_z t} \psi(t), \quad E_1(t) = 2\alpha t - \frac{\Delta(\varepsilon_1 \varepsilon_2 t)}{\varepsilon_1 \varepsilon_2}, \quad E_2(t) = 4Et + 2\alpha t + \frac{\Delta(\varepsilon_1 \varepsilon_2 t)}{\varepsilon_1 \varepsilon_2},$$

and notice that

$$i \frac{d}{dt} \psi_{\text{I}} = H_{\text{I}} \psi_{\text{I}},$$

where

$$\begin{aligned} H_{\text{I}}(t) &= -(E + \alpha)\sigma_z + e^{i(E+\alpha)\sigma_z t} H(t) e^{-i(E+\alpha)\sigma_z t} \\ &= \varepsilon_1 u(\varepsilon_1 \varepsilon_2 t) \begin{pmatrix} 0 & e^{iE_1(t)} + e^{iE_2(t)} \\ e^{-iE_1(t)} + e^{-iE_2(t)} & 0 \end{pmatrix}. \end{aligned}$$

We will assume without loss of generality that  $T = 1$ . For  $E \in \mathbb{R}$ , define

$$A(E) = \begin{pmatrix} 0 & e^{iE} \\ e^{-iE} & 0 \end{pmatrix}, \quad B(E) = \begin{pmatrix} 0 & -ie^{iE} \\ ie^{-iE} & 0 \end{pmatrix}. \quad (\text{V.15})$$

In terms of these new notations, we can rewrite  $H_{\text{I}}(t) = \varepsilon_1 u(\varepsilon_1 \varepsilon_2 t) A(E_1(t)) + \varepsilon_1 u(\varepsilon_1 \varepsilon_2 t) A(E_2(t))$ ,  $t \in [0, \frac{1}{\varepsilon_1 \varepsilon_2}]$ .

In the usual first order RWA setting, one neglects the term containing the factor  $A(E_2)$ , which is highly oscillating compared to the first one. A standard method to justify this, is to use a change of variables close to the identity (see, e.g., [Rou08] and [SVM07]). Inspired by this, we introduce the notation

$$f_1(t) = \frac{d}{dt} E_1(t), \quad f_2(t) = \frac{d}{dt} E_2(t), \quad (\text{V.16})$$

and we apply the unitary change of variables

$$\tilde{\psi}_{\text{I}}(t) = \exp\left(i\varepsilon_1 \frac{u(\varepsilon_1 \varepsilon_2 t)}{f_2(t)} B(E_2(t))\right) \psi_{\text{I}}(t). \quad (\text{V.17})$$

The dynamics of  $\tilde{\psi}_I$  are characterized by the Hamiltonian

$$\begin{aligned} \tilde{H}_I(t) &= i \frac{d}{dt} \left( \cos \left( \varepsilon_1 \frac{u(\varepsilon_1 \varepsilon_2 t)}{f_2(t)} \right) I + i \sin \left( \varepsilon_1 \frac{u(\varepsilon_1 \varepsilon_2 t)}{f_2(t)} \right) B(E_2(t)) \right) \\ &\quad \left( \cos \left( \varepsilon_1 \frac{u(\varepsilon_1 \varepsilon_2 t)}{f_2(t)} \right) I - i \sin \left( \varepsilon_1 \frac{u(\varepsilon_1 \varepsilon_2 t)}{f_2(t)} \right) B(E_2(t)) \right) \\ &\quad + \left( \cos \left( \varepsilon_1 \frac{u(\varepsilon_1 \varepsilon_2 t)}{f_2(t)} \right) I + i \sin \left( \varepsilon_1 \frac{u(\varepsilon_1 \varepsilon_2 t)}{f_2(t)} \right) B(E_2(t)) \right) H_I \\ &\quad \left( \cos \left( \varepsilon_1 \frac{u(\varepsilon_1 \varepsilon_2 t)}{f_2(t)} \right) I - i \sin \left( \varepsilon_1 \frac{u(\varepsilon_1 \varepsilon_2 t)}{f_2(t)} \right) B(E_2(t)) \right). \end{aligned}$$

Notice that the first term can be rewritten as  $-\varepsilon_1 u(\varepsilon_1 \varepsilon_2 t) A(E_2(t)) + O(\varepsilon_1^2)$ , so that  $\tilde{H}_I(t) = \varepsilon_1 A(E_1(t)) + O(\varepsilon_1^2 + \varepsilon_1^2 \varepsilon_2)$ , where the notation  $O(\cdot)$  is defined as follows.

**Definition V.14.** Let  $R$  be a  $(\varepsilon_1, \varepsilon_2)$ -parameterized function in the following sense: for every  $\varepsilon_1, \varepsilon_2 > 0$ ,  $R_{\varepsilon_1, \varepsilon_2}$  is a real-valued function defined on the interval  $[0, \frac{1}{\varepsilon_1 \varepsilon_2}]$ . We say that  $R = O(g(\varepsilon_1, \varepsilon_2))$  with  $g : \mathbb{R}_+^2 \rightarrow \mathbb{R}_+$  if there exist  $\delta, C > 0$  such that for every  $(\varepsilon_1, \varepsilon_2) \in (0, \delta)^2$  and  $t \in [0, \frac{1}{\varepsilon_1 \varepsilon_2}]$ , we have  $|R_{\varepsilon_1, \varepsilon_2}(t)| \leq Cg(\varepsilon_1, \varepsilon_2)$ .

**Remark V.15.** — We have  $|\psi_I - \tilde{\psi}_I| = O(\varepsilon_1)$ . Moreover, from the hypotheses of Theorem V.3, we have  $u(0) = u(1) = 0$ , thus  $\tilde{\psi}_I(0) = \psi_I(0)$  and  $\tilde{\psi}_I(\frac{1}{\varepsilon_1 \varepsilon_2}) = \psi_I(\frac{1}{\varepsilon_1 \varepsilon_2})$ .

— Let  $\psi_{\text{rwa}}$  be the solution of the Schrödinger equation with initial condition  $\psi_I(0)$  and Hamiltonian  $\varepsilon_1 u(\varepsilon_1 \varepsilon_2 t) A(E_1(t))$ . Then it turns out that  $|\psi_{\text{rwa}}(\frac{1}{\varepsilon_1 \varepsilon_2}) - \tilde{\psi}_I(\frac{1}{\varepsilon_1 \varepsilon_2})| = O(\varepsilon_1 / \varepsilon_2)$  (see Lemma V.29). To prove convergence as  $(\varepsilon_1, \varepsilon_2) \rightarrow 0$  in a suitable asymptotic regime, it would thus be enough to show that the dynamics of  $\varepsilon_1 u(\varepsilon_1 \varepsilon_2 t) A(E_1(t))$  induce a transition between  $(0, 1)$  and  $(1, 0)$  up to a phase, in the regime  $\varepsilon_1 \ll \varepsilon_2$ . Nevertheless this is not the case (recall that ‘standard’ adiabatic theorem cannot be applied since  $\varepsilon_1$  is not fixed) as illustrated in Figure V.4.

### V.3.2 Idea of the proof

We aim at providing correction terms to the Hamiltonian  $\varepsilon_1 u(\varepsilon_1 \varepsilon_2 t) A(E_1(t))$ , in order to improve the order of the averaging approximation. For this we will repeat a procedure similar to the one introduced in Equation (V.17). At each step the expression of the obtained effective Hamiltonian is more complicated but provides a more accurate estimate of the final state. Then it will be possible to apply adiabatic theory to prove transition for the effective Hamiltonian. More precisely, we will prove the following theorem.

**Theorem V.16.** Let  $\alpha \in (v_0, v_1)$  and assume that  $E + \alpha > 0$  and  $4E - 3\Delta'(s) > 2\alpha$  for every  $s \in [0, 1]$ . Then, for every  $N_0 \in \mathbb{N}$  there exists a Hamiltonian  $H_{\text{RWA}}$  of the form

$$H_{\text{RWA}}(t) = \varepsilon_1 h_1(\varepsilon_1 \varepsilon_2 t) A(E_1(t)) + \varepsilon_1^2 h_2(\varepsilon_1 \varepsilon_2 t) B(E_2(t)) + \varepsilon_1^2 h_3(\varepsilon_1 \varepsilon_2 t) \sigma_z, \quad (\text{V.18})$$

with  $h_1, h_2, h_3$  polynomials in  $(\varepsilon_1, \varepsilon_2)$  with coefficients in  $C^\infty([0, 1], \mathbb{R})$ , such that the solution  $\psi_{\text{RWA}}$  of the Cauchy problem

$$i \frac{d}{dt} \psi_{\text{RWA}} = H_{\text{RWA}} \psi_{\text{RWA}}, \quad \psi_{\text{RWA}}(0) = \psi_I(0),$$

satisfies  $|\psi_{\text{RWA}}(\frac{1}{\varepsilon_1 \varepsilon_2}) - \psi_I(\frac{1}{\varepsilon_1 \varepsilon_2})| = O(\varepsilon_1^2 \varepsilon_2 + \varepsilon_1^{N_0-1} / \varepsilon_2)$ . More precisely, there exist  $h_{j,p,q} \in C^\infty([0, 1], \mathbb{R})$ , for  $j = 1, 2, 3$ ,  $p = 0, \dots, N_0$ , and  $q = 0, 1$ , such that

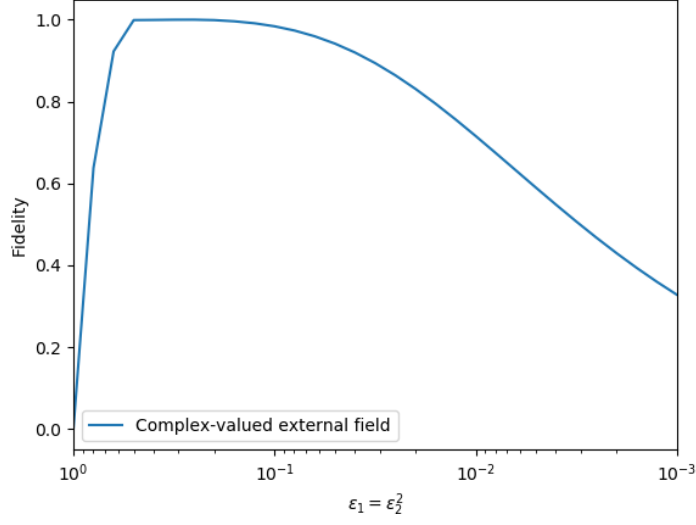


Figure V.4 – Taking  $v_0 = -0.5$ ,  $v_1 = 0.5$ ,  $E = 0$ , and  $\alpha = 0$ , we observe that the fidelity does not converge to 1 as  $(\varepsilon_1, \varepsilon_2) \rightarrow 0$  in the regime  $\varepsilon_1 \ll \varepsilon_2$ . The plot corresponds to the choice  $\varepsilon_1 = \varepsilon_2^2$ .

1.  $h_1 = u + \sum_{p=1}^{N_0} \sum_{q=0}^1 \varepsilon_1^p \varepsilon_2^q h_{1,p,q}$  with  $h_{1,p,0}(0) = h_{1,p,0}(1) = 0$ ,
2.  $h_2 = \sum_{p=0}^{N_0} \sum_{q=0}^1 \varepsilon_1^p \varepsilon_2^q h_{2,p,q}$  with  $h_{2,p,0}(0) = h_{2,p,0}(1) = 0$ ,
3.  $h_3 = \sum_{p=0}^{N_0} \sum_{q=0}^1 \varepsilon_1^p \varepsilon_2^q h_{3,p,q}$  with  $h_{3,p,0}(0) = h_{3,p,0}(1) = 0$ .

After that, we will prove that  $H_{\text{RWA}}$  induces a transition between eigenstates with an error of order  $O(\varepsilon_2/\varepsilon_1)$ , which will be enough to prove Theorem V.3.

### V.3.3 The rotating wave approximation

**Definition V.17.** Define the algebra  $\mathcal{S}$  of slow functions as the set of all  $(\varepsilon_1, \varepsilon_2)$ -parameterized functions  $f$  (in the sense of Definition V.14) such that for every  $t \in [0, \frac{1}{\varepsilon_1 \varepsilon_2}]$ ,  $f_{\varepsilon_1, \varepsilon_2}(t) = g(\varepsilon_1 \varepsilon_2 t)$  for some smooth  $g : [0, 1] \rightarrow \mathbb{R}$  independent of  $(\varepsilon_1, \varepsilon_2)$ . The quantity  $\sup_{t \in [0, \frac{1}{\varepsilon_1 \varepsilon_2}]} |f_{\varepsilon_1, \varepsilon_2}(t)|$  is independent of  $(\varepsilon_1, \varepsilon_2)$  and provides a norm, endowing  $\mathcal{S}$  with the structure of Banach algebra.

**Remark V.18.** — The functions  $f_1$  and  $f_2$  defined in (V.16) are slow.

- $\mathcal{S}$  is isometric to the Banach algebra  $\mathcal{C}^\infty([0, 1], \mathbb{R})$ .
- Given  $f \in \mathcal{S}$ , its  $t$ -derivative  $\dot{f}$  defined by  $f_{\varepsilon_1, \varepsilon_2}(t) = \frac{d}{dt} f_{\varepsilon_1, \varepsilon_2}(t)$  is such that  $\frac{1}{\varepsilon_1 \varepsilon_2} \dot{f} \in \mathcal{S}$ .

For every  $j \in \mathbb{Z}$ , let us introduce the notations

$$\begin{aligned} \Lambda_j &= (j+1)E_1 - jE_2, & \tilde{\phi}_j &= jE_1 - jE_2, \\ \lambda_j &= (j+1)f_1 - jf_2, & \phi_j &= jf_1 - jf_2. \end{aligned} \tag{V.19}$$



**Definition V.19.** Define the set

$$G = \{\pm Z(\Lambda_p), \pm \cos(\Phi_p)\sigma_z, \pm \sin(\Phi_p)\sigma_z \mid Z \in \{A, B\}, p \in \mathbb{Z}\}.$$

We say that an element of  $G$  is oscillating if its associated integer  $p$  is different from 0.

**Lemma V.20.**  $G$  has the following stability properties:

1.  $\forall p \in \mathbb{Z}, \forall X \in G, \cos(\Phi_p)X$  and  $\sin(\Phi_p)X$  are in  $\text{span}_{\mathbb{R}}G$ ;
2.  $\forall X, Y \in G, i[X, Y] \in \text{span}_{\mathbb{R}}G$ ;
3.  $\forall X, Y \in G, XYX \in \text{span}_{\mathbb{R}}G$ .

*Proof.* The first point is a consequence of the fact that  $\{\Phi_p \mid p \in \mathbb{Z}\}$  is a group for the addition. Thus, for every  $p, q \in \mathbb{Z}$ ,

$$2 \cos(\Lambda_p) \cos(\Phi_q)\sigma_z = \cos(\Phi_p + \Phi_q)\sigma_z + \cos(\Phi_p - \Phi_q)\sigma_z \in \text{span}_{\mathbb{R}}G.$$

Moreover,  $2 \cos(\Phi_p)A(\Lambda_q) = A(\Lambda_{q+p}) + A(\Lambda_{q-p}) \in \text{span}_{\mathbb{R}}G$ . The remaining cases can be checked similarly.

For the second point, for every  $E, E', E'' \in \mathbb{R}$ ,

$$\begin{aligned} i[A(E), A(E')] &= -2 \sin(E - E')\sigma_z, \\ i[A(E), \cos(E')\sigma_z] &= 2 \cos(E')B(E) = B(E + E') + B(E - E'), \\ i[\cos(E')\sigma_z, \cos(E'')\sigma_z] &= 0. \end{aligned}$$

Using the fact that, for every  $p \in \mathbb{Z}, A(\Lambda_p - \pi/2) = B(\Lambda_p)$ , we obtain that  $i[A(\Lambda_p), G] \in \text{span}_{\mathbb{R}}G$ . Similar results can easily be obtained for  $B(\Lambda_p), \cos(\Phi_p)\sigma_z$ , and  $\sin(\Phi_p)\sigma_z$ .

The last point relies on the relations

$$\begin{aligned} A(E)A(E')A(E) &= A(2E - E'), \\ \cos(E')A(E)\sigma_z A(E) &= -\cos(E')\sigma_z, \\ 2 \cos^2(E')\sigma_z A(E)\sigma_z &= A(E) + \frac{1}{2}(A(E + 2E') + A(E - 2E')), \\ 2 \cos^2(E') \cos(E'')\sigma_z^3 &= \left( \cos(E'') + \frac{1}{2}(\cos(2E' + E'') + \cos(2E' - E'')) \right) \sigma_z. \end{aligned}$$

□

**Definition V.21.** Define the vector space  $\mathcal{E}$  as the set of entire series in  $(\varepsilon_1, \varepsilon_2)$  with coefficients in the set  $\text{span}_{\mathcal{S}}G$ , i.e.,

$$\left\{ \sum_{j,k \geq 0} \varepsilon_1^j \varepsilon_2^k \sum_{g \in H_{j,k}} s_g g \mid H_{j,k} \subset G \text{ finite, } s_g \in \mathcal{S}, \sum_{j,k \geq 0} \varepsilon_1^j \varepsilon_2^k \sum_{g \in H_{j,k}} |s_g| < \infty \text{ for } (\varepsilon_1, \varepsilon_2) \text{ small enough} \right\}.$$

### V.3.3.1 The elimination procedure

In order to generalize (V.17), we introduce the operation of elimination of an oscillating term of a coefficient of  $\mathcal{E}$ .

**Definition V.22.** Define the operation  $\text{Pr} : G \rightarrow G$  by the relations  $\text{Pr}(\pm A(\Lambda_p)) = \pm B(\Lambda_p)$ ,  $\text{Pr}(\pm B(\Lambda_p)) = \mp A(\Lambda_p)$ ,  $\text{Pr}(\pm \cos(\Phi_p)\sigma_z) = \pm \sin(\Phi_p)\sigma_z$ , and  $\text{Pr}(\pm \sin(\Phi_p)\sigma_z) = \mp \cos(\Phi_p)\sigma_z$ .

**Definition V.23.** Let  $H \in \mathcal{E}$  and  $Z(E)$  be an oscillating term of  $G$  ( $E = \Lambda_p$  if  $Z \in \{A, B\}$  or  $E = \Phi_p$  if  $Z(E) \in \{\cos(E)\sigma_z, \sin(E)\sigma_z\}$ ). Suppose that  $f = \dot{E}$  (which is necessarily slow) is nowhere vanishing. Fix  $j \geq 1$ ,  $k \geq 0$ ,  $s \in \mathcal{S}$  and let  $c = \varepsilon_1^j \varepsilon_2^k s$ . The operation of elimination of  $cZ(E)$  from  $H$  is defined as

$$\begin{aligned} \text{El}(c, Z(E))(H) &= i \frac{d}{dt} \left[ \exp \left( i(c/f) \text{Pr}(Z)(E) \right) \right] \exp \left( -i(c/f) \text{Pr}(Z)(E) \right) \\ &\quad + \exp \left( i(c/f) \text{Pr}(Z)(E) \right) H \exp \left( -i(c/f) \text{Pr}(Z)(E) \right), \quad (\text{V.20}) \\ \tilde{\text{El}}(c, Z(E))(\psi) &= \exp \left( ic/f \text{Pr}(Z)(E) \right) \psi. \end{aligned}$$

In fact, the elimination procedure is the generalization of the change of variables in Equation (V.17). It transforms the Hamiltonian dynamics  $i \frac{d}{dt} \psi = H \psi$  into the dynamics  $i \frac{d}{dt} \eta = \text{El}(c, Z(E))(H) \eta$ , where  $\eta = \tilde{\text{El}}(c, Z(E)) \psi$ . The term *elimination* is motivated by the following lemma, stating that the procedure described above generates in the transformed Hamiltonian only terms of degree higher than  $\varepsilon_1^j \varepsilon_2^k$ .

**Lemma V.24.** Take  $H, Z(E), j, k, c$  as in Definition V.23. Then  $\text{El}(c, Z(E))(H) \in \mathcal{E}$ . Besides, if  $H = O(\varepsilon_1)$  then  $\text{El}(c, Z(E))(H + cZ(E)) = H + O(\varepsilon_1^{j+1} \varepsilon_2^k)$ .

*Proof.* First recall that for each matrix  $M$  such that  $M^2 = I$  and each  $c \in \mathbb{R}$ ,  $\exp(icM) = \cos(c)I + i \sin(c)M$ . As  $A(E)^2 = B(E)^2 = \sigma_z^2 = I$ , we can give an explicit expression for  $\text{El}(c, Z(E))(H)$ .

Let us start from the case  $Z(E) = A(\Lambda_p)$ , for which we have

$$\text{El}(c, A(\Lambda_p))(H) = J_1 + J_2 + J_3, \quad (\text{V.21})$$

where

$$\begin{aligned} J_1 &= i \frac{d}{dt} (c/f) \left( -\sin(c/f)I + i \cos(c/f)B(\Lambda_p) \right) \left( \cos(c/f)I - i \sin(c/f)B(\Lambda_p) \right) \\ &= -\frac{d}{dt} (c/f) B(\Lambda_p), \\ J_2 &= -\sin(c/f) \frac{d}{dt} B(\Lambda_p) \left( \cos(c/f)I - i \sin(c/f)B(\Lambda_p) \right) \\ &= -f \sin(c/f) A(\Lambda_p) \left( \cos(c/f)I - i \sin(c/f)B(\Lambda_p) \right), \\ J_3 &= \left( \cos(c/f)I + i \sin(c/f)B(\Lambda_p) \right) H \left( \cos(c/f)I - i \sin(c/f)B(\Lambda_p) \right). \end{aligned}$$

The term  $J_1$  is obviously an element of  $\mathcal{E}$ . Besides,  $\cos(c/f)$  and  $\sin(c/f)$  are entire series in  $\varepsilon_1, \varepsilon_2$  with coefficients in  $\mathcal{S}$ . Thus,

$$J_2 = -f \sin(c/f) \cos(c/f) A(\Lambda_p) - f \sin^2(c/f) \sigma_z$$

is also an element of  $\mathcal{E}$ . The last term to be considered is

$$J_3 = \cos^2(c/f) H + \cos(c/f) \sin(c/f) i [B(\Lambda_p), H] + \sin^2(c/f) B(\Lambda_p) H B(\Lambda_p).$$

Thanks to Lemma V.20,  $J_3$  is then the sum of elements of  $\mathcal{E}$ .

Let us now assume that  $H = O(\varepsilon_1)$  and focus on the order of each term (in the case  $Z(E) = A(\Lambda_p)$ ). We notice that  $J_1 = O(\varepsilon_1^{j+1}\varepsilon_2^{k+1})$  as  $\frac{d}{dt}(s/f) = O(\varepsilon_1\varepsilon_2)$  and  $J_2 = -cA(\Lambda_p) + O(\varepsilon_1^{j+1}\varepsilon_2^k)$ . Finally,  $J_3 = H + (c/f)i[B(\Lambda_p), H] + O(\varepsilon_1^{j+1}\varepsilon_2^k)$ . As  $H = O(\varepsilon_1)$ , we get  $(c/f)i[B(\Lambda_p), H] = O(\varepsilon_1^{j+1}\varepsilon_2^k)$ . Thus

$$\text{El}(c, Z(E))(H + cZ(E)) = -cZ(E) + H + cZ(E) + O(\varepsilon_1^{j+1}\varepsilon_2^k).$$

The same computations as above work for the case  $Z(E) = B(\Lambda_p)$ .

In the case  $Z(E) = \cos(\Phi_p)\sigma_z$  we have

$$\text{El}(c, \cos(\Phi_p)\sigma_z)(H) = J_1 + J_2 + J_3, \quad (\text{V.22})$$

where

$$J_1 = -\frac{d}{dt}(c/f) \sin(\Phi_p)\sigma_z,$$

$$J_2 = -c \cos(\Phi_p)\sigma_z,$$

$$J_3 = (\cos(c/f \sin(\Phi_p))I + i \sin(c/f \sin(\Phi_p))\sigma_z)H(\cos(c/f \sin(\Phi_p))I - i \sin(c/f \sin(\Phi_p))\sigma_z).$$

Note that  $\sin(c/f \sin(\Phi_p))$  and  $\cos(c/f \sin(\Phi_p))$  can be developed as entire series in  $\varepsilon_1, \varepsilon_2$  with coefficients in  $\mathcal{S} \cos(\Phi_q)$  and  $\mathcal{S} \sin(\Phi_q)$  for  $q \in \mathbb{Z}$ . Lemma V.20 ensures that  $\text{El}(c, \cos(\Phi_p)\sigma_z)(H)$  is an element of  $\mathcal{E}$ . The computations of the order of the terms when  $H = O(\varepsilon_1)$  are similar to those made above, and one can apply the same reasoning to  $\text{El}(c, \sin(\Phi_p)\sigma_z)(H)$ .  $\square$

A key assumption of Lemma V.24 above is that  $f$  is nowhere vanishing. The following result ensures that this is the case for all frequencies of the oscillating terms in  $G$ .

**Lemma V.25.** *Let  $j \in \mathbb{Z}$  be nonzero. Then the functions  $\lambda_j$  and  $\phi_j$ , defined as in (V.19), are nowhere vanishing in  $[0, \frac{1}{\varepsilon_1\varepsilon_2}]$ .*

*Proof.* Let us first prove that

$$2f_1(t) < f_2(t), \quad \forall t \in \left[0, \frac{1}{\varepsilon_1\varepsilon_2}\right], \quad (\text{V.23})$$

where we recall that  $f_1, f_2$  are defined in (V.16). Indeed,

$$2f_1(t) - f_2(t) = 2\alpha - 4E - 2\Delta'(\varepsilon_1\varepsilon_2t) - \Delta'(\varepsilon_1\varepsilon_2t) < 2v_1 - 4E - 6v_0,$$

where we used the inequality  $\alpha < v_1$  and the fact that, according to the hypotheses of Theorem V.3,  $\Delta'$  is increasing from  $2v_0$  to  $2v_1$ . The inequality  $2v_1 - 4E - 6v_0 = 2(E + v_1) - 6(E + v_0) \leq 0$ , corresponding to the assumption  $3(E + v_0) \geq E + v_1$  of Theorem V.3, concludes the proof of (V.23).

Moreover,

$$f_2(t) = 4E + 2\alpha + \Delta'(t) \geq 4(E + v_0) \geq \frac{4(E + v_1)}{3} > 0, \quad \forall t \in \left[0, \frac{1}{\varepsilon_1\varepsilon_2}\right].$$

In particular,  $f_1 - f_2 = -4E - 2\Delta' \leq -4(E + v_0) < 0$ . This implies that  $\phi_j$  never vanishes for  $j \neq 0$ . Finally, for  $j > 0$ ,  $\lambda_j = (j+1)f_1 - jf_2 = (j-1)(f_1 - f_2) + 2f_1 - f_2 < 0$ , and, similarly,  $\lambda_j = (j+1)(f_1 - f_2) + f_2 > 0$  for  $j < 0$ .  $\square$

### V.3.3.2 Algorithm description

We can now introduce an algorithm to simplify the Hamiltonian  $H_I$ . The cleaning operation  $\text{cl}_{\bar{p}}(p_0, q_0)$ , with  $p_0 \leq \bar{p}$ , consists in eliminating from  $H_I$  all oscillating terms of degree  $\varepsilon_1^p \varepsilon_2^q$  for

$$\begin{cases} p \leq \bar{p} \\ q < q_0 \end{cases} \quad \text{and} \quad \begin{cases} p \leq p_0 \\ q = q_0 \end{cases}$$

in lexicographic order on  $(p, q)$ .

The algorithm is constructed by induction, as follows:

- $\text{cl}_p(0, 0) = H_I$ ;
- for  $0 \leq p' < p$ ,  $\text{cl}_p(p' + 1, q)$  is obtained from  $\text{cl}_p(p', q)$  by eliminating one by one all its oscillating terms of degree  $(p' + 1, q)$ , using Lemma V.24;
- $\text{cl}_p(0, q + 1) = \text{cl}_p(p, q)$ . Notice that, by construction, there is no term of degree  $(0, q + 1)$  in  $\text{cl}_p(p, q)$ .

Associated with the transformed Hamiltonian  $\text{cl}_{p_0}(p, q)$ , we define  $\tilde{\text{cl}}_{p_0}(p, q)$  the variable obtained iteratively from  $\psi_I$  by applying, at every use of Lemma V.24, the corresponding transformation  $\tilde{\text{El}}$ .

**Remark V.26.** *According to Lemma V.24, each elimination procedure produces only terms of higher degree, thus the algorithm yielding  $\text{cl}_{p_0}(p, q)$  ends after a finite number of steps.*

When we apply the algorithm, we first deal with monomials of the type  $\varepsilon_1^p \varepsilon_2^0$ ,  $p \geq 1$ . The following lemma provides a useful property concerning their corresponding coefficients.

**Lemma V.27.** *Define  $\mathcal{S}_0 = \{s \in \mathcal{S} \mid s(0) = s(\frac{1}{\varepsilon_1 \varepsilon_2}) = 0\}$ . Given  $p, p', q \in \mathbb{N}$  with  $p' \leq p$ , consider the decomposition  $\text{cl}_p(p', q) = H_1 + \varepsilon_2 H_2$ , where  $H_1$  is an entire series in  $\varepsilon_1$  with coefficient in  $\text{span}_{\mathcal{S}} G$  and  $H_2 \in \mathcal{E}$  ( $H_1$  collects all the monomials of the type  $\varepsilon_1^n \varepsilon_2^0$ ). Then the coefficients of  $H_1$  are in  $\text{span}_{\mathcal{S}_0} G$ .*

*Proof.* Let us first consider the case  $q = 0$ . Then  $H = H_1 + \varepsilon_2 H_2$  and we want to eliminate an element  $cZ(E)$  with  $c = \varepsilon_1^{p+1} s$  and  $s \in \mathcal{S}_0$  using Formula (V.20). Notice that

$$-\frac{d}{dt}(c/f)\text{Pr}(Z)(E) \quad \text{and} \quad \exp\left(i(c/f)\text{Pr}(Z)(E)\right)\varepsilon_2 H_2 \exp\left(-i(c/f)\text{Pr}(Z)(E)\right)$$

only consist of monomials of the type  $\varepsilon_1^n \varepsilon_2^m$  with  $m \geq 1$ .

On the other hand, the terms  $J_1$  and  $J_2$  in Equations (V.21) and (V.22) (and the corresponding ones for  $Z(E) = B(E)$  and  $Z(E) = \sin(E)\sigma_z$ ) are clearly in  $\text{span}_{\mathcal{S}_0} G$ . Besides, the coefficients of

$$\exp\left(i(c/f)\text{Pr}(Z)(E)\right)H_1 \exp\left(-i(c/f)\text{Pr}(Z)(E)\right)$$

also stay in  $\text{span}_{\mathcal{S}_0} G$ , as  $\mathcal{S}_0$  is a subalgebra.

In the case  $q \neq 0$ , the elimination of a term of degree  $(p, q)$  with  $q \geq 1$  does not impact the monomials of the type  $\varepsilon_1^n \varepsilon_2^0$ , according to Lemma V.24.  $\square$

Let  $G_0$  be the set of non-oscillating elements of  $G$ .

**Lemma V.28.** *Assume that (V.23) holds. Then we have*

$$\text{cl}_{N_0}(N_0, 1) = \varepsilon_1 H_{N_0} + \varepsilon_1^{N_0+1} H_{r, N_0} + \varepsilon_1^2 \varepsilon_2 H'_{N_0} + \varepsilon_1^{N_0+1} \varepsilon_2 H'_{N_0, r} + \varepsilon_1^3 \varepsilon_2^2 H''_r,$$

where

1.  $H_{N_0}$  is a polynomial of degree  $N_0 - 1$  in  $\varepsilon_1$  with coefficients in  $\text{span}_{\mathcal{S}_0} G_0$ ,
2.  $H'_{N_0}$  is a polynomial of degree  $N_0 - 2$  in  $\varepsilon_1$  with coefficients in  $\text{span}_{\mathcal{S}} G_0$ ,
3.  $H_{r,N_0}$  is an entire series in  $\varepsilon_1$  with coefficients in  $\text{span}_{\mathcal{S}_0} G$ ,
4.  $H'_{r,N_0}$  is an entire series in  $\varepsilon_1$  with coefficients in  $\text{span}_{\mathcal{S}} G$ ,
5.  $H''_r$  is an entire series in  $\varepsilon_1, \varepsilon_2$  with coefficients in  $\text{span}_{\mathcal{S}} G$ .

*Proof.* Points 1 and 3 follow from Lemma V.27, while points 2, 4, and 5 follow from Lemma V.24.  $\square$

Noticing that, in particular,  $\text{cl}_{N_0}(N_0, 1) = \varepsilon_1 H_{N_0} + \varepsilon_1^2 \varepsilon_2 H'_{N_0} + O(\varepsilon_1^3 \varepsilon_2^2 + \varepsilon_1^{N_0})$ , we introduce the truncation  $H_{\text{RWA}} = \varepsilon_1 H_{N_0} + \varepsilon_1^2 \varepsilon_2 H'_{N_0}$  of  $\text{cl}_{N_0}(N_0, 1)$  and we denote by  $\psi_{\text{RWA}}$  the solution of

$$i \frac{d}{dt} \psi_{\text{RWA}} = H_{\text{RWA}} \psi_{\text{RWA}}, \quad \psi_{\text{RWA}}(0) = \psi_{N_0}(0), \quad (\text{V.24})$$

where  $\psi_{N_0} = \tilde{\text{cl}}_{N_0}(N_0, 1)$ . Notice that, even if we are using the same notation  $\psi_{\text{RWA}}$ , we are considering here a RWA of higher-order than the one discussed in Remark V.15.

**Lemma V.29.** *We have the following estimates:*

1.  $|\psi_{N_0}(\frac{1}{\varepsilon_1 \varepsilon_2}) - \psi_{\text{I}}(\frac{1}{\varepsilon_1 \varepsilon_2})| = O(\varepsilon_1^2 \varepsilon_2)$ ;
2.  $|\psi_{N_0}(\frac{1}{\varepsilon_1 \varepsilon_2}) - \psi_{\text{RWA}}(\frac{1}{\varepsilon_1 \varepsilon_2})| = O(\varepsilon_1^2 \varepsilon_2 + \varepsilon_1^{N_0-1} / \varepsilon_2)$ .

*Proof.* By Lemma V.27, all the changes of variable used for obtaining  $\text{cl}_{N_0}(N_0, 0)$  from  $H_{\text{I}}$  are of the form  $\tilde{\text{El}}(c, Z(E))(\psi)$  with  $c = \varepsilon_1^p s$ ,  $s \in \mathcal{S}_0$ . Thus  $\psi_{\text{I}}(0) = \tilde{\text{cl}}(N_0, 0)(0)$  and  $\psi_{\text{I}}(\frac{1}{\varepsilon_1 \varepsilon_2}) = \tilde{\text{cl}}_{N_0}(N_0, 0)(\frac{1}{\varepsilon_1 \varepsilon_2})$ . Such changes of variable preserve the state at the initial and final time. After that we applied finitely many changes of variable of the form  $\psi \mapsto \exp(i\varepsilon_1^p \varepsilon_2^q s Z(E))\psi$  with  $p \geq 2$  and  $q = 1$ . Thus

$$\sup_{t \in [0, \frac{1}{\varepsilon_1 \varepsilon_2}]} |\tilde{\text{cl}}_{N_0}(N_0, 0)(t) - \psi_{N_0}(t)| = O(\varepsilon_1^2 \varepsilon_2), \quad (\text{V.25})$$

which concludes the proof of the first estimate.

Notice that

$$\begin{aligned} \frac{d}{dt} |\psi_{N_0} - \psi_{\text{RWA}}|^2 &= 2 \text{Re} i \langle \psi_{N_0} - \psi_{\text{RWA}} | \text{cl}_{N_0}(N_0, 1) \psi_{N_0} - H_{\text{RWA}} \psi_{\text{RWA}} \rangle \\ &= 2 \text{Re} (i \langle \psi_{N_0} - \psi_{\text{RWA}} | H_{\text{RWA}} (\psi_{N_0} - \psi_{\text{RWA}}) \rangle \\ &\quad + i \langle \psi_{N_0} - \psi_{\text{RWA}} | (\text{cl}_{N_0}(N_0, 1) - H_{\text{RWA}}) \psi_{N_0} \rangle) \\ &\leq |\psi_{N_0} - \psi_{\text{RWA}}| O(\varepsilon_1^3 \varepsilon_2^2 + \varepsilon_1^{N_0}). \end{aligned}$$

Thus,

$$2 \frac{d}{dt} |\psi_{N_0} - \psi_{\text{RWA}}| \leq O(\varepsilon_1^3 \varepsilon_2^2 + \varepsilon_1^{N_0}),$$

and we conclude by integrating over  $[0, \frac{1}{\varepsilon_1 \varepsilon_2}]$ .  $\square$

This concludes the proof of Theorem V.16.

### V.3.4 Two scales adiabatic approximation

The goal of this part is to prove the following lemma:

**Lemma V.30.** *There exists  $\delta > 0$  such that the solution  $\psi_{\text{RWA}}$  of (V.24) satisfies  $|\psi_{\text{RWA}}(\frac{1}{\varepsilon_1\varepsilon_2}) - (e^{i\theta}, 0)| \leq M\varepsilon_2/\varepsilon_1$  for some  $\theta \in \mathbb{R}$  (possibly depending on  $\varepsilon_1, \varepsilon_2, \alpha$ ) for  $(\varepsilon_1, \varepsilon_2) \in (0, \delta)^2$ .*

With a slight abuse of notation, let us say in this section that a  $(\varepsilon_1, \varepsilon_2)$ -parametric function  $f$  is a  $O(g(\varepsilon_1, \varepsilon_2))$  (respectively, a  $\Omega(g(\varepsilon_1, \varepsilon_2))$ ) if there exist  $M, \delta > 0$  such that

$$\forall \varepsilon_1, \varepsilon_2 \in (0, \delta)^2, \forall s \in [0, 1], |f_{\varepsilon_1, \varepsilon_2}(s)| \leq Mg(\varepsilon_1, \varepsilon_2) \quad (\text{respectively, } |f_{\varepsilon_1, \varepsilon_2}(s)| \geq Mg(\varepsilon_1, \varepsilon_2)) \quad (\text{V.26})$$

Recall that

$$H_{\text{RWA}}(t) = \varepsilon_1 h_1(\varepsilon_1 \varepsilon_2 t) A(E_1(t)) + \varepsilon_1^2 h_2(\varepsilon_1 \varepsilon_2 t) B(E_2(t)) + \varepsilon_1^2 h_3(\varepsilon_1 \varepsilon_2 t) \sigma_z,$$

with  $h_1, h_2,$  and  $h_3$  given by Theorem V.16. We introduce the unitary change of variables  $\psi_{\text{slow}}(t) = U(t)\psi_{\text{RWA}}(t)$  with

$$U(t) = \begin{pmatrix} e^{i(\alpha t - \frac{\Delta(\varepsilon_1 \varepsilon_2 t)}{2\varepsilon_1 \varepsilon_2})} & 0 \\ 0 & e^{-i(\alpha t - \frac{\Delta(\varepsilon_1 \varepsilon_2 t)}{2\varepsilon_1 \varepsilon_2})} \end{pmatrix}.$$

The notation  $\psi_{\text{slow}}$  is motivated by the fact that the Hamiltonian corresponding to its evolution is slow in the sense that it only depends on the *slow variable*  $s = \varepsilon_1 \varepsilon_2 t$ , also known as *macroscopic* or *reduced time*. More precisely,  $i \frac{d}{dt} \psi_{\text{slow}}(t) = H_{\text{slow}}(\varepsilon_1 \varepsilon_2 t) \psi_{\text{slow}}(t)$ , where

$$H_{\text{slow}}(s) = \varepsilon_1 h_1(s) \sigma_x + \varepsilon_1^2 h_2(s) \sigma_y + \left( \alpha - \frac{\Delta'(s)}{2} + \varepsilon_1^2 h_3(s) \right) \sigma_z. \quad (\text{V.27})$$

We cannot directly apply a ‘standard adiabatic theorem’ to describe the evolution of  $\psi_{\text{slow}}$  because the adiabatic path depends on  $(\varepsilon_1, \varepsilon_2)$ .

The eigenvalues of  $H_{\text{slow}}(s)$  are

$$\pm \omega_{\varepsilon_1, \varepsilon_2}(s) = \pm \sqrt{(\varepsilon_1 h_1(s))^2 + (\varepsilon_1^2 h_2(s))^2 + (\alpha - \Delta'(s)/2 + \varepsilon_1^2 h_3(s))^2}, \quad s \in [0, 1].$$

Using a Taylor series development, we have  $\omega_{\varepsilon_1, \varepsilon_2} = \Omega(\varepsilon_1)$ . Thus, for  $(\varepsilon_1, \varepsilon_2)$  small enough,  $\omega_{\varepsilon_1, \varepsilon_2}$  does not vanish. As a consequence, we can introduce the spectral projector  $P_{\varepsilon_1, \varepsilon_2}(s)$  of  $H_{\text{slow}}(s)$  on the negative eigenvalue. Consider  $\gamma_{\varepsilon_1, \varepsilon_2} : [0, 1] \rightarrow S^2$  such that  $H_{\text{slow}}(s) = \omega_{\varepsilon_1, \varepsilon_2}(s) \gamma_{\varepsilon_1, \varepsilon_2}(s) \cdot \vec{\sigma}$  where  $\vec{\sigma} = a_1 \sigma_x + a_2 \sigma_y + a_3 \sigma_z$ . We want to approximate  $P_{\varepsilon_1, \varepsilon_2}$  and its derivatives by the spectral projector on the negative eigenvalue for the simplified Hamiltonian  $\tilde{H}_{\text{slow}} = \varepsilon_1 u \sigma_x + (\alpha - \Delta'/2) \sigma_z$  and its derivatives.

**Lemma V.31.** *Let  $-\tilde{\omega}_{\varepsilon_1}(s)$  be the negative eigenvalue of the Hamiltonian  $\tilde{H}_{\text{slow}}(s) = \varepsilon_1 u(s) \sigma_x + (\alpha - \Delta'(s)/2) \sigma_z$  and  $\tilde{P}_{\varepsilon_1}(s)$  be the spectral projector on  $-\tilde{\omega}_{\varepsilon_1}(s)$ ,  $s \in [0, 1]$ . Then*

$$\begin{aligned} & - |P_{\varepsilon_1, \varepsilon_2} - \tilde{P}_{\varepsilon_1}| = O(\varepsilon_1), \\ & - \left| \frac{d}{ds} P_{\varepsilon_1, \varepsilon_2} - \frac{d}{ds} \tilde{P}_{\varepsilon_1} \right| = O(1), \\ & - \left| \frac{d^2}{ds^2} P_{\varepsilon_1, \varepsilon_2} - \frac{d^2}{ds^2} \tilde{P}_{\varepsilon_1} \right| = O(1/\varepsilon_1). \end{aligned}$$

*Proof.* First, remark that for every nonnegative integer  $\nu$

$$\frac{d^\nu}{ds^\nu} H_{\text{slow}}(s) = \frac{d^\nu}{ds^\nu} \tilde{H}_{\text{slow}}(s) + \frac{d^\nu}{ds^\nu} R(\varepsilon_1, \varepsilon_2, s), \quad (\text{V.28})$$

where  $\frac{d^\nu}{ds^\nu} R(\varepsilon_1, \varepsilon_2, s) = O(\varepsilon_1^2)$ .

For  $H \in \mathfrak{isu}_2 \setminus \{0\}$ , define the orthogonal projector  $P(H)$  as the projector on the negative eigenvalue of  $H$ . The map  $P$  is  $C^\infty$  and positively homogeneous of degree 0 on  $\mathfrak{isu}_2 \setminus \{0\}$ .

For every  $r > 0$ , let  $B_r$  be the Euclidean ball of center 0 and radius  $r$  in  $\mathcal{M}_2(\mathbb{C})$ . Denote by  $\mathcal{K}$  the compact set  $i\mathfrak{su}_2 \cap \partial B_1$ . The differential  $dP$  is positively homogeneous of degree  $-1$ , since

$$\forall H, h \in i\mathfrak{su}_2, dP_H(h) = dP_{\frac{h}{|H|}}\left(\frac{h}{|H|}\right).$$

As a consequence, for  $H \in \mathfrak{su}_2 \setminus B_r$ ,  $|dP_H| \leq \frac{\sup_{L \in \mathcal{K}} |dP_L|}{r}$ . Thus, there exists a universal constant  $C > 0$  such that  $P$  is  $\frac{C}{r}$ -Lipschitz continuous on  $i\mathfrak{su}_2 \setminus B_r$ .

Moreover, consider  $r(\varepsilon_1, \varepsilon_2) := \inf_{s \in [0, 1]} \omega_{\varepsilon_1, \varepsilon_2}(s)/2 = \Omega(\varepsilon_1)$ . As  $H_{\text{slow}} - \tilde{H}_{\text{slow}} = O(\varepsilon_1^2)$ , for  $\varepsilon_1, \varepsilon_2$  small enough we can assume that the segment  $[H_{\text{slow}}(s), \tilde{H}_{\text{slow}}(s)] \cap B_{r(\varepsilon_1, \varepsilon_2)}$  is the empty set for every  $s \in [0, 1]$ . Then, applying  $P$  to the equality (V.28) for  $\nu = 0$ , we obtain

$$|P_{\varepsilon_1, \varepsilon_2}(s) - \tilde{P}_{\varepsilon_1}(s)| \leq \frac{C}{r(\varepsilon_1, \varepsilon_2)} |R(\varepsilon_1, \varepsilon_2, s)| \leq M' \varepsilon_1, \quad \forall s \in [0, 1].$$

For the second point, we have  $\frac{d}{ds} P_{\varepsilon_1, \varepsilon_2}(s) = dP_{H_{\text{slow}}(s)}\left(\frac{d}{ds} H_{\text{slow}}(s)\right)$ . As  $dP$  is positively homogeneous of degree  $-1$ ,  $d^2 P$  is positively homogeneous of degree  $-2$ . Thus  $H \mapsto dP_H$  is  $\frac{C'}{r^2}$ -Lipschitz continuous on  $i\mathfrak{su}_2 \setminus B_r$  with  $C' = \sup_{L \in \mathcal{K}} |d^2 P_L|$ . Thus, for  $\varepsilon_1, \varepsilon_2$  small enough,

$$\left| dP_{H_{\text{slow}}}\left(\frac{d}{ds} H_{\text{slow}}\right) - dP_{\tilde{H}_{\text{slow}}}\left(\frac{d}{ds} H_{\text{slow}}\right) \right| \leq \frac{C'}{r(\varepsilon_1, \varepsilon_2)^2} |H_{\text{slow}} - \tilde{H}_{\text{slow}}|,$$

and

$$\left| dP_{\tilde{H}_{\text{slow}}}\left(\frac{d}{ds} H_{\text{slow}}\right) - dP_{\tilde{H}_{\text{slow}}}\left(\frac{d}{ds} \tilde{H}_{\text{slow}}\right) \right| \leq \frac{C}{r(\varepsilon_1, \varepsilon_2)} \left| \frac{d}{ds} H_{\text{slow}} - \frac{d}{ds} \tilde{H}_{\text{slow}} \right|.$$

Thus we get

$$dP_{H_{\text{slow}}}\left(\frac{d}{ds} H_{\text{slow}}\right) = dP_{\tilde{H}_{\text{slow}}}\left(\frac{d}{ds} \tilde{H}_{\text{slow}}\right) + O(1).$$

The third point is obtained by the same kind of argument.  $\square$

**Remark V.32.** The Hamiltonian  $\tilde{H}_{\text{slow}}(s) = \varepsilon_1 u(s) \sigma_x + (\alpha - \Delta'(s)/2) \sigma_z$  is given by the first order RWA. The fact that  $\varepsilon_1$  appears in front of the pulse is obviously of utter importance for the estimation of the RWA error but also means that the ‘adiabatic path’ is shrinking to the conical eigenvalue intersection. In fact, it is worse than just the shrinking of the spectral gap, as the derivative of the spectral projector is blowing up near the conical intersection (see Figures V.5 and V.6).

Define  $\tilde{\gamma}_{\varepsilon_1}(s)$ ,  $s \in [0, 1]$ , by the relation  $\tilde{H}_{\text{slow}}(s) = \tilde{\omega}_{\varepsilon_1}(s) \tilde{\gamma}_{\varepsilon_1}(s) \cdot \vec{\sigma}$  and denote by  $(\tilde{\theta}, \tilde{\phi})$  the spherical coordinates of  $\tilde{\gamma}_{\varepsilon_1}$ . Hence  $X = \sin(\tilde{\theta}/2) e_1 - e^{i\tilde{\phi}} \cos(\tilde{\theta}/2) e_2$  is an eigenvector of  $\tilde{H}_{\text{slow}}(s)$  associated with the negative eigenvalue and

$$\tilde{P}_{\varepsilon_1} = \begin{pmatrix} \sin^2(\tilde{\theta}/2) & -e^{-i\tilde{\phi}} \sin(\tilde{\theta}/2) \cos(\tilde{\theta}/2) \\ -e^{i\tilde{\phi}} \sin(\tilde{\theta}/2) \cos(\tilde{\theta}/2) & \cos^2(\tilde{\theta}/2) \end{pmatrix}. \quad (\text{V.29})$$

**Lemma V.33.** Under the assumptions of Theorem V.3, we have:

1.  $|P_{\varepsilon_1, \varepsilon_2}(0) - P_{e_1}| = O(\varepsilon_1^2 \varepsilon_2)$  and  $|P_{\varepsilon_1, \varepsilon_2}(1) - P_{e_2}| = O(\varepsilon_1^2 \varepsilon_2)$ , where  $P_{e_i}$  is the orthogonal projector on  $\mathbb{C}e_i$ ;
2.  $\int_0^1 \left| \left( \frac{d}{ds} P_{\varepsilon_1, \varepsilon_2} \right)(s) \right|^2 ds = O(1/\varepsilon_1)$ ;
3.  $\int_0^1 \left| \left( \frac{d^2}{ds^2} P_{\varepsilon_1, \varepsilon_2} \right)(s) \right| ds = O(1/\varepsilon_1)$ ;

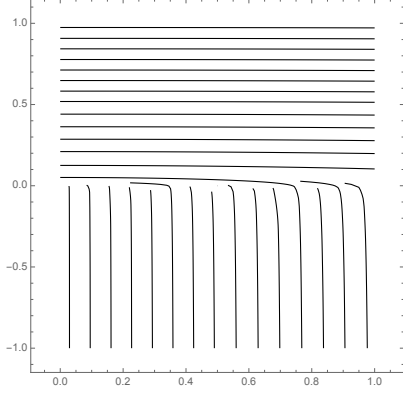


Figure V.5 – Eigendirection corresponding to the negative eigenvalue of  $\tilde{H}_{\text{slow}}$  as a function of  $(u, \Delta') \in \mathbb{R}^2$ , for  $\varepsilon_1 = 0.01$  and  $\alpha = 0$ .

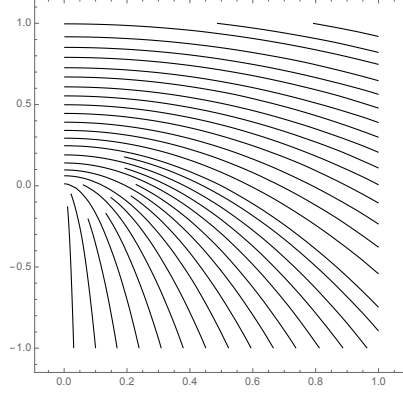


Figure V.6 – Eigendirection corresponding to the negative eigenvalue of  $\tilde{H}_{\text{slow}}$  as a function of  $(u, \Delta') \in \mathbb{R}^2$ , for  $\varepsilon_1 = 1$  and  $\alpha = 0$ .

$$4. \int_0^1 \left| \frac{1}{\omega_{\varepsilon_1, \varepsilon_2}(s)^2} \frac{d}{ds} P_{\varepsilon_1, \varepsilon_2}(s) \right| \left| \frac{d}{ds} H_{\text{slow}}(s) \right| ds = O(1/\varepsilon_1^2).$$

*Proof.* Point 1 is a simple consequence of points 1, 2 and 3 in Theorem V.16.

Concerning the other three points, thanks to Lemma V.31 we are left to prove the corresponding estimates for  $\tilde{P}_{\varepsilon_1, \varepsilon_2}$  and  $\tilde{H}_{\text{slow}}$ . We recall that  $\tilde{\theta}(s) = \arccos\left((\alpha - \Delta'(s)/2)/\tilde{\omega}_{\varepsilon_1}(s)\right)$ . We can bound the transverse velocity of  $\tilde{\gamma}_{\varepsilon_1}(s)$  by its total velocity

$$\left| \tilde{\omega}_{\varepsilon_1}(s) \frac{d}{ds} \tilde{\theta}(s) \right| \leq \sqrt{(\varepsilon_1 \frac{d}{ds} u(s))^2 + (\Delta''(s)/2)^2} = O(1),$$

thus  $\frac{d}{ds} \tilde{\theta} = O(1/\varepsilon_1)$ .

Using formula (V.29), it is clear that  $|\frac{d}{ds} \tilde{P}_{\varepsilon_1}(s)| \lesssim |\frac{d}{ds} \tilde{\theta}|$  and  $|\frac{d^2}{ds^2} \tilde{P}_{\varepsilon_1}(s)| \lesssim |\frac{d}{ds} \tilde{\theta}|^2 + |\frac{d^2}{ds^2} \tilde{\theta}|$ , where  $\lesssim$  stands for inequality up to an universal multiplicative constant. As  $\Delta'' \geq 0$ ,  $\tilde{\theta}$  is increasing and

$$\int_0^1 \left| \frac{d}{ds} \tilde{\theta}(s) \right|^2 ds \leq \sup_{s \in [0,1]} \left| \frac{d}{ds} \tilde{\theta}(s) \right| \int_0^1 \left| \frac{d}{ds} \tilde{\theta}(s) \right| ds \leq \pi \sup_{s \in [0,1]} \left| \frac{d}{ds} \tilde{\theta}(s) \right| = O(1/\varepsilon_1).$$

Moreover, bounding the transverse acceleration of  $\tilde{\gamma}_{\varepsilon_1}(s)$  by its total acceleration, we have

$$\left| 2 \frac{d}{ds} \tilde{\omega}_{\varepsilon_1}(s) \frac{d}{ds} \tilde{\theta}(s) - \tilde{\omega}_{\varepsilon_1}(s) \frac{d^2}{ds^2} \tilde{\theta}(s) \right| \leq \sqrt{(\varepsilon_1 \frac{d^2}{ds^2} u(s))^2 + (\Delta'''(s)/2)^2} = O(1).$$

As  $\frac{d}{ds} \tilde{\omega}_{\varepsilon_1}(s) \leq \sqrt{(\varepsilon_1 \frac{d}{ds} u(s))^2 + (\Delta''(s)/2)^2} = O(1)$ , we have

$$\int_0^1 \left| \frac{d}{ds} \tilde{\omega}_{\varepsilon_1}(s) \frac{d}{ds} \tilde{\theta}(s) \right| = O(1),$$



leading to

$$\int_0^1 \left| \frac{d^2}{ds^2} \tilde{\omega}_{\varepsilon_1}(s) \right| ds = O(1/\varepsilon_1).$$

Concerning point 4, notice the integral  $\int_0^1 \frac{1}{\omega_{\varepsilon_1, \varepsilon_2}(s)^2} \frac{d}{ds} P_{\varepsilon_1, \varepsilon_2}(s) \left\| \frac{d}{ds} H_{\text{slow}}(s) \right\| ds$  can be upper bounded, up to a multiplicative constant, by

$$\int_0^1 \left( \frac{1}{\tilde{\omega}_{\varepsilon_1}(s)^2} \left| \frac{d}{ds} \tilde{\omega}_{\varepsilon_1}(s) \right| \left| \frac{d}{ds} \tilde{\theta}(s) \right| + \frac{1}{\tilde{\omega}_{\varepsilon_1}(s)} \left( \frac{d}{ds} \tilde{\theta}(s) \right)^2 \right) ds,$$

which is of order  $1/\varepsilon_1^2$ .  $\square$

To conclude the proof of Lemma V.30, we deduce from [Teu03, Corollary 2.3] the adiabatic estimate

$$\begin{aligned} \left| \psi_{\varepsilon_1, \varepsilon_2}^\alpha \left( \frac{1}{\varepsilon_1 \varepsilon_2} \right) - (e^{i\theta}, 0) \right| \leq \varepsilon_1 \varepsilon_2 \left[ \frac{\left| \frac{d}{ds} P_{\varepsilon_1, \varepsilon_2}(1) \right|}{\omega_{\varepsilon_1, \varepsilon_2}(1)} + \frac{\left| \frac{d}{ds} P_{\varepsilon_1, \varepsilon_2}(0) \right|}{\omega_{\varepsilon_1, \varepsilon_2}(0)} \right. \\ \left. + \int_0^1 \left( \frac{2 \left| \frac{d}{ds} P_{\varepsilon_1, \varepsilon_2}(s) \right|^2}{\omega_{\varepsilon_1, \varepsilon_2}(s)} + \frac{\left| \frac{d^2}{ds^2} P_{\varepsilon_1, \varepsilon_2}(s) \right|}{\omega_{\varepsilon_1, \varepsilon_2}(s)} + \frac{\left| \frac{d}{ds} P_{\varepsilon_1, \varepsilon_2}(s) \right| \left\| \frac{d}{ds} H_{\text{slow}}(s) \right\|}{2\omega_{\varepsilon_1, \varepsilon_2}(s)^2} \right) ds \right], \end{aligned}$$

for some  $\theta \in \mathbb{R}$ . Finally, Lemma V.30 together with Theorem V.16 conclude the proof of Theorem V.3 for a given  $\alpha$  and  $\delta$ . To get uniformity on the range of  $\alpha$ , notice that the algorithm does not depend on  $\alpha$  elsewhere than in the expression of  $E_1$  and  $E_2$  (see (V.16)). For the adiabatic part, if we restrict  $\alpha$  to a compact interval  $I \subset (v_0, v_1)$ , the estimates of Lemma V.33 can be taken uniform with respect to  $\alpha$ . The uniformity with respect to  $\delta$  is straightforward.

**Remark V.34.** *Now that we have detailed the whole proof, we want to stress some of its key points.*

1. *The changes of variables applied iteratively in order to eliminate the oscillating terms of the Hamiltonian induce a very small error (of order  $\varepsilon_2 \varepsilon_1^2$ ) on the initial and the final state (Lemma V.29), whereas the error is of order  $\varepsilon_1$  if one look at the entire trajectory.*
2. *The frequencies which appear during the algorithm are of very special type ( $pf_1 - pf_2$  and  $(p+1)f_1 - pf_2$  for  $p$  integer) allowing us to perform as many changes of variables as we need and to give a simple condition implying that all such frequencies are nonzero.*
3. *Each change of variables yields a more complicated Hamiltonian. Fortunately, when we study the adiabatic dynamics of such an Hamiltonian, we can neglect all the terms except for those appearing in the first order RWA.*
4. *The first order RWA induces a population transfer in the limit  $\varepsilon_2 \ll \varepsilon_1$ .*

## V.4 Numerical simulations

We present in this section some numerical simulations illustrating the results stated in Theorem V.3. In all simulations we use the chirp scheme presented in Remark V.5 with  $E = 1$ ,  $v_0 = -0.5$ , and  $v_1 = 0.5$ .

Figure V.7 shows the behavior of the distance from the target state as a function of  $\varepsilon_1, \varepsilon_2$  represented in log scale. The AA error appears clearly, reflecting the fact that one needs  $\varepsilon_2 \ll \varepsilon_1$  in order to have a fidelity close to 1. The figure also shows that the strategy has better performances than those anticipated theoretically in Theorem V.3.

Figure V.8 shows the fidelity as a function of  $\alpha$ , while  $\varepsilon_1$  and  $\varepsilon_2$  (and hence  $\mathcal{T}$ ) are fixed.

Figure V.9 shows the fidelity as a function of the reduced time for three values of  $\alpha$ , while  $\varepsilon_1$  and  $\varepsilon_2$  (and hence  $\mathcal{T}$ ) are fixed. We clearly see that the RWA produces large oscillations (of magnitude of order  $\varepsilon_1$ ), which become much smaller at the endpoints, as described in Remark V.34, point 3.

Finally, Figure V.10 illustrates the conflict between the AA and RWA. At  $\mathcal{T} = 0.05$  fixed, for smaller  $\varepsilon_1$  we observe that the RWA is more accurate as the thick line (1st order RWA) is closer to the highly oscillating one (the trajectory  $\psi_{\varepsilon_1, \varepsilon_2}^0$ ). Nevertheless as  $\varepsilon_1$  decreases, the ratio  $\varepsilon_2/\varepsilon_1$  increases and the AA becomes less accurate.

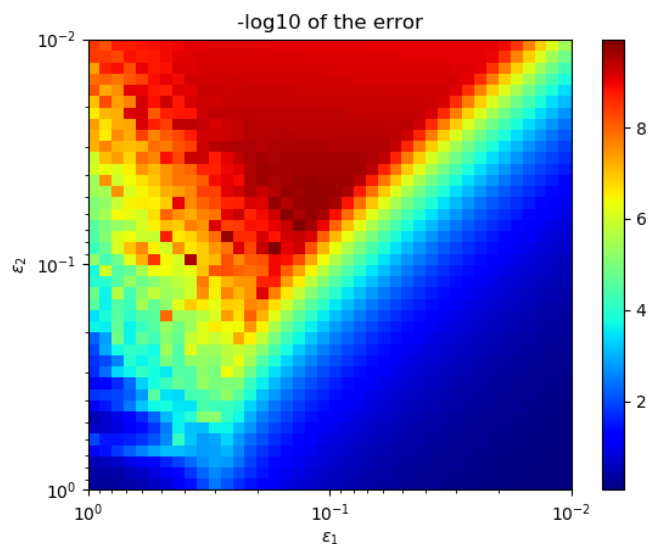


Figure V.7 – Log of the distance from  $\psi_{\varepsilon_1, \varepsilon_2}^0(\frac{1}{\varepsilon_1 \varepsilon_2})$  to the orbit of  $(1, 0)$ .

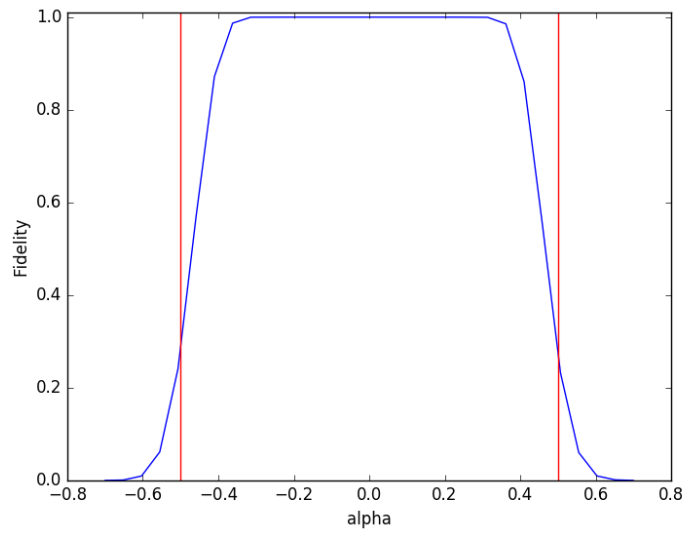


Figure V.8 – Population transfer as a function of  $\alpha$  for  $E = 1$ ,  $\varepsilon_1 = 0.5$  and  $\varepsilon_2 = 0.1$ .

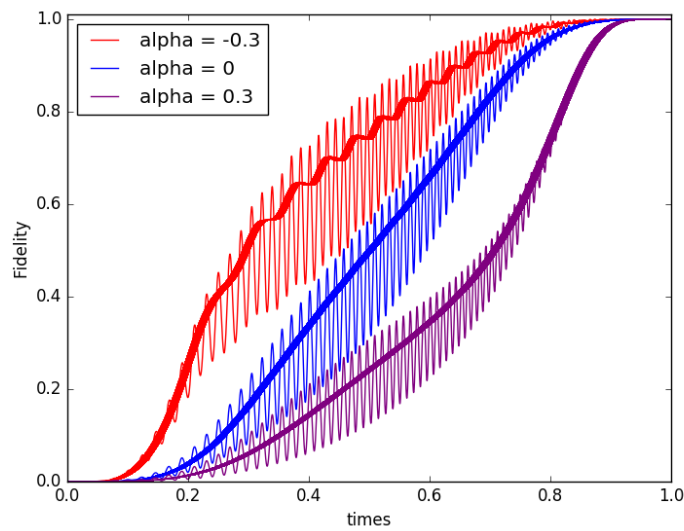


Figure V.9 –  $\varepsilon_1 = 0.5$ ,  $\varepsilon_2 = 0.1$  and  $\alpha = 0$ . In thick line are the trajectories corresponding to the equivalent 1st order RWA system.

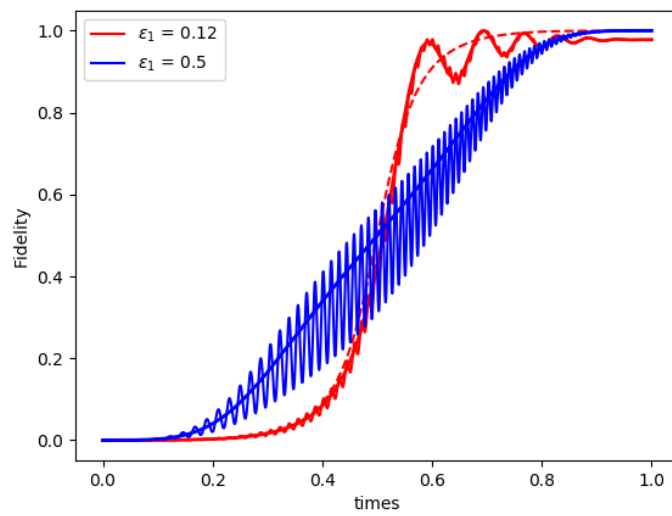


Figure V.10 –  $\varepsilon_1\varepsilon_2 = 0.05$ ,  $\alpha = 0$ . In thick line are the trajectories corresponding to the equivalent 1st order RWA system and in dotted line the theoretical AA trajectories.



## Chapter VI

# Chattering Phenomenon in Quantum Optimal Control

This chapter is taken from the following submitted letter (also referred as [Rob+22a]):

R. Robin, U. Boscain, M. Sigalotti, and D. Sugny. *Chattering phenomenon in quantum optimal control*. 2022. arXiv: 2206.13868 [quant-ph]

We present a quantum optimal control problem which exhibits a chattering phenomenon. This is the first instance of such a process in quantum control. Using the Pontryagin Maximum Principle and a general procedure due to V. F. Borisov and M. I. Zelikin, we characterize the local optimal synthesis, which is then globalized by a suitable numerical algorithm. We illustrate the importance of detecting chattering phenomena because of their impact on the efficiency of numerical optimization procedures.

## VI.1 Contribution

### VI.1.1 Introduction

Consider the experiment in which a ball bounces up and down on the ground. We assume that the impact with the ground is inelastic and that the ball is only subjected to the gravity. In the ideal case in which the ball changes its speed instantaneously at each bounce, an infinite number of bounces is performed in the finite time of the process. Chattering thus refers to an observable (here the speed) having very fast oscillations, which lead in the mathematical limit to an infinite number of jumps in a finite-time interval [ZB94]. This type of process can also be found in quantum physics. Examples are the quantum Zeno effect and dynamical decoupling in which a repeated observation of the system and a periodic sequence of instantaneous pulses prevent, respectively, its time evolution [MS77] or its coupling with the environment [VKL99]. The possibility of chattering was also established in Optimal Control Theory (OCT). OCT was founded in the sixties by the pioneering study of Pontryagin and his co-workers [Pon+74], who introduced the Pontryagin Maximum Principle (PMP). OCT is a rigorous framework to design control protocols for driving a dynamical system from a given initial state into a desired target state, while minimizing energy or other resources. Chattering was found in this field by A.T. Fuller in a planar system [Ful60; Bor00; SL12]. In OCT, it consists of an optimal con-

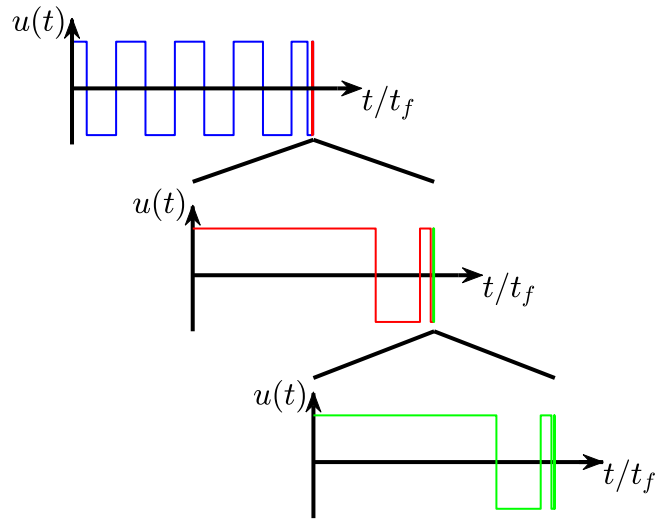


Figure VI.1 – Schematic description of the time scale invariance of the optimal control in the chattering process. The optimal solution of the quantum control problem described in this chapter is plotted in the top panel. Near the final time  $t_f$ , the control switches infinitely many times with an asymptotic invariant structure by time dilation as represented on the two lower panels (successive zooms around  $t = t_f$  given by the different plot colors).

control that switches infinitely many times over a finite time period [ZB94]. This observation runs counter to common experience for which control is viewed as a continuous or piecewise continuous function, while for chattering, the control is no longer piecewise continuous but lies in a larger class of functions [BSS21]. In Fuller’s example, the number of switchings accumulates with a geometric progression at the final control time. The control law has a time scale invariance near the final time as schematically represented in Fig. VI.1. At first Fuller’s example was considered a curiosity in optimal control but it gradually became clear that this type of phenomenon was very widespread in the control of dynamical systems, as rigorously shown few years later by I. Kupka [Kup90]. While optimal trajectories exhibiting the chattering phenomenon have been found in practically relevant examples from medicine [SL12] to classical [ZB94] and space dynamics [ZTC16], this phenomenon has, to the best of our knowledge, not yet been studied in quantum control [Gla+15; Koc+22]. In particular, the existence, the role and the ubiquity of this process in quantum systems remain an open question. We point out that such control schemes are interesting from a fundamental point of view even if they turn out to not be feasible in experiments. They may also have a rather severe impact on the numerical search of optimal solutions [BHH75]. For example, the chattering phenomenon has recently been shown to lead to numerical instabilities in optimization procedures, preventing the design of efficient controls [ZTC16]. It is therefore important to understand why chattering is occurring and how it can be avoided [Cap+18]. We propose in this chapter to begin the description of this phenomenon in quantum control by studying a simple but fundamental quantum system. We introduce on this key example a systematic procedure to design the optimal control protocol. The impact on the efficiency of optimization procedures is also described.

### VI.1.2 Model

Let us consider the control of a three-level quantum system described by a pure state belonging to a Hilbert space spanned by the states  $|1\rangle$ ,  $|2\rangle$ , and  $|3\rangle$ . As in a standard STIRAP process [Vit+17], in a suitable rotating frame and in the rotating wave approximation, the dynamics of the system are controlled by two pulses of Rabi frequencies  $\Delta$  and  $u$  that couple, respectively, states  $|1\rangle$  and  $|2\rangle$  and states  $|2\rangle$  and  $|3\rangle$ . The resulting dynamics are

$$\dot{\mathbf{X}} = (\Delta\Omega_3 + u(t)\Omega_1)\mathbf{X}, \quad (\text{VI.1})$$

where  $\mathbf{X}$  is a vector of real coordinates  $(x_1, x_2, x_3)$  with the condition  $x_1^2 + x_2^2 + x_3^2 = 1$  (see Section VI.2.1). The two skew-symmetric matrices  $\Omega_1$  and  $\Omega_3$  are given by

$$\Omega_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \Omega_3 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

generating respectively the rotations along the  $x_1$ - and  $x_3$ -directions. We assume that  $\Delta$  is a constant and that the only control parameter is  $u(t)$ . In the case for which the control cannot exceed a certain physical bound, we have  $|u(t)| \leq u_0$  for some  $u_0 > 0$ . A time rescaling results in the multiplication of the two parameters  $\Delta$  and  $u_0$  of the problem by a positive scalar, which leads to the normalization  $u_0 = 1$ .

Starting from any state  $\mathbf{X}_0$  on the unit sphere, the goal of the control is to steer the system to the state  $(0, 0, 1)$ , still denoted by  $|3\rangle$ , while minimizing the population transfer to the state  $|1\rangle$ . This control protocol is interesting in practice if, e.g., the state  $|1\rangle$  is the only state of the system subject to an unwanted relaxation process. Notice that this latter is not modeled by Eq. (VI.1). The control protocol can be formalized as an optimal control problem by introducing the cost functional  $\mathcal{C} = \int_0^{t_f} x_1^2(t) dt$  to be minimized, where  $t_f$  is the control time, which is not fixed. We stress that a mathematically equivalent problem would have been obtained by considering the task of reaching any fixed state from  $|3\rangle$  minimizing the population transfer to the state  $|1\rangle$ . The existence of a control reaching the desired target in finite time and achieving the minimum of the cost  $\mathcal{C}$  is not obvious. It is a consequence of the general theory developed in the book [ZB94], as explained in Section VI.2.4. We give below an argument showing that, once its existence has been established, such a control has a chattering behaviour.

### VI.1.3 Description of the optimal control

The main tool to prove the existence of chattering is the PMP, which is a first-order necessary condition for optimality (see [BSS21] for details). The PMP can be stated by introducing the Pontryagin Hamiltonian

$$H_P = \mathbf{P} \cdot (\Delta\Omega_3 + u\Omega_1)\mathbf{X} + p^0 x_1^2,$$

where  $\mathbf{X}$  is a point on the unit sphere,  $\mathbf{P}$  the adjoint state of coordinates  $(p_1, p_2, p_3)$ , and  $p^0$  a constant equal either to 0 or to  $-\frac{1}{2}$ . If  $\mathbf{X}(t)$  is an optimal trajectory with corresponding control  $u(t)$ , then  $\mathbf{X}(t)$  is *extremal*, namely, there exists  $\mathbf{P}(t)$  such that  $(\mathbf{P}(t), p^0) \neq (0, 0)$ ,  $\dot{\mathbf{X}} = \frac{\partial H_P}{\partial \mathbf{P}}$ ,  $\dot{\mathbf{P}} = -\frac{\partial H_P}{\partial \mathbf{X}}$ , and  $H_P(\mathbf{X}(t), \mathbf{P}(t), u(t)) = \max_{v \in [-1, 1]} H_P(\mathbf{X}(t), \mathbf{P}(t), v) = 0$ . If  $p^0 = 0$  (resp.  $p^0 = -\frac{1}{2}$ ) the extremal is called *abnormal* (resp. *normal*).

The maximization condition of the PMP can be solved by introducing the switching function  $\Phi(t) = \mathbf{P}(t) \cdot \Omega_1 \mathbf{X}(t)$ . If  $\Phi(t)$  is different from zero then  $H_P$  is maximum when the control, called



*bang*, is a constant control of maximum amplitude of the form  $u(t) = \text{sign}[\Phi(t)]$ . When  $\Phi(t)$  vanishes at an isolated point and changes sign, the control switches from  $\pm 1$  to  $\mp 1$  leading to bang-bang protocols. These extremals are said to be *regular*, otherwise they are called *singular*, as for instance when  $\Phi$  is zero on a given time interval [BSS21].

Singular extremals can be characterized as follows. A simple computation shows that  $\dot{\Phi} = \Delta \mathbf{P} \cdot \Omega_2 \mathbf{X}$ , where  $\Omega_2$  is the generator of the rotations along the  $x_2$ -axis. If both  $\Phi$  and  $\dot{\Phi}$  vanish at time  $t$  then either  $\mathbf{P}(t) = 0$  or  $\Omega_1 \mathbf{X}(t)$  and  $\Omega_2 \mathbf{X}(t)$  are linearly dependent, that is  $\mathbf{X}(t)$  lies on the equator  $x_3 = 0$ . If  $\mathbf{P}$  is equal to zero on a time interval then  $p^0 = -\frac{1}{2}$  and, moreover,  $0 = \dot{\mathbf{P}} = \frac{1}{2} \frac{\partial x_1^2}{\partial \mathbf{X}}$ , leading to  $x_1 = 0$  on the same interval. Since  $\dot{x}_1 = -\Delta x_2$ , it follows that  $\mathbf{X} = |3\rangle$ .

We now show that it is enough to consider normal extremals, i.e., that any optimal trajectory corresponding to an abnormal extremal also corresponds to a normal one. Indeed, if an optimal trajectory reaches  $|3\rangle$  in a finite time  $t_f$  then all its extensions that stay on  $|3\rangle$  for larger times are also optimal. As a consequence, the optimal trajectory corresponds to a vanishing control for all times larger than  $t_f$ , hence it is singular and we have  $\mathbf{P}(T) = 0$ . Such an extremal cannot be abnormal (otherwise  $(\mathbf{P}(T), p^0)$  would be zero). We deduce that for any optimal solution reaching  $|3\rangle$ , it can be assumed without loss of generality that  $T = t_f$ , the extremal is normal, and  $\mathbf{P}(t_f) = 0$ .

The last step consists in showing that an optimal trajectory cannot be bang-bang in  $[0, t_f]$ . By contradiction, assume that  $u$  is constantly equal to  $+1$  or to  $-1$  in an interval  $[t_f - \varepsilon, t_f]$  for some  $\varepsilon > 0$ . Then  $\Phi$  is smooth on  $[t_f - \varepsilon, t_f]$  and an explicit computation based on the dynamics of  $\mathbf{P}$

$$\dot{p}_1 = -\Delta p_2 + x_1, \quad \dot{p}_2 = \Delta p_1 - u p_3, \quad \dot{p}_3 = u p_2 \quad (\text{VI.2})$$

shows that  $\Phi(t_f) = \dot{\Phi}(t_f) = \Phi^{(2)}(t_f) = \Phi^{(3)}(t_f) = 0$  and  $\Phi^{(4)}(T) = -u\Delta^2$  (see Section VI.2.3). Then  $\Phi(t)$  has opposite sign with respect to  $u$  in a small left neighborhood of  $t_f$ , contradicting the fact that  $u(t) = \text{sign}[\Phi(t)]$ . Hence the only option to reach  $|3\rangle$  is a chattering process in which the control switches infinitely many times in a finite time-interval.

Having determined the chattering nature of the optimal control, the next goal is to find the position and the times of the switching points. This set of points is called the switching curve. This is not an easy task and an exact analytic expression cannot be derived. Notice however that close to the state  $|3\rangle$  the system can be described by the two coordinates  $x_1$  and  $x_2$  with the dynamics

$$\begin{cases} \dot{x}_1 = -\Delta x_2, \\ \dot{x}_2 = \Delta x_1 - u\sqrt{1 - x_1^2 - x_2^2}. \end{cases} \quad (\text{VI.3})$$

Taking only the dominant terms, the system can be locally approximated as

$$\begin{cases} \dot{x}_1 = -\Delta x_2, \\ \dot{x}_2 = -u, \end{cases} \quad (\text{VI.4})$$

which is not the standard linear approximation. In the control literature, system (VI.4) is referred to as the *nilpotent approximation* of (VI.3). The dynamics of (VI.4) together with the cost  $\int_0^{t_f} x_1^2(t) dt$  ( $t_f$  free) and the origin  $x_1 = x_2 = 0$  as target state, yield the classical Fuller problem, up to the change of coordinates  $x_1 \rightarrow x_1/\Delta$ ,  $x_2 \rightarrow x_2$ . The chattering trajectories of the Fuller model can be described exactly [SL12] and the main results about this example are recalled in Section VI.2.2. We then deduce the following properties for the linear system (VI.4) under study. The optimal solution is bang-bang with an infinite number of switchings near the

origin. The switching curve is defined by

$$x_1 = \text{sign}[x_2]\xi\Delta x_2^2, \quad (\text{VI.5})$$

where  $\xi = \sqrt{\frac{\sqrt{33}-1}{24}} \simeq 0.44623$ . The switching times  $t_k$  are given by a geometric progression of the form  $\frac{T-t_k}{T-t_{k-1}} = \frac{1}{\alpha}$  with  $\alpha = \sqrt{\frac{1+2\xi}{1-2\xi}} \simeq 4.1301$ .

In order to relate the optimal syntheses of (VI.3) and (VI.4), it is not enough to notice that they differ by terms of order higher than one. Instead, we can apply the results of [ZB94], which permit to conclude that a system has a Fuller-like optimal synthesis provided that it differs from the Fuller model by terms that are small while applying suitable non-isotropic dilations. On the basis of this study which is described in Section VI.2.4, we state the following result.

**Proposition VI.1.** *For every point  $\mathbf{X}_0$  sufficiently close to  $|3\rangle$ , an optimal solution of (VI.1) connecting  $\mathbf{X}_0$  to  $|3\rangle$  corresponds to a control having infinitely many discontinuities accumulating at the first time  $T$  at which  $|3\rangle$  is reached (and possibly staying at  $|3\rangle$  for larger times). The optimal synthesis is characterized by a switching curve  $\Gamma$  passing through  $|3\rangle$ , whose expression in coordinates  $(x_1, x_2)$  is of the form*

$$x_1 = \begin{cases} \lambda_1(x_2)x_2^2 & \text{if } x_2 > 0, \\ \lambda_0(x_2)x_2^2 & \text{if } x_2 < 0, \end{cases}$$

where  $\lambda_0$  and  $\lambda_1$  are  $C^1$  function satisfying  $\lambda_0(0) = -\lambda_1(0) = \Delta\xi$ . The optimal control is  $-1$  above  $\Gamma$  and  $+1$  below it.

### VI.1.4 Numerical simulations

The optimal synthesis for our control problem can be computed numerically starting from that of the Fuller model. When we are sufficiently close to  $|3\rangle$ , we approximate the switching curve of the quantum system by that of its approximation (VI.4). We consider a point of the latter curve of coordinates  $(\xi\Delta x_{20}^2, x_{20})$ , with  $x_{20} > 0$  (noticing that the same method could be used for  $x_{20} < 0$ ). The third component of  $\mathbf{X}^{(0)}$  is obtained from  $x_{30} = \sqrt{1 - x_{10}^2 - x_{20}^2}$  and the adjoint state from the condition  $\mathbf{P} \cdot \Omega_1 \mathbf{X} = 0$  together with  $H_P = 0$  (since  $t_f$  is free [BSS21]). The dynamics of the PMP allows to propagate backward in time  $\mathbf{X}$  and  $\mathbf{P}$ . Starting from an initial point with  $x_{20} > 0$ , one integrates the equations taking  $u = -1$ . When the corresponding switching function vanishes, the control switches from  $-1$  to  $+1$ . Then one goes on by integrating the equations with  $u = +1$  up to the next switching time and so on. An optimal control law can be obtained for each value of  $x_{20}$ . Even if the result applied above to characterize the optimal synthesis (see Theorem VI.2 in Section VI.2.4) can be used only for initial points  $\mathbf{X}_0$  close to  $|3\rangle$ , numerical simulations show that optimal trajectories have the same structure everywhere in the north hemisphere.

The trajectory starting from a given initial state (say  $|2\rangle$ ) can then be determined by a Newton algorithm to estimate the right parameter  $x_{20}$  from which the backward propagation arrives at the initial state. The parameter  $x_{20}$  is not unique because the forward optimal trajectory has infinitely many switching points near the target. In practice, the choice of  $x_{20}$  is dictated by the required precision on the final state. For  $\Delta = 10$  and a precision of  $10^{-3}$ , numerical simulations lead to  $x_{20} = 6.9 \times 10^{-4}$ . Figure VI.2 depicts the optimal trajectory on the unit sphere. The corresponding control  $u(t)$  and switching function  $\Phi(t)$  are displayed in Fig. VI.3.

The switching curves are reconstructed numerically by varying the small parameter  $x_{20}$  and collecting all the switching points of the corresponding trajectories obtained by integrating the

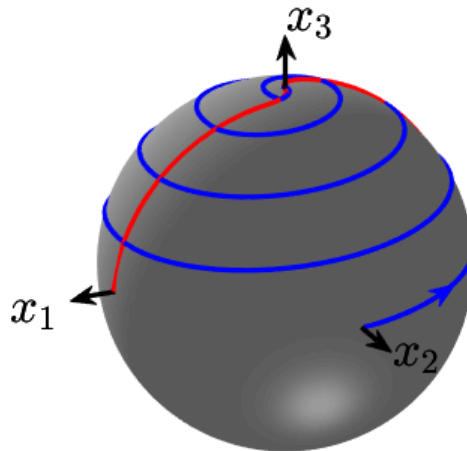


Figure VI.2 – Plot of the optimal trajectory from  $\mathbf{X}(0) = (0, 1, 0)$  to  $\mathbf{X}(t_f) = |3\rangle = (0, 0, 1)$  on the sphere  $x_1^2 + x_2^2 + x_3^2 = 1$ . The switching curves are plotted in red. The parameter  $\Delta$  is set to 10.

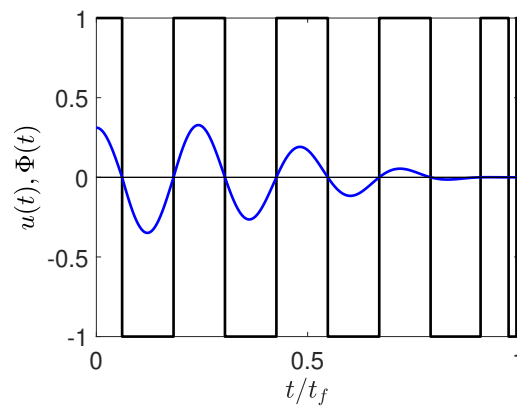


Figure VI.3 – Time evolution of the control  $u(t)$  (bold black line) and of the switching function  $\Phi(t)$  (blue line) for the optimal trajectory of Fig. VI.2. The control switches from  $\pm 1$  to  $\mp 1$  when the switching function changes sign.

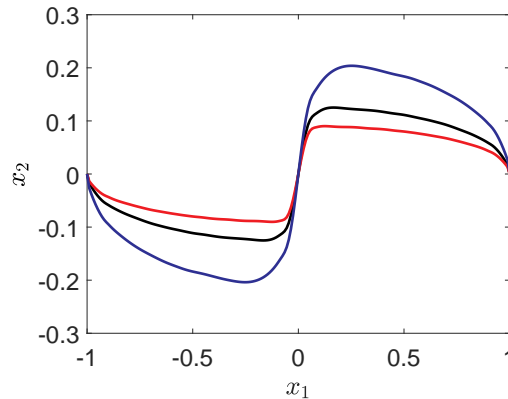


Figure VI.4 – Switching curves corresponding to three different values of  $\Delta$ . This parameter is set respectively to 6, 10 and 14 for

extremal flow backwards in time. The result is represented in Fig. VI.2 and VI.4<sup>1</sup>. A comparison between the results of the quantum system and those of its approximation (VI.4) can be found in the Section VI.2.2.

We observe in Fig. VI.4 that the points  $(\pm 1, 0, 0)$  belong to the switching curves of the quantum system, independently of the value of  $\Delta$ . The fact that the switching curve cannot exit the north hemisphere other than through the point  $(\pm 1, 0, 0)$  can be shown by applying a general technique described in [BC03, Section 4.3] (see also [BP04]). The reason is that the expression of the derivatives of the switching function  $\Phi$  (computed from the PMP) can be used to deduce that the control switchings from  $-1$  to  $+1$  occurs only in the regions  $\{x_2 > 0, x_3 > 0\}$  and  $\{x_2 < 0, x_3 < 0\}$ , while the control switchings from  $+1$  to  $-1$  are possible only in  $\{x_2 < 0, x_3 > 0\}$  and  $\{x_2 > 0, x_3 < 0\}$ . This result is explicitly derived in Section VI.2.3.

The preceding geometric analysis gives the optimal control protocol with a very high numerical precision. Due to their complexity, quantum control problems are usually solved by numerical optimization algorithms in which the control is assumed to be a piecewise constant function [Gla+15]. At this point, an intriguing question is to study to which extent the solutions derived from these algorithms can approximate the chattering phenomenon of the optimal strategy. The numerical simulations presented below use a direct optimal method with the software BOCOP [BGM11] with a fixed control time equal to the geometric one. In Fig. VI.5, we observe that the chattering process of the optimal solution can only be reproduced approximately by the numerical optimization. Additional results illustrating this aspect can be found in Section VI.2.5. As could be expected, the fineness of the time discretization corresponding here to  $t_f/N$ , where  $N$  is the number of time steps, is a key factor to improve the efficiency of the protocol and to reproduce the control shape. However, for  $N=400$ , we observe an erratic structure of the control. Without a precise understanding of the optimal control strategy, this switching accumulation could be misinterpreted as numerical instabilities or artifacts, while it is due to the very structure of the optimal control. Reasonable efficiencies of the control protocols are achieved for quite small values of  $N$ . Such sub-optimal strategies could be a possible option to bypass the problem

1. Note that it makes sense to integrate backwards the extremal flow since, according to the results in [ZB94], the map  $\Upsilon$  associating with a switching point the subsequent one is hyperbolic and has  $\Gamma$  as stable manifold towards [3]. As a result, the  $k$ -th iteration of the inverse of  $\Upsilon$  tends towards  $\Gamma$  as  $k$  grows.

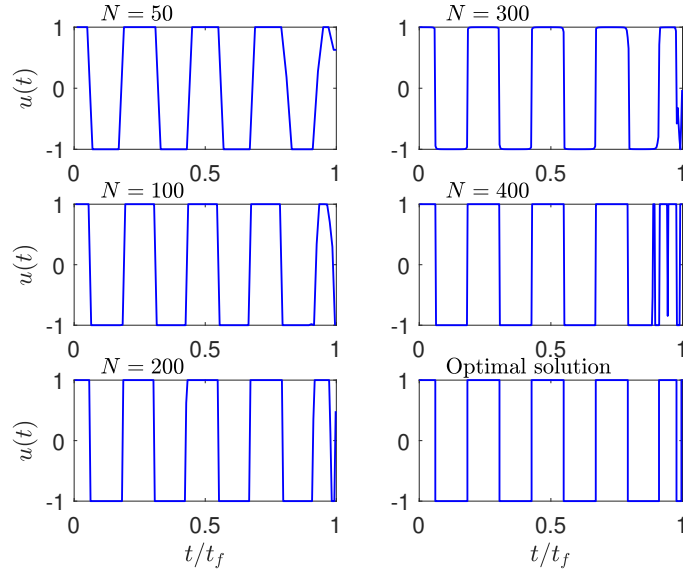


Figure VI.5 – Comparison between the optimal solution obtained by our numerical integration of the PMP (bottom right panel) and a numerical control designed by a direct approach for different time steps  $N$ . The control time is fixed to 2.59 in the numerical optimization.

due to chattering in the optimization procedure.

### VI.1.5 Conclusion

In summary, through the analysis of a fundamental quantum control problem, we have shown that the chattering phenomenon can appear in quantum optimal control. Such a process can play an unrecognized in numerical optimization algorithms. In this example, we point out that fast oscillations occur in numerical optimization procedures. In accordance with the existing results on the ubiquity of the Fuller phenomena in OCT [Kup90], we expect chattering to appear in many further examples, especially for higher-dimensional problems. Actually, one can give sufficient conditions on the relations between the commutators of the uncontrolled and the controlled Hamiltonians to guarantee the existence of chattering solutions of the PMP (see [Kup90] and [ZB94, Chapter 4] for technical details). Such conditions render the chattering phenomenon more and more frequent as the dimension grows, and our example shows that no obstacle to their appearance comes from the quantum structure of the control problem. It should be noticed, however, that one cannot infer from the cited general conditions in high dimensions the optimality of the chattering trajectories, unlike for the two-dimensional system considered in this chapter.

## VI.2 Technical results

This section gives technical details about the derivation of the physical system (Sec. VI.2.1), chattering phenomenon in the classical Fuller model (Sec. VI.2.2), and the switching function for the quantum optimal control problem studied in the chapter (Sec. VI.2.3). It also provides a theorem giving a sufficient condition for an optimal control problem to have a Fuller-like chat-

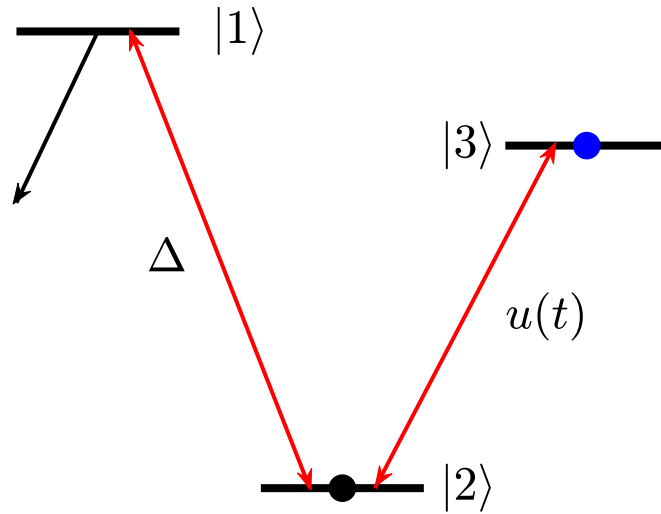


Figure VI.6 – Schematic representation of the quantum system with the coupling  $\Delta$  and  $u(t)$  between the states  $|1\rangle$  and  $|2\rangle$  and  $|2\rangle$  and  $|3\rangle$  (red arrows). The black and blue dots indicate respectively the initial and the target states. The black arrow represents the relaxation process.

tering phenomenon (Sec. VI.2.4) and some additional results about the numerical optimization procedure using the software BOCOP (Sec. VI.2.5).

### VI.2.1 The model system

We consider a three-level quantum system in a  $\Lambda$  configuration whose dynamics are governed by the Schrödinger equation. The system is described by a pure state  $|\psi(t)\rangle$  which belongs to a three-dimensional Hilbert space spanned by the basis  $\{|1\rangle, |2\rangle, |3\rangle\}$ . The system is subject to a pump and a Stokes pulses coupling, respectively, states  $|1\rangle$  and  $|2\rangle$  and states  $|2\rangle$  and  $|3\rangle$ . They are assumed to be on-resonance with the corresponding frequency transitions. There is no direct coupling between levels  $|1\rangle$  and  $|3\rangle$ . A schematic representation of the control problem is given in Fig. VI.6.

The time evolution of the system is given by the Schrödinger equation

$$i \frac{\partial}{\partial t} |\psi(t)\rangle = H |\psi(t)\rangle$$

where units such that  $\hbar = 1$  are used. In the interaction representation and in the rotating-wave approximation, the Hamiltonian of the system can be written as

$$H = \begin{pmatrix} 0 & \Delta & 0 \\ \Delta & 0 & u \\ 0 & u & 0 \end{pmatrix}$$

where  $\Delta$  and  $u$  represent the Rabi frequencies of the two pulses. We denote by  $c_1$ ,  $c_2$ , and  $c_3$  the complex coefficients of the state  $|\psi(t)\rangle$ , and we introduce the real coefficients  $x_1, \dots, x_6$  defined by  $c_1 = x_1 + ix_4$ ,  $c_2 = x_5 - ix_2$ ,  $c_3 = x_3 + ix_6$ . Straightforward computations from the Schrödinger equations show that the variables  $x_1$ ,  $x_2$ , and  $x_3$  are decoupled from  $x_4$ ,  $x_5$ , and  $x_6$ . For our purposes it is sufficient to study the dynamics of the first set of variables, which turn out to be

$$\begin{cases} \dot{x}_1 = -\Delta x_2, \\ \dot{x}_2 = \Delta x_1 - ux_3, \\ \dot{x}_3 = ux_2. \end{cases}$$

## VI.2.2 The Fuller model

This paragraph briefly describes the main results that can be established for the classical Fuller model. The interested reader will find the proofs of the different statements in textbooks of mathematical control theory [SL12; ZB94].

The Fuller model is a linear optimal control problem in  $\mathbb{R}^2$ . The dynamics of the state  $(x, y)$  are governed by the differential equations

$$\begin{cases} \dot{x} = y, \\ \dot{y} = u, \end{cases} \quad (\text{VI.6})$$

where the control  $u = u(t)$  is subject to the constraint  $u(t) \in [-1, 1]$ . Starting from the state  $(x_0, y_0)$ , the goal of the control is to reach the origin  $(0, 0)$ , while minimizing the cost functional  $\mathcal{J} = \int_0^{t_f} x^2(t) dt$ , in which the control time  $t_f$  is not fixed.

The existence of optimal solutions to the problem above is not obvious, and can be proved by showing that the Hamilton–Jacobi–Bellman equation satisfied by the value function has a classical solution (see [SL12, Theorem 5.1.1 and Example 5.1.2]).

The optimal trajectories have the discrete symmetry

$$(x(t), y(t), u(t)) \mapsto (-x(t), -y(t), -u(t)),$$

and a scaling symmetry defined by the family of transformations

$$(x(t), y(t), u(t)) \mapsto (x_\lambda(t) = \lambda^2 x(t/\lambda), y_\lambda(t) = \lambda y(t/\lambda), u_\lambda(t) = u(t/\lambda)),$$

where  $\lambda$  is a positive parameter. This means that if  $(x(t), y(t), u(t))$  is an optimal solution then  $(x_\lambda(t), y_\lambda(t), u_\lambda(t))$  is also solution, with initial state  $(\lambda^2 x_0, \lambda y_0)$  and cost  $\mathcal{J}_\lambda = \lambda^5 \mathcal{J}$ . We deduce that if  $(\bar{x}, \bar{y})$  is a switching point for an optimal trajectory, i.e., the corresponding control goes from  $\pm 1$  to  $\mp 1$  when the trajectory crosses  $(\bar{x}, \bar{y})$ , then the optimal synthesis has a switching curve of equation  $x = -\xi \text{sign}[y]y^2$ , where  $\xi$  is a positive constant such that the curve passes through  $(\bar{x}, \bar{y})$ .

The second step of the analysis consists in applying the Pontryagin Maximum Principle, which is a necessary condition for optimality [Pon+74]. The Pontryagin Hamiltonian can be expressed as

$$H_P = p_x y + p_y u + p^0 x^2,$$

where  $(p_x, p_y)$  is the adjoint state and  $p^0$  is a constant multiplier equal either to 0 or to  $-\frac{1}{2}$ . If  $(x(t), y(t))$  is an optimal trajectory with corresponding control  $u(t)$  then there exist  $(p_x(t), p_y(t))$  and  $p^0$  such that  $(p_x(t), p_y(t), p^0) \neq (0, 0, 0)$ ,  $\dot{x} = \frac{\partial H_P}{\partial p_x}$ ,  $\dot{y} = \frac{\partial H_P}{\partial p_y}$ ,  $\dot{p}_x = -\frac{\partial H_P}{\partial x}$ ,  $\dot{p}_y = -\frac{\partial H_P}{\partial y}$ , and  $H_P(x(t), y(t), p_x(t), p_y(t), u(t)) = \max_{v \in [-1, 1]} H_P(x(t), y(t), p_x(t), p_y(t), v) = 0$ .

The switching function  $\Phi$  is given by  $\Phi(t) = p_y(t)$ . Integrating system (VI.6) from the state  $(-\xi y_0^2, y_0)$  with  $y_0 > 0$  and  $u = +1$ , and imposing that the corresponding switching function vanishes at a point  $(\xi y_1^2, y_1)$  with  $y_1 > 0$ , we obtain that  $\xi$  is solution of a polynomial equation of order 4, given by

$$\xi = \sqrt{\frac{\sqrt{33} - 1}{24}} \simeq 0.44623. \quad (\text{VI.7})$$

The optimal trajectory has the following symmetries: denoting by  $t_k$  the  $k$ -th switching time, by  $t_f^{\text{Ful}}$  the minimum time at which the optimal trajectory reaches  $(0, 0)$ , and introducing the parameter  $\alpha = \sqrt{\frac{1+2\xi}{1-2\xi}} \simeq 4.13016$ , we have

$$t_f^{\text{Ful}} - t_{k+1} = \frac{t_f^{\text{Ful}} - t_k}{\alpha}, \quad \frac{x(t_f^{\text{Ful}} - t_{k+1})}{x(t_f^{\text{Ful}} - t_k)} = -\frac{1}{\alpha^2}, \quad \frac{y(t_f^{\text{Ful}} - t_{k+1})}{y(t_f^{\text{Ful}} - t_k)} = -\frac{1}{\alpha}.$$

Finally, starting from the state  $(-\xi y_0^2, y_0)$ ,  $y_0 > 0$ , it can be shown that  $t_f^{\text{Ful}} = \frac{1+\alpha}{\alpha-1} y_0$ . An example of optimal trajectory is plotted in Fig. VI.7. Figure VI.8 compares the results obtained from the quantum system to its linear approximation. We consider the projection onto the  $(x_1, x_2)$ - plane of the optimal trajectory. This comparison highlights that the two solutions are very close to each other near the target state.

### VI.2.3 Properties of the switching function for the three-level quantum system

We focus in this paragraph on the switching function  $\Phi$  for the three-level quantum system. Using the Hamiltonian equations for the adjoint state in the normal case (Eq. (VI.2)), one computes that  $\Phi$  and its time derivatives can be expressed as

$$\begin{aligned} \Phi &= p_3 x_2 - p_2 x_3, \\ \dot{\Phi} &= \Delta(x_1 p_3 - x_3 p_1), \\ \ddot{\Phi} &= -\Delta^2 \Phi + \Delta u(x_1 p_2 - x_2 p_1) - \Delta x_1 x_3. \end{aligned}$$

Moreover, on a segment where  $\Phi \neq 0$  and, therefore,  $u$  is constantly equal to  $+1$  or  $-1$ , one has

$$\begin{aligned} \Phi^{(3)} &= -(\Delta^2 + 1)\dot{\Phi} - 2\Delta u x_1 x_2 + \Delta^2 x_2 x_3, \\ \Phi^{(4)} &= -(\Delta^2 + 1)\ddot{\Phi} + \Delta(\Delta^2 + 2)x_1 x_3 + \Delta^2 u(3x_2^2 - 2x_1^2 - x_3^2). \end{aligned}$$

Note that, if  $x_2 \neq 0$  and  $\Phi = 0$ , then  $p_3 = \frac{x_3}{x_2} p_2$  and hence

$$\dot{\Phi} = \Delta \frac{x_3}{x_2} (x_1 p_2 - x_2 p_1).$$

Notice also that, since  $H_P = 0$ , then, using again that  $\Phi = 0$ , we have

$$x_1 p_2 - x_2 p_1 = \frac{x_1^2}{2} \geq 0.$$

Hence, for  $x_2 x_3 \neq 0$  and  $x_1 \neq 0$ , the derivative  $\dot{\Phi}$  has the same sign as  $x_2 x_3$ , which means that switches only occur from  $-1$  to  $+1$  if  $x_2 x_3 > 0$  and from  $+1$  to  $-1$  if  $x_2 x_3 < 0$ . If  $x_2 \neq 0$



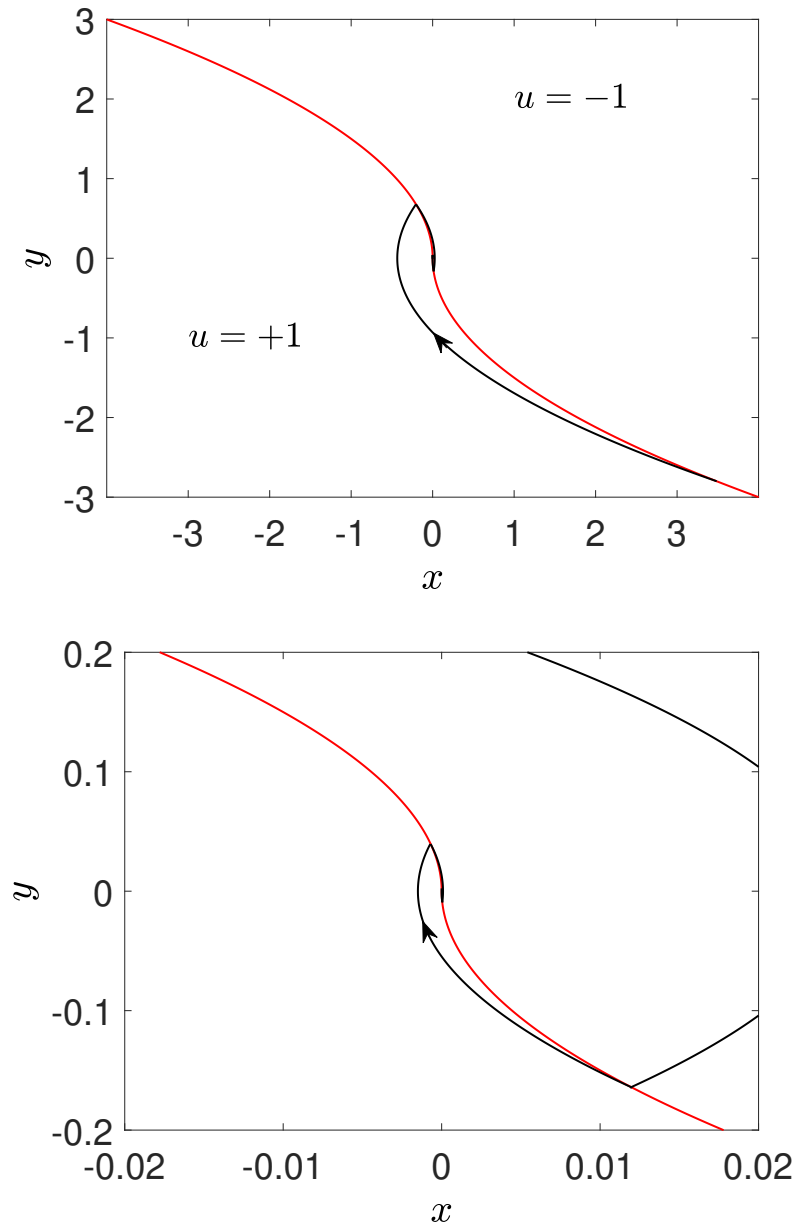


Figure VI.7 – (top) Optimal trajectory for the Fuller model (black solid line). The switching curves are plotted in red. (bottom) Zoom of the top panel near the origin

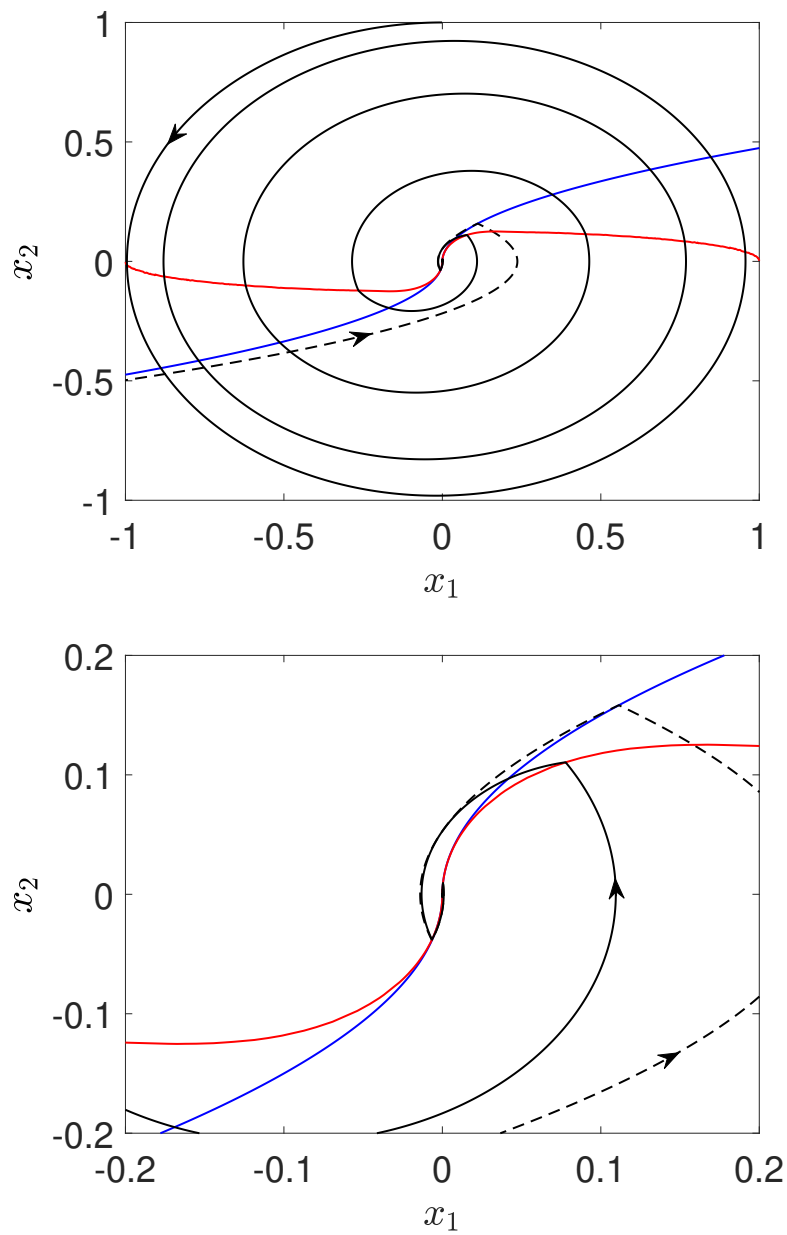


Figure VI.8 – Plot onto the  $(x_1, x_2)$ - plane of the optimal trajectory of the quantum control problem (solid black line). The dashed line depicts the solution of the linear approximation. The red and blue curves represent respectively the switching curves for the quantum and Fuller systems. The bottom panel is a zoom of the top one.

and  $x_1 = 0$  then  $\dot{\Phi} = 0$ , which implies that  $p_1 = 0$ . Therefore  $\ddot{\Phi} = 0$  and

$$\Phi^{(3)} = \Delta^2 x_2 x_3.$$

It follows that everywhere on the set  $\{x \mid x_2 x_3 > 0\}$  switches only occur from  $-1$  to  $+1$ , while on  $\{x \mid x_2 x_3 < 0\}$  they only occur from  $+1$  to  $-1$ .

## VI.2.4 A sufficient condition for chattering

In this section we present an adaptation of the results presented in Chapter 3 of the book [ZB94] by Zelikin and Borisov, concerning sufficient conditions for the appearance of Fuller-like chattering in a two-dimensional optimal synthesis.

We consider a control system of the form

$$\begin{cases} \dot{x} = \Delta y + \phi_1^x(x, y) + u\phi_2^x(x, y), \\ \dot{y} = u + \phi_1^y(x, y) + u\phi_2^y(x, y). \end{cases} \quad (\text{VI.8})$$

We also suppose that  $\phi_i^x, \phi_i^y$  are smooth and small in the following sense: denoting by  $g_\kappa$  the anisotropic dilatation  $g_\kappa(x, y) = (\kappa^2 x, \kappa y)$ , one has

$$\limsup_{\kappa \rightarrow 0^+} \frac{|\phi_i^x(g_\kappa(x, y))|}{\kappa} < \infty, \quad \limsup_{\kappa \rightarrow 0^+} \frac{|\phi_i^y(g_\kappa(x, y))|}{\kappa^2} < \infty, \quad (\text{H})$$

for every  $(x, y) \in \mathbb{R}^2$ . Then, the optimal control problem

$$\begin{cases} \int_0^T x^2(t) dt \longrightarrow \min \\ T > 0 \text{ free, } u \in L^\infty([0, T], [-1, 1]) \\ t \mapsto (x(t), y(t)) \in W^{1, \infty}([0, T], S^2) \text{ solution of (VI.8)} \\ (x(0), y(0)) = (x_0, y_0) \end{cases} \quad (\text{VI.9})$$

satisfies the following properties.

**Theorem VI.2.** *For every  $(x_0, y_0)$  in a sufficiently small neighbourhood of  $(0, 0)$ , Problem (VI.9) admits a solution. Such a solution reaches  $(0, 0)$  in finite time and its corresponding control has infinitely many discontinuities accumulating at the final time. Moreover, there is no other solution of (VI.9) up to prolongation by a constant trajectory at  $(0, 0)$ . The optimal synthesis is characterized by a switching curve of the form*

$$\Gamma = \begin{cases} x = \lambda_1(y)y^2, & y > 0, \\ x = \lambda_0(y)y^2, & y < 0, \end{cases} \quad (\text{VI.10})$$

where  $\lambda_0$  and  $\lambda_1$  are  $\mathcal{C}^1$  function satisfying  $\lambda_0(0) = -\lambda_1(0) = \xi\Delta$ , where  $\xi$  is as in (VI.7). The optimal control is  $-1$  above  $\Gamma$  and  $+1$  below it.

All the proof material is in the book [ZB94] by M. I. Zelikin and V. F. Borisov. Nevertheless, since this theorem is not explicitly stated as such, we propose to retrace the ingredients of the proof and point out the relevant statements in [ZB94].

*Proof.* First, let us highlight the fact that the existence of a solution of (VI.9) is not obvious. Indeed, it is a priori possible that the cost decreases by reaching the target later in time. As a consequence we are lacking compactness for existence of the solution of the problem with free

final time. Existence of optimal trajectories can be obtained once an extremal synthesis has been constructed, using a field-of-extremals argument ([ZB94, Theorem 3.3]).

The Hamiltonian given by the PMP for Problem (VI.9) is

$$H_P(x, y, p_x, p_y, p_0, u) = p_x (\Delta y + \phi_1^x(x, y) + u\phi_2^x(x, y)) \\ + p_y (u + \phi_1^y(x, y) + u\phi_2^y(x, y)) + p_0 x^2.$$

Hence, the adjoint equations for normal extremals ( $p_0 = -1/2$ ) are

$$\begin{cases} \dot{p}_x = x - p_x \partial_x (\phi_1^x + u\phi_2^x) - p_y \partial_x (\phi_1^y + u\phi_2^y), \\ \dot{p}_y = -\Delta p_x - p_x \partial_y (\phi_1^x + u\phi_2^x) - p_y \partial_y (\phi_1^y + u\phi_2^y). \end{cases} \quad (\text{VI.11})$$

If we define  $(z_1, z_2, z_3, z_4) = ((p_y + p_x \phi_2^x + p_y \phi_2^y)/\Delta^2, -p_x/\Delta, -x/\Delta, -y)$ , we get

$$\begin{cases} \dot{z}_1 = z_2 + f_1(z, u), & \dot{z}_2 = z_3 + f_2(z, u), \\ \dot{z}_3 = z_4 + f_3(z, u), & \dot{z}_4 = u + f_4(z, u), \\ u = \text{sign}[z_1]. \end{cases} \quad (\text{VI.12})$$

Introducing  $\tilde{g}_\kappa(z_1, z_2, z_3, z_4) = (\kappa^4 z_1, \kappa^3 z_2, \kappa^2 z_3, \kappa z_4)$ , we have the smallness property

$$\limsup_{\kappa \rightarrow 0} \frac{|f_i(\tilde{g}_\kappa(z), u)|}{\kappa^{5-i}} < \infty \quad \text{for } 1 \leq i \leq 4, \quad (\text{VI.13})$$

which generalises (H). As a consequence, this system is of the form given in Equation (3.5) of [ZB94]. The smallness condition (VI.13) allows to apply [ZB94][Proposition 4.1], which guarantees, under a rank condition discussed below, that the conclusions of Theorem 3.3 in [ZB94] hold true.

The study of this dynamical system around  $z = 0$  is performed in Chapter 3 of [ZB94] thanks to a blow-up procedure. The (degenerate) hyperbolicity of the fixed point  $z = 0$  is established, as well as the existence of a two-dimensional invariant contracting manifold  $\Sigma$  corresponding to the trajectories converging to  $z = 0$  [ZB94, Section 3.4 - 3.8]. This manifold is given by the trajectories of the non-smooth system (VI.12) switching on curves of the form

$$\hat{\Gamma}^0 = \{(0, \mu_0(\kappa)\kappa^3, \lambda_0(\kappa)\kappa^2, \kappa) \mid \kappa < 0\}, \quad (\text{VI.14})$$

$$\hat{\Gamma}^1 = \{(0, \mu_1(\kappa)\kappa^3, \lambda_1(\kappa)\kappa^2, \kappa) \mid \kappa > 0\}, \quad (\text{VI.15})$$

where  $\mu_i, \lambda_i$  are smooth functions satisfying (cf. [ZB94, Lemma 3.3])

$$\lambda_0(0) \in \left(-\frac{\Delta}{2}, 0\right), \quad \lambda_1(0) \in \left(0, \frac{\Delta}{2}\right), \quad \mu_0(0) = \frac{1}{2}\lambda_0(0)^2, \quad \mu_1(0) = \frac{1}{2}\lambda_1(0)^2. \quad (\text{VI.16})$$

Besides, since  $\lambda_0(0)$  and  $\lambda_1(0)$  are solutions of the same polynomial system than in the case of the following Füller dynamics

$$\begin{cases} \dot{x} = \Delta y, \\ \dot{y} = u, \end{cases} \quad (\text{VI.17})$$

we get  $\lambda_0(0) = -\lambda_1(0) = \xi\Delta$ .

Let us now check that the projection  $\pi : (z_1, z_2, z_3, z_4) \mapsto (z_3, z_4)$  restricted to  $\Sigma$  is a  $\mathcal{C}^1$  mapping with a Jacobian matrix of maximal rank on  $\Sigma \setminus \hat{\Gamma}$ , as required in [ZB94, Theorem 3.3]. Let us

denote by  $\mathcal{F}(\kappa, t)$  the solution of System (VI.12) at time  $t$  with initial condition parameterized by  $\kappa$  as in (VI.14) (an analogous argument applying for the points of  $\Sigma$  parameterized by the trajectories of System (VI.12) starting from  $\hat{\Gamma}^1$ ). Let  $t$  be such that  $\mathcal{F}(\kappa, t) \notin \hat{\Gamma}^0 \cup \hat{\Gamma}^1$  for all  $0 < s < t$ . Then  $\mathcal{F}(\kappa, s)$  is given by the solution of (VI.12) with  $u = -1$  and  $\pi(\mathcal{F}(\kappa, s))$  is the solution of System (VI.11) with constant control  $u = -1$  and initial condition  $(x_0, y_0) = (\lambda_0(\kappa)\kappa^2, \kappa)$ . Therefore,  $(\kappa, s) \mapsto \pi(\mathcal{F}(\kappa, s))$  is smooth. Let us check that  $\partial_\kappa \pi(\mathcal{F}(\kappa, t))$  and  $\partial_t \pi(\mathcal{F}(\kappa, t))$  are linearly independent. Using the fact that the flow of (VI.11) is a diffeomorphism, it is enough to check that the two vectors

$$\partial_\kappa \pi(\mathcal{F}(\kappa, 0)) = (\lambda_0'(\kappa)\kappa^2 + 2\lambda_0(\kappa)\kappa, 1)$$

and

$$\partial_t \pi(\mathcal{F}(\kappa, 0)) = (\kappa + f_3(z, -1), -1 + f_4(z, -1)),$$

are linearly independent. Using (VI.16), this is the case for  $\kappa$  small enough. We can thus apply [ZB94][Theorem 3.3], which allows to conclude the proof of Theorem VI.2.  $\square$

## VI.2.5 Numerical optimization procedure

We apply in this paragraph a direct optimization method to design numerically the solution of the optimal control problem. We use the open access optimal solver BOCOP [BGM11]. A direct optimization approach is a procedure in which the state and the time are discretized, transforming the initial optimal control problem into a nonlinear constrained optimization problem. In this optimization protocol, the optimal control problem is slightly modified. The goal is to minimize the cost  $\mathcal{C} = \int_0^{t_f} x_1^2(t) dt$ , while reaching a state as close as possible of the target  $|3\rangle$  in a fixed time  $t_f$ . In order to have the fairest comparison possible between the two approaches, we estimate as follows the time of the optimal solution presented in the previous section as follows. We assume that the control time is the sum of  $t_f^{\text{ful}}$  and of the nonlinear part used to steer the system from  $(0, 1, 0)$  to  $\mathbf{X}^{(0)}$ . In the Fuller's example, it can be shown that, starting from a point on a switching curve of the form  $\text{sign}[x_{20}] \xi \Delta x_{20}^2, x_{20}$ , the total time to reach exactly the target is  $t_f^{\text{ful}} = \frac{\alpha+1}{\alpha-1} |x_{20}|$ . In the quantum control problem, we obtain respectively for the nonlinear and linear times, 2.5890 and 0.0011, which leads to  $t_f = 2.5901$ . We set  $t_f$  to 2.59 in the numerical optimization procedure. The time subdivision is regular and given by the number of time steps  $N$ , going from  $N = 50$  to  $N = 400$ . The other optimization parameters are set to their default or recommended values in BOCOP.

In addition of the controls displayed in Fig. VI.5, we report in Fig. VI.9 additional numerical results about the distance to the target and the cost  $\mathcal{C}$  with respect to  $N$ . When  $N$  becomes larger, the numerical optimal process seems to converge towards the optimal solution both in terms of final state and cost. For quite small values of  $N$ , we observe that the efficiency of the control is already reasonable. We point out that such sub-optimal strategy could be a possible option to bypass the chattering phenomenon in the optimization procedure.

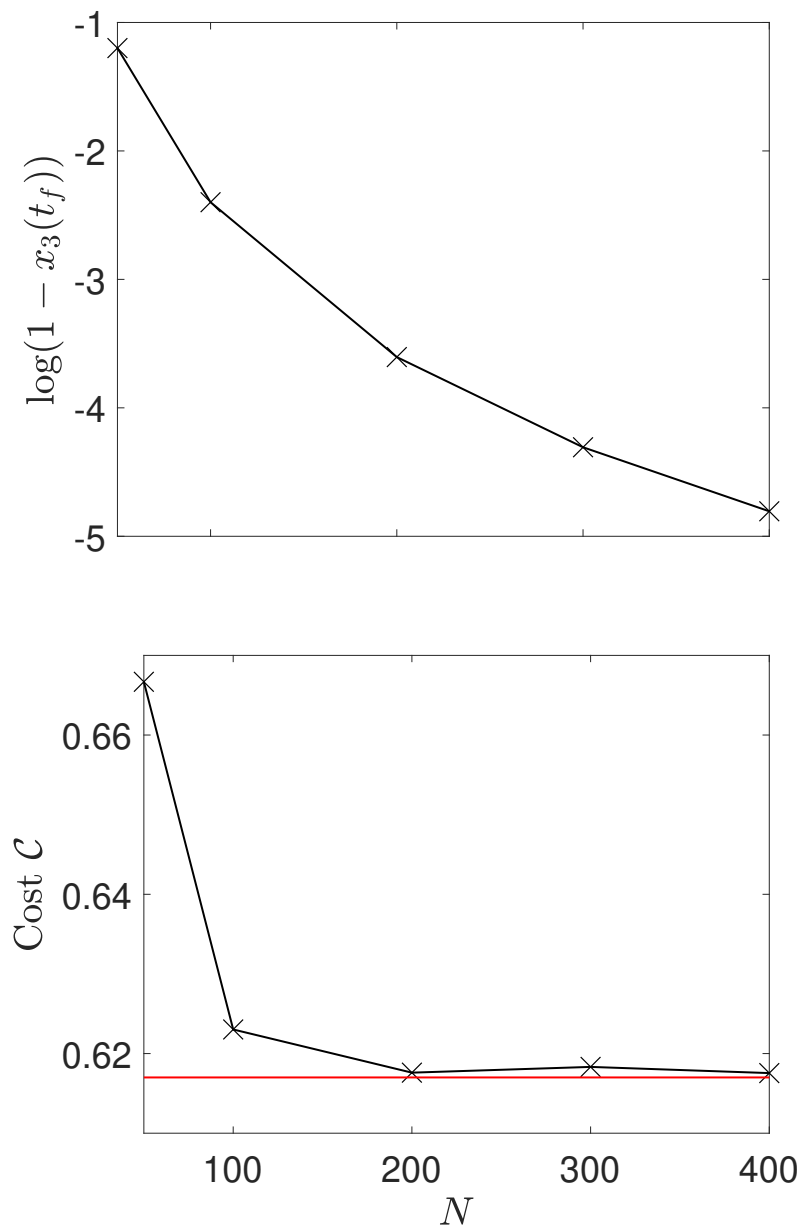


Figure VI.9 – Evolution of the distance to the target (top) and of the cost  $\mathcal{C}$  (bottom) as a function of the number of time steps (crosses). In the bottom panel, the horizontal solid line (in red) indicates the efficiency of the optimal solution. The solid black line is just to guide the lecture in the two panels.



Third Part: Controllability of one  
dimensional fluid dynamics  
equations





## Chapter VII

# Small-time global null controllability of generalized Burgers' equations

This chapter is taken from the following submitted article (also referred as [Rob22]):

R. Robin. *Small-time global null controllability of generalized Burgers' equations*. 2022. arXiv: 2206.05931 [math]

We refer to Section I.7 for an introduction to the motivations behind the study of these equations.

In this chapter, we study the small-time global null controllability of the generalized Burgers' equations  $y_t + (|y|^\gamma)_x - y_{xx} = u(t)$  on the segment  $[0, 1]$ . The scalar control  $u(t)$  is uniform in space and plays a role similar to the pressure in higher dimension. We set a right Dirichlet boundary condition  $y(t, 1) = 0$ , and allow a left boundary control  $y(t, 0) = v(t)$ . Under the assumption  $\gamma > 3/2$  we prove that the system is small-time global null controllable. Our proof relies on the return method and a careful analysis of the shape and dissipation of a boundary layer.

## VII.1 Introduction

### VII.1.1 Description of the system

For a given  $T > 0$ , we are concerned with the following generalized Burgers' equations on the segment  $[0, 1]$ :

$$\begin{cases} y_t + (|y|^\gamma)_x - y_{xx} = u(t) & \text{on } (0, T) \times (0, 1), \\ y(t, 0) = v(t) & \text{on } (0, T), \\ y(t, 1) = 0 & \text{on } (0, T), \\ y(0, x) = y_0(x) & \text{on } (0, 1), \end{cases} \quad (E_\gamma)$$

where  $u(t)$  is an interior control which does not depend on space, and  $v(t)$  is a boundary control. We are interested in the small-time global null controllability. That is, for any initial (possibly large) datum  $y_0$  and any (possibly small) final time  $T$ , can we find some controls  $u$  and  $v$  such that the solution of Eq.  $(E_\gamma)$  is steered to 0 in time  $T$ ?

### VII.1.2 Statement of our main result

We have to provide a reasonable definition for the solutions of Eq.  $(E_\gamma)$ . Namely, let us consider the following generalisation:

$$\begin{cases} y_t + (|y|^\gamma)_x - y_{xx} = u(t) & \text{on } (0, T) \times (0, 1), \\ y(t, 0) = v(t) & \text{on } (0, T), \\ y(t, 1) = w(t) & \text{on } (0, T), \\ y(0, x) = y_0(x) & \text{on } (0, 1), \end{cases} \quad (F_\gamma)$$

where the scalar controls are  $u \in L^\infty(0, T)$  and  $v, w \in H^{1/4}(0, T) \cap L^\infty(0, T)$ . We also assume  $\gamma \geq 1$  and  $y_0 \in L^\infty(0, 1)$ , then there exists a unique solution of  $(F_\gamma)$  in  $\mathcal{C}^0([0, T]; L^2) \cap L^2(0, T; H^1) \cap L^\infty((0, T) \times (0, 1))$ . We do not provide the proof here, it is based on apriori energy estimates and a fixed point argument (see [Lio69]). For the interested reader, we mention that well-posedness in less regular space is studied in [Bek96; LZZ19].

Let us now state our main contribution.

**Theorem VII.1.** *Suppose  $\gamma > 3/2$ ,  $y_0 \in L^\infty(0, 1)$  and  $T > 0$ . Then, there exist  $u \in L^\infty(0, T)$  and  $v \in H^{1/4}(0, T) \cap L^\infty(0, T)$  steering the solution  $y$  of  $(E_\gamma)$  to the null state in time  $T$ .*

**Remark VII.2.** *We are not able to tackle the case  $\gamma \in (1, 3/2]$ , and we believe that our method cannot be used to tackle this (entire) range of  $\gamma$ . We comment this point in Section VII.6.*

**Remark VII.3.** *If one replaces the non-linearity  $(|y|^\gamma)_x$  in Eq.  $(E_\gamma)$  with  $|y|^{\gamma-1}y_x$ , Theorem VII.1 remains valid for  $\gamma > 2$ . Adaptation of the proof is straightforward except for Section VII.4. In that section, our proof does not extend to  $\gamma \leq 2$  (there is a sign issue in Eq. (VII.71)).*

The sketch of the proof is the following: to reach the null state in arbitrary small-time, we take advantage of the return method introduced by Coron in [Cor92] (see also [Cor09]), and more specifically, the three-stages strategy developed by Marbach in [Mar14].

- Hyperbolic stage, first part: we introduce the steady state  $\vartheta$  of  $(E_\gamma)$  with  $u = 0$  and  $v = \theta \gg 1$ :

$$\begin{cases} \vartheta_{xx} = (\vartheta^\gamma)_x, \\ \vartheta(0) = \theta, \quad \vartheta(1) = 0. \end{cases} \quad (\text{VII.1})$$

Note that  $\vartheta$  exhibits a boundary layer near the right endpoint. Using the hyperbolic nature of the equation when it is governed by the non-linear term, we prove that we can approximate  $\vartheta$  in small-time.

- Hyperbolic stage, second part: we use the pressure-like term to drive our system to a neighborhood of the null state up to a boundary residue around  $x = 1$ .
- Passive stage: we do not apply any control and wait for the dissipation of the boundary residue in small-time. The assumption  $\gamma > 3/2$  is crucial for this stage.
- Parabolic stage: we provide a local exact controllability result around zero using a fixed point method.

The proof of Theorem VII.1 is then obtained by the combination of these stages resumed in Theorems VII.7 and VII.9 to VII.11.

### VII.1.3 Preliminaries

#### VII.1.3.1 Comparison principle

Let us recall the comparison principle for semi-linear parabolic equations. Let us suppose that  $y$  (resp.  $\tilde{y}$ ) is a solution of  $(F_\gamma)$  with controls  $u, v, w$  (resp.  $\tilde{u}, \tilde{v}, \tilde{w}$ ) and initial condition  $y_0$  (resp.  $\tilde{y}_0$ ), such that a.e.,

$$\begin{aligned} u &\leq \tilde{u}, & v &\leq \tilde{v}, \\ w &\leq \tilde{w}, & y_0 &\leq \tilde{y}_0. \end{aligned}$$

Then,

$$y \leq \tilde{y}.$$

We refer to [PS07]. One can also easily extend the proof given by Marbach in [Mar14].

#### VII.1.3.2 Study of the steady states

It is crucial in our proof to have a reasonable description of the steady states  $\vartheta$  defined by Eq. (VII.1).

**Lemma VII.4.** *Let  $\theta > 0$ , then Eq. (VII.1) admits a unique solution. Besides,*

$$\vartheta_x = |\vartheta|^\gamma + \vartheta_x(1) \quad \text{with} \quad -\theta^\gamma - \theta < \vartheta_x(1) < -\theta^\gamma \quad (\text{VII.2})$$

and  $\theta \mapsto \vartheta_x(1)$  is decreasing.

*Proof.* To prove the existence, let us consider the application given as follows: for any  $(\theta, C) \in \mathbb{R}^2$ , we associate the (local) solution  $y$  of

$$y(0) = \theta, \quad y_x = |y|^\gamma + C. \quad (\text{VII.3})$$

If  $C = -\theta^\gamma$ , then  $y$  is the constant function  $\theta$ . Conversely, if  $C \leq -\theta^\gamma - \theta$ ,  $y$  reaches 0 in  $(0, 1)$ . Hence, by the intermediate value theorem, there exists  $C$  such that  $y(1) = 0$ . Besides, as  $C \mapsto y(1)$  is an increasing function, we proved the uniqueness of the solution of Eq. (VII.1).

Moreover, if one consider the solutions  $(y_1, y_2)$  of Eq. (VII.3) for  $\theta_1 < \theta_2$  with the same  $C$ , then  $\forall x \in [0, 1]$ ,  $y_1(x) \leq y_2(x)$ . Hence,  $\theta \mapsto \vartheta_x(1)$  is increasing.  $\square$

Note that Eq. (VII.2) also shows that  $\vartheta$  is a non-negative, decreasing and concave function.

**Remark VII.5.**

— In the case of the usual Burgers' equation,

$$\vartheta(x) = \hat{\theta} \tanh(\hat{\theta}(1-x)), \quad (\text{VII.4})$$

where  $\hat{\theta}$  is the unique solution of  $\hat{\theta} \tanh(\hat{\theta}) = \theta$ .

— For the linear case, i.e. with  $\gamma = 1$  (which is not included in the assumption of Theorem VII.1), we have

$$\vartheta(x) = \frac{(e - e^x)}{e - 1} \theta. \quad (\text{VII.5})$$

For a general  $\gamma$ , we are not aware of an explicit expression for  $\vartheta$ . If  $\gamma > 1$ ,  $\vartheta$  presents a boundary layer at the right endpoint (see Fig. VII.1) which is characterized in the next lemma.

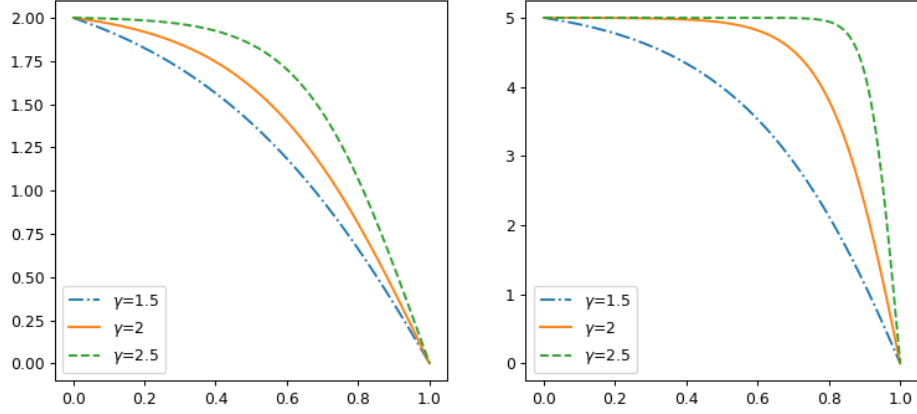


Figure VII.1 – Steady states  $\vartheta$  with  $\theta = 2$  and  $\theta = 5$  for different values of  $\gamma$ .

**Lemma VII.6.** *Let  $\gamma > 1$  and  $\vartheta$  the solution of Eq. (VII.1). Then  $\vartheta$  exhibits a boundary layer of size  $\frac{1}{\theta^{\gamma-1}}$  around  $x = 1$ . More precisely, the following estimates hold:*

1. *Let  $a \in (0, 1)$ , then for all  $\theta > 0$ ,*

$$\vartheta(x) \geq a\theta, \quad \forall x \in [0, 1 - \frac{a}{1 - a^\gamma} \theta^{1-\gamma}]. \tag{VII.6}$$

2. *Let  $\varepsilon > 0$  and  $\alpha < \gamma - 1$ , then there exists  $C_{\varepsilon, \alpha} > 0$  such that*

$$\vartheta(x) \geq \theta - \varepsilon, \quad \forall x \in [0, 1 - C_{\varepsilon, \alpha} \theta^{-\alpha}]. \tag{VII.7}$$

*Proof.* For  $x \in [0, 1]$ , according to Eq. (VII.2),

$$\vartheta_x(x) = \vartheta(x)^\gamma + \vartheta_x(1) \leq \vartheta(x)^\gamma - \theta^\gamma < 0.$$

Thus, for any  $x^* \in (0, 1)$ ,

$$\int_{x^*}^1 \frac{-\vartheta_x(x) dx}{\theta^\gamma - \vartheta^\gamma(x)} \geq 1 - x^*.$$

By the change of variable formula with  $z = \vartheta(x)$ ,

$$\int_0^{\vartheta(x^*)} \frac{dz}{\theta^\gamma - z^\gamma} \geq 1 - x^*. \tag{VII.8}$$

To prove the first point, let  $x^* \in (0, 1)$  be the unique solution to the equation  $\vartheta(x^*) = a$ . Injecting into Equation (VII.8), we get

$$a\theta \frac{1}{\theta^\gamma(1 - a^\gamma)} \geq \int_0^{\vartheta(x^*)} \frac{dz}{\theta^\gamma - z^\gamma} \geq 1 - x^*.$$

Thus,

$$x^* \geq 1 - \frac{a}{1-a^\gamma} \theta^{1-\gamma}.$$

For the second point, let  $x^* \in (0, 1)$  be the unique solution to the equation  $\vartheta(x^*) = \theta - \varepsilon$ . Let us use (VII.8) again, and apply the change of variable  $s = \theta z$ ,

$$\int_0^{\vartheta(x^*)} \frac{dz}{\theta^\gamma - z^\gamma} = \theta^{1-\gamma} \int_0^{1-\varepsilon/\theta} \frac{ds}{1-s^\gamma}.$$

For  $\gamma > 1$ , we have  $\frac{1}{1-s} > \frac{1}{1-s^\gamma}$ . Thus,

$$-\theta^{1-\gamma} \ln\left(\frac{\varepsilon}{\theta}\right) > 1 - x^*.$$

Using  $-\theta^{1-\gamma} \ln\left(\frac{\varepsilon}{\theta}\right) = o(\theta^{-\alpha})$ , for a given couple  $(\alpha, \varepsilon)$ , there exists  $C_{\varepsilon, \alpha}$  such that

$$x^* \geq 1 - C_{\varepsilon, \alpha} \theta^{-\alpha}.$$

□

## VII.2 Hyperbolic stage, first part: toward a very stable steady state

### VII.2.1 The control strategy

The goal of this section is to prove the following proposition.

**Proposition VII.7.** *For a given  $y_0 \in L^\infty(0, 1)$ ,  $\eta > 0$  and  $T > 0$ , one can find  $\theta_0 > 0$  such that the following holds: for any  $\theta \geq \theta_0$ , there exist  $u, v \in L^\infty(0, T) \times (L^\infty(0, T) \cap H^{1/4}(0, T))$  such that the solution  $y$  of  $(E_\gamma)$  satisfies*

$$\vartheta(x) \leq y(T, x), \quad (\text{VII.9})$$

and

$$y(T, x) \leq \theta + \eta. \quad (\text{VII.10})$$

For the proof of Theorem VII.7, we use the following controls on  $[0, T]$

$$u(t) = \begin{cases} \frac{\theta + 2\|y_0\|_{L^\infty}}{T'} & \text{for } t \leq T', \\ 0 & \text{for } t > T', \end{cases} \quad (\text{VII.11})$$

and

$$v(t) = \begin{cases} \frac{(\theta + \|y_0\|_{L^\infty})t}{T'} & \text{for } t \leq T', \\ \theta + \frac{\|y_0\|_{L^\infty}(\frac{T}{2} - t)}{\frac{T}{2} - T'} & \text{for } T' < t \leq T/2, \\ \theta & \text{for } t > T/2, \end{cases} \quad (\text{VII.12})$$

where  $T' < \frac{T}{2}$  will be chosen small enough later. We denote  $y$  the solution of  $(E_\gamma)$  associated with these controls. We prove separately the lower bound (VII.9) and the upper bound (VII.10).

**Remark VII.8.** *In what follows, we use  $C$  for a constant which is independent of  $\theta$  and may be different from one line to the next.*

## VII.2.2 Lower bound

The idea to handle the lower bound is to use a very small time  $T'$ . Let us introduce

$$\underline{v}(t) = \frac{(\theta + 2\|y_0\|_{L^\infty})t}{T'} - \|y_0\|_{L^\infty} \leq v(t) \text{ for all } t \leq T', \quad (\text{VII.13})$$

and set

$$T'_0 = \frac{T'\|y_0\|_{L^\infty}}{(\theta + 2\|y_0\|_{L^\infty})}. \quad (\text{VII.14})$$

Leading to,  $\forall t \leq T'_0$ ,  $\underline{v}(t) \leq 0$ .

We define the following subsolution  $\underline{y}$  of  $y$  on  $(0, T'_0)$ :

$$\begin{cases} \underline{y}_t + (|\underline{y}|^\gamma)_x - \underline{y}_{xx} = u(t) & \text{on } (0, T'_0) \times (0, 1), \\ \underline{y}(t, 0) = \underline{v}(t) & \text{on } (0, T'_0), \\ \underline{y}(t, 1) = \underline{v}(t) & \text{on } (0, T'_0), \\ \underline{y}(0, x) = -\|y_0\|_{L^\infty} & \text{on } (0, 1). \end{cases} \quad (\text{VII.15})$$

We easily check that  $\underline{y}(t) = \underline{v}(t)$  for all  $t \leq T'_0$ , which means in particular that  $\underline{y}(T'_0) = 0$ .

Let us now study the solution  $\underline{y}^{\text{lin}}$  of the following heat equation on the semi-infinite space domain  $(-\infty, 1)$  with Dirichlet boundary condition on  $x = 1$ :

$$\begin{cases} \underline{y}_t^{\text{lin}} - \underline{y}_{xx}^{\text{lin}} = u(t) & \text{on } (T'_0, T') \times (-\infty, 1), \\ \lim_{x \rightarrow -\infty} \underline{y}^{\text{lin}}(t, x) = \underline{v}(t) & \text{on } (T'_0, T'), \\ \underline{y}^{\text{lin}}(t, 1) = 0 & \text{on } (T'_0, T'), \\ \underline{y}^{\text{lin}}(0, x) = 0 & \text{on } (-\infty, 1). \end{cases} \quad (\text{VII.16})$$

Using the usual representation formula, we introduce  $\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-x^2} dx$ , and compute

$$\begin{aligned} \underline{y}^{\text{lin}}(t, x_1) &= \int_{T'_0}^t \frac{1}{\sqrt{4\pi(t-s)}} \int_{-\infty}^1 (e^{-\frac{(x_1-x_2)^2}{4(t-s)}} - e^{-\frac{(x_1+x_2-2)^2}{4(t-s)}}) u(s) dx_2 ds \\ &= \int_{T'_0}^t \frac{\theta + 2\|y_0\|_{L^\infty}}{T'} \text{erf}\left(\frac{1-x}{\sqrt{t-s}}\right) ds. \end{aligned} \quad (\text{VII.17})$$

Note that for  $t \in [T'_0, T']$ ,

$$\underline{y}_x^{\text{lin}}(t, x) \leq 0, \quad \text{and} \quad 0 \leq \underline{y}^{\text{lin}}(t, x) \leq \underline{v}(t). \quad (\text{VII.18})$$

Besides, using that erf is an increasing function,

$$(\theta + \|y_0\|_{L^\infty}) \text{erf}\left(\frac{1-x}{\sqrt{T'}}\right) \leq \underline{y}^{\text{lin}}(T', x). \quad (\text{VII.19})$$

Using the upper bound

$$\vartheta(x) \leq \min [M, \vartheta_x(1)(1-x)], \quad x \in [0, 1] \quad (\text{VII.20})$$

combined with the fact that  $\lim_{z \rightarrow \infty} \text{erf}(z) = 1$ , we can find  $T'$  small enough so that

$$\vartheta(x) \leq \underline{y}^{\text{lin}}(T', x), \quad \text{for all } x \in [0, 1]. \quad (\text{VII.21})$$

Going back to the non-linear equation, we extend  $\underline{y}$  to  $(T'_0, T')$  by

$$\begin{cases} \underline{y}_t + (\underline{y}^\gamma)_x - \underline{y}_{xx} = u(t) & \text{on } (T'_0, T') \times (0, 1), \\ \underline{y}(t, 0) = \underline{y}^{\text{lin}}(t, 0) & \text{on } (T'_0, T'), \\ \underline{y}(t, 1) = 0 & \text{on } (T'_0, T'), \\ \underline{y}(0, x) = 0 & \text{on } (0, 1). \end{cases} \quad (\text{VII.22})$$

Note that  $\underline{y}$  is non-negative on  $(T'_0, T') \times (0, 1)$ . We consider  $\delta = \underline{y} - \underline{y}^{\text{lin}}$ , which is solution of

$$\begin{cases} \delta_t - \delta_{xx} = -((\underline{y}^{\text{lin}} + \delta)^\gamma)_x & \text{on } (T'_0, T') \times (0, 1), \\ \delta(t, 0) = 0 & \text{on } (T'_0, T'), \\ \delta(t, 1) = 0 & \text{on } (T'_0, T'), \\ \delta(0, x) = 0 & \text{on } (0, 1). \end{cases} \quad (\text{VII.23})$$

Let us prove that  $\delta \geq 0$  thanks to an energy estimate. We multiply Eq. (VII.23) by  $w = \min(\delta, 0)$ , the negative part of  $\delta$ , and we integrate in space to get

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \|w(t)\|_{L^2}^2 + \|w_x(t)\|_{L^2}^2 &= - \int_0^1 w((\underline{y}^{\text{lin}} + w)^\gamma)_x \\ &= - \int_0^1 w((\underline{y}^{\text{lin}} + w)^\gamma - (\underline{y}^{\text{lin}})^\gamma + (\underline{y}^{\text{lin}})^\gamma)_x \\ &= - \int_0^1 \gamma w \underline{y}_x^{\text{lin}} (\underline{y}^{\text{lin}})^{\gamma-1} + \int_0^1 w_x((\underline{y}^{\text{lin}} + w)^\gamma - (\underline{y}^{\text{lin}})^\gamma). \end{aligned}$$

The first term of the last line is non-positive, hence using Cauchy-Schwartz and Young inequality, we obtain

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \|w(t)\|_{L^2}^2 + \|w_x(t)\|_{L^2}^2 &\leq \|w_x(t)\|_{L^2} \|(\underline{y}^{\text{lin}} + w)^\gamma - (\underline{y}^{\text{lin}})^\gamma\|_{L^2} \\ &\leq \frac{1}{2} \|w_x(t)\|_{L^2}^2 + \frac{\gamma^2}{2} (\theta + \|y_0\|_{L^\infty})^2 \|w\|_{L^2}^2. \end{aligned}$$

By applying Gronwall inequality, we get

$$\|w(t)\|_{L^2}^2 \leq \|w(0)\|_{L^2}^2 e^{\gamma^2(\theta + \|y_0\|_{L^\infty})^2 t}, \quad (\text{VII.24})$$

which implies  $w = 0$ .

Hence, we proved that for  $T'$  small enough,  $y(T') \geq \vartheta$ . Besides,  $\vartheta$  is a subsolution for the controls defined in Eqs. (VII.11) and (VII.12) on  $(T', T)$ . As a consequence, for all  $t \geq T'$ ,  $y(t, x) \geq \vartheta(x)$ , which concludes the proof of the lower bound (VII.9) of Theorem VII.7.



### VII.2.3 Upper bound

We set

$$\bar{v}(t) = \begin{cases} \|y_0\|_{L^\infty} + \frac{(\theta+2\|y_0\|_{L^\infty})t}{T'} & \geq v(t) \quad \text{for } t \leq T', \\ v(t) & \text{for } t > T'. \end{cases} \quad (\text{VII.25})$$

Let us consider the supersolution  $\bar{y}$  on  $(0, T)$ <sup>1</sup>

$$\begin{cases} \bar{y}_t + (\bar{y}^\gamma)_x - \bar{y}_{xx} = u(t) & \text{on } (0, T) \times (0, 1), \\ \bar{y}(t, 0) = \bar{v}(t) & \text{on } (0, T), \\ \bar{y}(t, 1) = \bar{v}(t) & \text{on } (0, T), \\ \bar{y}(0, x) = \|y_0\|_\infty & \text{on } (0, 1). \end{cases} \quad (\text{VII.26})$$

Note that  $\forall t \in [0, T']$ ,  $\bar{y}(t) = \bar{v}(t)$ . Using the comparison principle, we obtain

$$\theta \leq \bar{y}(t) \leq \theta + 3\|y_0\|_{L^\infty} \quad \text{for all } t \geq T'. \quad (\text{VII.27})$$

Let us now denote by  $\delta(t, x)$  the solution of

$$\begin{cases} \delta_t + \gamma\theta^{\gamma-1}\delta_x - \delta_{xx} = -((\theta + \delta)^\gamma - \gamma\theta^{\gamma-1}\delta)_x & \text{on } (\frac{T}{2}, T) \times (0, 1), \\ \delta(t, 0) = 0 & \text{on } (\frac{T}{2}, T), \\ \delta(t, 1) = 0 & \text{on } (\frac{T}{2}, T), \\ \delta(\frac{T}{2}, x) = 3\|y_0\|_{L^\infty} & \text{on } (0, 1). \end{cases} \quad (\text{VII.28})$$

Using the comparison principle, we observe that, for  $t \in [\frac{T}{2}, T]$ ,  $\delta(t, x) \geq \bar{y}(t, x) - \theta$ . To study the evolution of  $\delta$ , we introduce the weight

$$A(x) = e^{\frac{\gamma}{2}\theta^{\gamma-1}(1-x)}. \quad (\text{VII.29})$$

Thus,

$$A_x = -\frac{\gamma}{2}\theta^{\gamma-1}A. \quad (\text{VII.30})$$

We multiply the first line of (VII.28) by  $A\delta$  and we integrate on space. The terms of the left-hand side are

$$\int_0^1 \delta_t A \delta dx = \frac{1}{2} \frac{d}{dt} \|\delta\|_{L^2(A dx)}^2, \quad (\text{VII.31})$$

$$\int_0^1 \gamma\theta^{\gamma-1}\delta_x A \delta dx = \int_0^1 (\gamma\theta^{\gamma-1})^2 \delta^2 A, \quad (\text{VII.32})$$

$$-\int_0^1 \delta_{xx} A \delta dx = \int_0^1 (\delta_x)^2 A dx - \int_0^1 \frac{\gamma}{2}\theta^{\gamma-1}\delta\delta_x A dx. \quad (\text{VII.33})$$

Hence, the left-hand side of Equation (VII.28) becomes

$$\frac{1}{2} \frac{d}{dt} \|\delta\|_{L^2(A dx)}^2 + \frac{\gamma^2\theta^{2(\gamma-1)}}{2} \|\delta\|_{L^2(A dx)}^2 + \|\delta_x\|_{L^2(A dx)}^2. \quad (\text{VII.34})$$

---

1. As  $\bar{v}$  is not in  $H^{1/4}(0, T)$ , we formally define  $\bar{y}$  on both time intervals  $(0, T')$  and  $(T', T)$ .

We multiply the right-hand side of Eq. (VII.28) by  $A\delta$  and integrate in space to get

$$-\int_0^1 A\delta((\theta + \delta)^\gamma - \gamma\theta^{\gamma-1}\delta)_x = \int_0^1 A_x\delta((\theta + \delta)^\gamma - \gamma\theta^{\gamma-1}\delta) \quad (\text{VII.35})$$

$$+ \int_0^1 A\delta_x((\theta + \delta)^\gamma - \gamma\theta^{\gamma-1}\delta). \quad (\text{VII.36})$$

As  $z \mapsto z^\gamma$  is convex,  $(\theta + \delta)^\gamma - \gamma\theta^{\gamma-1}\delta \geq 0$ . As a consequence, the right-hand side integral in line (VII.35) is non-positive. Besides,

$$\int_0^1 A\delta_x((\theta + \delta)^\gamma - \gamma\theta^{\gamma-1}\delta) \leq \frac{1}{2}\|\delta_x\|_{L^2(Adx)}^2 + \frac{1}{2}\|(\theta + \delta)^\gamma - \gamma\theta^{\gamma-1}\delta\|_{L^2(Adx)}^2. \quad (\text{VII.37})$$

Let us estimate the second term of the right-hand side. For  $\gamma \geq 2$ , we can use the fact that the second derivative of  $z \mapsto z^\gamma$  is increasing to get

$$(\theta + \delta)^\gamma - \gamma\theta^{\gamma-1}\delta \leq \frac{\delta^2}{2}\gamma(\gamma-1)(\theta + 3\|y_0\|_{L^\infty})^{\gamma-2}. \quad (\text{VII.38})$$

Whereas, for  $\gamma \leq 2$ , we get

$$(\theta + \delta)^\gamma - \gamma\theta^{\gamma-1}\delta \leq \frac{\delta^2}{2}\gamma(\gamma-1)\theta^{\gamma-2}. \quad (\text{VII.39})$$

Combining Eqs. (VII.34) and (VII.37) to (VII.39), we obtain the weighted energy estimate

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \|\delta(t)\|_{L^2(Adx)}^2 + \frac{1}{2} \|\delta_x\|_{L^2(Adx)}^2 + \frac{\gamma^2 \theta^{2(\gamma-1)}}{2} \|\delta(t)\|_{L^2(Adx)}^2 \\ \leq \gamma^2 (\gamma-1)^2 \tilde{\theta}^{2(\gamma-2)} \|\delta(t)\|_{L^\infty}^2 \|\delta(t)\|_{L^2(Adx)}^2, \end{aligned} \quad (\text{VII.40})$$

where

$$\tilde{\theta} = \begin{cases} \theta + 3\|y_0\|_{L^\infty} & \text{if } 3/2 < \gamma < 2, \\ \theta & \text{if } \gamma \geq 2. \end{cases} \quad (\text{VII.41})$$

We obtain an  $L^\infty$  norm estimate for  $\delta$  with Eq. (VII.27)

$$\|\delta(t)\|_{L^\infty}^2 \leq 9\|y_0\|_{L^\infty}^2. \quad (\text{VII.42})$$

Hence, there exists  $\theta_0$  (depending on  $\|y_0\|_{L^\infty}$  and  $\gamma$ ) which together with Eq. (VII.40) implies that for any  $\theta \geq \theta_0$

$$\frac{1}{2} \frac{d}{dt} \|\delta(t)\|_{L^2(Adx)}^2 + \frac{1}{2} \|\delta_x\|_{L^2(Adx)}^2 \leq -\frac{\gamma^2 \theta^{2(\gamma-1)}}{4} \|\delta(t)\|_{L^2(Adx)}^2. \quad (\text{VII.43})$$

Using the expression of  $A$  (VII.29), we estimate the initial condition

$$\|\delta(\frac{T}{2})\|_{L^2(Adx)}^2 \leq \|\delta(\frac{T}{2})^2 A\|_{L^\infty} \leq 9\|y_0\|_{L^\infty}^2 e^{\frac{3}{2}\theta^{\gamma-1}}. \quad (\text{VII.44})$$

Gronwall inequality applied to Eqs. (VII.43) and (VII.44) gives

$$\|\delta(t)\|_{L^2(Adx)}^2 \leq 9\|y_0\|_{L^\infty}^2 e^{\frac{\gamma}{2}\theta^{\gamma-1}} e^{-\frac{\gamma^2\theta^{2(\gamma-1)}}{4}(t-T')}. \quad (\text{VII.45})$$

Integrating Eq. (VII.43) on  $(\frac{3T}{4}, T)$ , we get

$$\begin{aligned} \|\delta_x\|_{L^2(\frac{3T}{4}, T; L^2(Adx))}^2 &\leq \|\delta(3T/4)\|_{L^2(Adx)}^2 \\ &\leq 9\|y_0\|_{L^\infty}^2 e^{\frac{\gamma}{2}\theta^{\gamma-1}} e^{-\frac{\gamma^2\theta^{2(\gamma-1)}}{4}(T/4)} \xrightarrow{\theta \rightarrow \infty} 0. \end{aligned} \quad (\text{VII.46})$$

Hence, there exists  $t^* \in [\frac{3T}{4}, T]$  such that

$$\|\delta(t^*)\|_{H^1}^2 \leq C\|\delta_x(t^*)\|_{L^2(Adx)}^2 \leq \frac{4C}{T}\|\delta_x\|_{L^2(\frac{3T}{4}, T; L^2(Adx))}^2. \quad (\text{VII.47})$$

Using a classical parabolic estimate on the heat equation (see Section VII.A), we get

$$\|\delta(T)\|_{H_0^1} \leq \|\delta(t^*)\|_{H_0^1} + \|((\theta + \delta)^\gamma - \theta^\gamma)_x\|_{L^2(t^*, T; L^2(dx))}^2 \quad (\text{VII.48})$$

$$\leq \frac{4C}{T}\|\delta_x\|_{L^2(\frac{3T}{4}, T; L^2(Adx))}^2 + \gamma^2(\theta + 3\|y_0\|_{L^\infty})^{2(\gamma-1)}\|\delta\|_{L^2(t^*, T; H_0^1)}^2. \quad (\text{VII.49})$$

Thanks to the exponential decrease in  $\theta$  of Eq. (VII.46), we obtain that

$$\|\delta(T)\|_{H_0^1} \xrightarrow{\theta \rightarrow \infty} 0, \quad (\text{VII.50})$$

which proves Eq. (VII.10) and conclude the proof of Theorem VII.7.

### VII.3 Hyperbolic stage, second part: toward a neighborhood of zero up to a boundary layer

Thanks to the previous part, we reduced our problem to the case where the initial condition  $y_0$  satisfies for some  $\theta, \eta > 0$ ,

$$\vartheta(x) \leq y_0(x) \leq \theta + \eta. \quad (\text{VII.51})$$

In this section we prove that we can steer the solution of the system  $(E_\gamma)$  to a small neighborhood of the null state up to a boundary residue near the right endpoint.

**Lemma VII.9.** *Let  $T > 0$  and  $\theta > 0$ . There exists  $T' \leq T$ ,  $u, v \in L^\infty(0, T') \times (L^\infty(0, T') \cap H^{1/4}(0, T'))$  such that for any  $y_0$  satisfying Eq. (VII.51), we have*

$$\vartheta(x) - \theta - \eta < y(T', x) < \eta. \quad (\text{VII.52})$$

*Proof.* Let us consider the controls  $u(t) = -\frac{\theta}{T'}$  and  $v(t) = \theta(1 - \frac{t}{T'})$  on  $[0, T']$  for some  $T' \leq T$  which will be chosen later. We denote by  $y$  the corresponding solution of Eq.  $(E_\gamma)$  with any initial condition satisfying Eq. (VII.51). Then, we define a subsolution  $\underline{y}$  and a supersolution  $\bar{y}$

by

$$\begin{cases} \underline{y}_t + (|\underline{y}|^\gamma)_x - \underline{y}_{xx} = u(t) & \text{on } (0, T') \times (0, 1), \\ \underline{y}(t, 0) = v(t) & \text{on } (0, T'), \\ \underline{y}(t, 1) = -\frac{\theta t}{T'} & \text{on } (0, T'), \\ \underline{y}(0, x) = \vartheta & \text{on } (0, 1), \end{cases} \quad (\text{VII.53})$$

and

$$\begin{cases} \bar{y}_t + (|\bar{y}|^\gamma)_x - \bar{y}_{xx} = u(t) & \text{on } (0, T') \times (0, 1), \\ \bar{y}(t, 0) = v(t) + \eta & \text{on } (0, T'), \\ \bar{y}(t, 1) = v(t) + \eta & \text{on } (0, T'), \\ \bar{y}(0, x) = \theta + \eta & \text{on } (0, 1). \end{cases} \quad (\text{VII.54})$$

We easily check that  $\bar{y}(t) = v(t) + \eta$ , thus  $\bar{y}(T') = \eta$ . This concludes the proof of the upper bound.

Let us now focus on the subsolution. We define  $\delta(t, x) = \underline{y}(t, x) - \vartheta(x) - \frac{\theta t}{T'}$ . Then,  $\delta$  is solution of

$$\begin{cases} \delta_t - \delta_{xx} = -(|\vartheta + \frac{\theta t}{T'} + \delta|^\gamma)_x & \text{on } (0, T') \times (0, 1), \\ \delta(t, 0) = 0 & \text{on } (0, T'), \\ \delta(t, 1) = 0 & \text{on } (0, T'), \\ \delta(0, x) = 0 & \text{on } (0, 1). \end{cases} \quad (\text{VII.55})$$

The  $H^{-1}(0, 1)$  norm of the source term can be estimated as follows:

$$\begin{aligned} \left\| \left( |\vartheta + \frac{\theta t}{T'} + \delta(t)|^\gamma \right)_x \right\|_{H^{-1}} &= \left\| \left( |\vartheta + \frac{\theta t}{T'} + \delta(t)|^\gamma \right)_x - \left( |\vartheta + \frac{\theta t}{T'}|^\gamma \right)_x + \left( |\vartheta + \frac{\theta t}{T'}|^\gamma \right)_x \right\|_{H^{-1}} \\ &\leq \left\| \left( |\vartheta + \frac{\theta t}{T'} + \delta(t)|^\gamma \right)_x - \left( |\vartheta + \frac{\theta t}{T'}|^\gamma \right)_x \right\|_{L^2} + C(\theta, \gamma) \\ &\leq C(\theta, \gamma) \|\delta(t)\|_{L^2} + C(\theta, \gamma). \end{aligned}$$

Thus, applying a regularity estimate for the heat equation reproduced in Section VII.A,

$$\|\delta(t)\|_{L^2}^2 + \|\delta\|_{L^2(0,t;H_0^1)}^2 \leq \int_0^t C(\theta, \gamma) (\|\delta(s)\|_{L^2}^2 + 1) ds. \quad (\text{VII.56})$$

Thanks to Gronwall inequality,

$$\|\delta(t)\|_{L^2}^2 \leq C(\theta, \gamma) t e^{C(\theta, \gamma)t}. \quad (\text{VII.57})$$

After an integration in time, we obtain

$$\|\delta\|_{L^2(0,t;H_0^1)}^2 \leq \frac{1}{C(\theta, \gamma)} \left( e^{C(\theta, \gamma)t} (C(\theta, \gamma)t - 1) + 1 \right) = O_{t \rightarrow 0}(t^2). \quad (\text{VII.58})$$

Now, using that there exists a constant  $C(\theta, \gamma)$  such that

$$\|(|\vartheta + \frac{\theta t}{T'} + \delta(t)|^\gamma)_x\|_{L^2} \leq C(\theta, \gamma) \|\delta(t)\|_{H^1} + C(\theta, \gamma), \quad (\text{VII.59})$$

together with Section VII.A again, we get

$$\|\delta(t)\|_{H_0^1}^2 \leq C(\theta, \gamma) \|\delta\|_{L^2(0,t;H_0^1)}^2 + C(\theta, \gamma)t. \quad (\text{VII.60})$$

As those estimates are uniform in  $T'$ , we have

$$\|\delta(T')\|_{H_0^1}^2 = O_{T' \rightarrow 0}(T'). \quad (\text{VII.61})$$

This implies that there exists  $T' < T$  such that

$$\|\delta(T')\|_{L^\infty} < \eta. \quad (\text{VII.62})$$

This concludes the proof of Theorem VII.9.  $\square$

## VII.4 Passive stage: dissipation of the boundary residue

In this stage, we start with an initial condition satisfying

$$\vartheta(x) - \theta - \eta < y(T', x) < \eta, \quad x \in [0, 1] \quad (\text{VII.63})$$

and we prove the dissipation of the residue  $\vartheta(x) - \theta$ .

**Proposition VII.10.** *For a given  $T > 0$ , there exist  $C(T)$  and  $\theta_0$  such that for every  $\theta \geq \theta_0$ , every  $\eta > 0$  in a neighborhood of zero and every initial condition  $y_0$  satisfying Eq. (VII.63), the solution of Eq. (E $_\gamma$ ) with null controls satisfies*

$$C(T)\eta \leq y(T) \leq \eta, \quad y(T) \in H_0^1(0, 1). \quad (\text{VII.64})$$

The upper bound is just a consequence of the comparison principle. Let us thus focus on the lower bound. If  $\gamma$  is smaller, the boundary layer is larger and the time needed for the dissipation increases. Fig. VII.2 illustrates this phenomenon. If  $\gamma \leq 3/2$ , our estimates are not sufficient to prove Theorem VII.10. In order to estimate the size of the boundary layer, we use the second point of Theorem VII.6. That is, for  $\frac{1}{2} < \alpha < \gamma - 1$ , there exists  $\tilde{C}$  such that

$$-2\eta - \theta \mathbb{1}_{x \geq 1 - \tilde{C}\theta^{-\alpha}} \leq \vartheta(x) - \theta - \eta. \quad (\text{VII.65})$$

We also set  $\varepsilon_1 = \gamma - 1 - \alpha > 0$  as we will need later to take  $\alpha$  close enough to  $\gamma - 1$ .

We introduce the following subsolution:

$$\begin{cases} \underline{y}_t + (|\underline{y}|^\gamma)_x - \underline{y}_{xx} = 0 & \text{on } (0, T) \times (0, 1), \\ \underline{y}(t, 0) = 0 & \text{on } (0, T), \\ \underline{y}(t, 1) = 0 & \text{on } (0, T), \\ \underline{y}(0) = -2\eta - \theta \mathbb{1}_{x \geq 1 - \tilde{C}\theta^{-\alpha}} & \text{on } (0, 1). \end{cases} \quad (\text{VII.66})$$

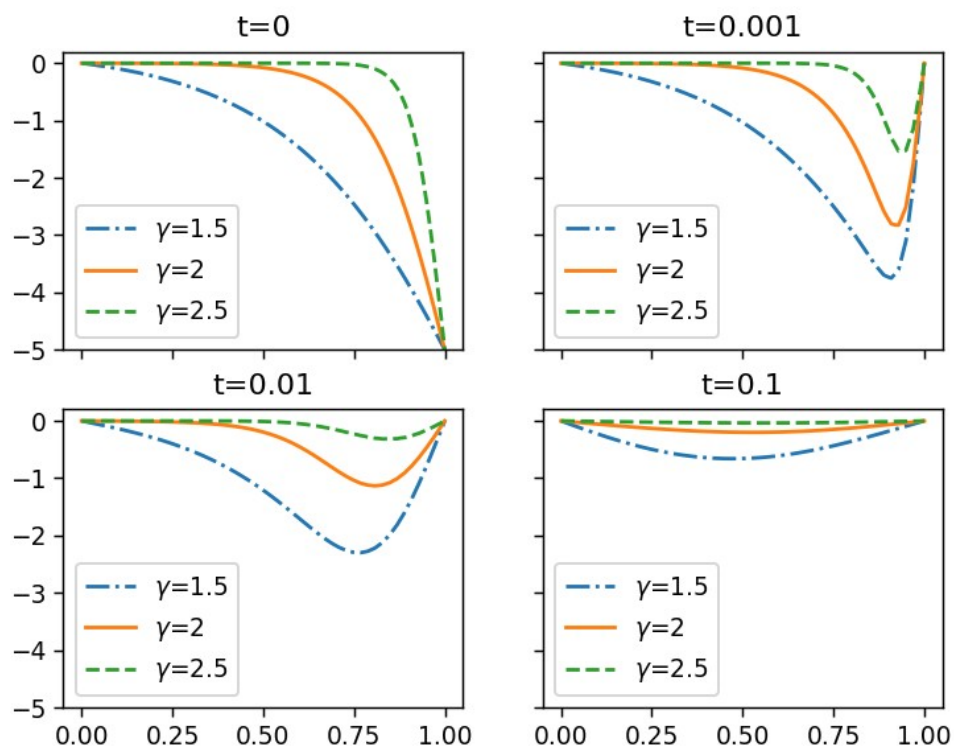
Note that  $\underline{y}$  is a non-positive solution. Besides

$$\|\underline{y}(0)\|_{L^1} = 2\eta + \tilde{C}\theta^{2-\gamma+\varepsilon_1}. \quad (\text{VII.67})$$

Hence, if  $\gamma > 2$ , we have

$$\lim_{\theta \rightarrow \infty} \|\underline{y}(0)\|_{L^1} = 2\eta. \quad (\text{VII.68})$$

We use a smoothing result by Carlen and Loss [CL95, Theorem 1] for strictly convex conservation

Figure VII.2 – Simulation of the dissipation of the boundary layer residue for different values  $\gamma$ .

laws to deduce

$$\|\underline{y}(t)\|_{L^\infty} \leq \frac{\|\underline{y}(0)\|_{L^1}}{\sqrt{4\pi t}}. \quad (\text{VII.69})$$

As a consequence, there exists  $\theta_0$  such that, for any  $\theta \geq \theta_0$ ,

$$\|\underline{y}(T)\|_{L^\infty} \leq \frac{3\eta}{\sqrt{4\pi T}}, \quad (\text{VII.70})$$

which concludes the proof for  $\gamma > 2$ .

In the case  $\gamma < 2$ , the boundary layer grows with  $\theta$ . The smoothing property of Eq. (VII.69) does not exploit the initial localisation near the boundary. Hence, to handle  $\gamma$  small, we use instead a weighted estimate.

We multiply Eq. (VII.66) by  $(x-1)$  and integrate on space to get

$$\frac{d}{dt} \int_0^1 (x-1)\underline{y}(t,x)dx + \int_0^1 |\underline{y}(t,x)|^\gamma dx - \underline{y}_x(0) = 0. \quad (\text{VII.71})$$

Note that  $\int_0^1 (x-1)\underline{y}(t,x)dx$  and  $-\underline{y}_x(0)$  are both non-negative as  $\underline{y} \leq 0$ . Besides,

$$\int_0^1 (x-1)(-2\eta\mathbb{1}_{x \geq 0} - \theta\mathbb{1}_{x \geq 1 - \tilde{C}\theta^{-\alpha}})dx = \frac{\tilde{C}^2\theta^{1-2\alpha}}{2} + \eta. \quad (\text{VII.72})$$

We integrate Eq. (VII.71) on  $(0, \frac{T}{2})$  and use Eq. (VII.72) to get the estimate

$$\|\underline{y}\|_{L^\gamma((0, \frac{T}{2}) \times (0,1))}^\gamma \leq \frac{\tilde{C}^2\theta^{1-2\alpha}}{2} + \eta. \quad (\text{VII.73})$$

Note that  $\theta^{1-2\alpha}$  goes to 0 as  $\theta$  goes to  $\infty$ . This is the case because we assumed  $\gamma > 3/2$ .

Using Eq. (VII.73), there exists  $C$  such that for  $\theta$  large enough, there exist  $t^* \in (0, \frac{T}{2})$  satisfying

$$\|\underline{y}(t^*)\|_{L^1(0,1)} < C\eta. \quad (\text{VII.74})$$

We cannot use the smoothing result from Carlen and Loss as it requires the flux function to be strongly convex (hence  $\gamma \geq 2$ ), but a similar result (without an explicit expression for the constant) is available in [FL99, Lemma 3.1] or [BBS20, Lemma 4.2]. Finally, we get

$$\|\underline{y}(T)\|_{L^\infty} \leq \frac{C}{\sqrt{T-t^*}} \|\underline{y}(t^*)\|_{L^1} \leq C'\eta. \quad (\text{VII.75})$$

This concludes the proof of Theorem VII.10

## VII.5 Parabolic stage: local null exact controllability

The local null exact controllability of semi-linear parabolic equations has been established for a wide variety of cases. For completeness, we provide a sketch of proof based on [FG07, Lemma 2].

**Lemma VII.11.** *There exists  $C^* > 0$  such that for every  $y_0 \in H_0^1$  satisfying  $\|y_0\|_{L^\infty} < \frac{1}{2}$  and*

$\|y_0\|_{L^2} \leq \frac{1}{2T} e^{-C^*/T}$ , there exists a control  $v \in H^{1/4}(0, T) \cap L^\infty(0, T)$  such that the solution of  $(E_\gamma)$  with controls  $u = 0$  and  $v$  satisfies  $y(T) = 0$ .

*Sketch of the proof.* First, we transform the boundary control problem into an internal control one. Namely, we consider the wider space domain  $[-1, 1]$  and a domain of the internal control  $\omega \subset (-1, 0)$  with non-empty interior. We denote  $Q = (0, T) \times [-1, 1]$ . Let  $w$  be an internal control acting on  $(0, T) \times \omega$ . We are interested in the local null exact controllability of

$$\begin{cases} y_t + (|y|^\gamma)_x - y_{xx} = w(t, x)\mathbb{1}_{(0, T) \times \omega} & \text{on } (0, T) \times (-1, 1), \\ y(t, 0) = 0 & \text{on } (0, T), \\ y(t, 1) = 0 & \text{on } (0, T), \\ y(0, x) = \tilde{y}_0(x) & \text{on } (-1, 1), \end{cases} \quad (\text{VII.76})$$

where  $\tilde{y}_0 \in H_0^1(0, 1)$  is the extension of  $y_0$  by 0 on  $(-1, 1)$ .

We set  $s \in (0, 1)$  and introduce the closed convex set

$$K = \{z \in H^s(Q) \mid \|z\|_{L^\infty(Q)} \leq 1\} \quad (\text{VII.77})$$

and the following set of admissible controls

$$\mathcal{A}_0 = \{w \in L^\infty((0, T) \times \omega) \mid \|w\|_{L^\infty(Q)} \leq \|y_0\|_{L^2} e^{C^*/T}\}, \quad (\text{VII.78})$$

where  $C^*$  will be defined later. Let us define  $\varphi$  for  $z \in K$  by

$$\varphi(z) = \text{sign}(z)|z|^{\gamma-1} \quad (\text{VII.79})$$

Note that  $\|\varphi(z)\|_{L^\infty} \leq 1$ . Let  $\mathcal{A} : H^s(Q) \rightarrow H^s(Q)$  be the set-value mapping associating with  $z \in H^s(Q)$  the set of solutions  $y$  of the linear equation

$$\begin{cases} y_t + \varphi(z)y_x - y_{xx} = w(t, x) & \text{on } (0, T) \times (-1, 1), \\ y(t, 0) = 0 & \text{on } (0, T), \\ y(t, 1) = 0 & \text{on } (0, T), \\ y(0, x) = \tilde{y}_0(x) & \text{on } (-1, 1), \end{cases} \quad (\text{VII.80})$$

with  $w \in \mathcal{A}_0$  and satisfying  $y(T) = 0$ .

According to [FG07; FZ00], there exists  $C^* > 0$  such that for any function  $\varphi(z)$  with  $\|\varphi(z)\|_{L^\infty} \leq 1$ , there exists  $w \in \mathcal{A}_0$  steering the solution of Eq. (VII.80) to zero, i.e.,  $\mathcal{A}(z)$  is not empty.

Let us now assume that  $y_0$  satisfies the assumption of Theorem VII.11 with  $C^*$ . We want to check that  $\mathcal{A}$  satisfies the hypotheses of Kakutani's fixed point theorem ([GD03, Chapter 2]).

Using [LSU68, Chapter 3], the solutions  $y$  of Eq. (VII.80) belong to

$$X = L^2(0, T; H^2(-1, 1)) \cap H^1(0, T; L^2(-1, 1))$$

and thus to  $H^1(Q)$ . Moreover, the maximum principle implies

$$\|y\|_{L^\infty(Q)} \leq \|y_0\|_{L^\infty(Q)} + T\|w\|_{L^\infty(Q)} \leq 1. \quad (\text{VII.81})$$

Hence,  $\mathcal{A}$  maps  $K$  into  $K$  and, for any  $z \in K$ ,  $\mathcal{A}(z)$  is a non-empty convex compact subset of  $H^s(Q)$ .

Let us now check that  $\mathcal{A}$  is upper hemicontinuous on  $K$ . In pursuite of this goal, we set



$\mu \in (H^s(Q))'$ . We have to check that

$$z \mapsto \sup_{y \in \mathcal{A}(z)} \langle \mu, y \rangle \tag{VII.82}$$

is upper semi-continuous.

Let  $(z_n)_n \in K^{\mathbb{N}}$  be a converging sequence toward  $z_\infty$  in  $H^s(Q)$ . Let us first observe that  $\varphi : K \rightarrow L^2(Q)$  is continuous. Indeed, for  $\gamma \geq 2$  we use the uniform boundedness of  $K$

$$\|\varphi(z_a) - \varphi(z_b)\|_{L^2(Q)} \leq \|z_a - z_b\|_{L^2(Q)}.$$

On the other-hand, for  $\gamma < 2$ ,  $x \mapsto \varphi(x)$  is  $\gamma - 1$  Hölder continuous:

$$\|\varphi(z_a) - \varphi(z_b)\|_{L^2(Q)} \leq \|(z_a - z_b)^{\gamma-1}\|_{L^2(Q)} \leq (2T)^{2-\gamma} \|z_a - z_b\|_{L^2(Q)}^{\gamma-1}.$$

Hence,  $\varphi(z_n) \rightarrow \varphi(z_\infty)$  in  $L^2(Q)$ .

By compactness of  $\mathcal{A}(z_n)$ , there exists  $y_n \in \mathcal{A}(z_n)$  such that

$$\sup_{y \in \mathcal{A}(z_n)} \langle \mu, y \rangle = \langle \mu, y_n \rangle. \tag{VII.83}$$

Using energy estimates ([LSU68, Chapter 3]),  $(y_n)_n$  is uniformly bounded in  $X$ . Thanks to Aubin-Lions Lemma [Aub63], up to a subsequence, there exists  $y_\infty \in X$  such that

$$y_{n,t} \rightharpoonup y_{\infty,t} \quad \text{weakly } L^2(Q), \tag{VII.84}$$

$$y_n \rightarrow y_\infty \quad \begin{cases} \text{weakly in } L^2(0, T; H^2(-1, 1)), \\ \text{strongly in } L^2(0, T; H_0^1(-1, 1)), \\ \text{in } C^0(0, T; L^2(-1, 1)). \end{cases} \tag{VII.85}$$

As  $\|\varphi(z_n)\|_{L^\infty(Q)} \leq 1$ , up to subsequence, there exists  $g \in L^\infty(Q)$  such that we have the weak star convergence  $\varphi(z_n) \xrightarrow{*} g$ . By uniqueness of the limit in the sense of distribution,  $g = \varphi(z)$ . Thus,  $\varphi(z_n)y_n \rightharpoonup \varphi(z_\infty)y_{\infty,x}$  in  $L^2(Q)$ .

As a consequence,  $y_\infty$  is solution of Eq. (VII.80) with  $\varphi(z_\infty)$  as coefficient. This shows that  $y_\infty \in \mathcal{A}(z_\infty)$ , i.e.,

$$\limsup_{n \rightarrow \infty} \sup_{y \in \mathcal{A}(z_n)} \langle \mu, y \rangle \leq \sup_{y \in \mathcal{A}(z_\infty)} \langle \mu, y \rangle. \tag{VII.86}$$

Hence, by Kakutani's fixed point theorem, there exists  $\hat{z} \in K$  such that  $\hat{z} \in \mathcal{A}(\hat{z})$ .

Let us now take the trace of  $\hat{z}$  on  $(0, T) \times \{0\}$ . As  $\hat{z} \in X$ , by [LM72, Theorem 2.1],  $\hat{z}(\cdot, 0) \in H^{3/4}(0, T)$ . This ensures that the restriction of  $\hat{z}$  to  $(0, T) \times (0, 1)$  is an admissible trajectory of the system  $(E_\gamma)$  and proves Theorem VII.11.  $\square$

## VII.6 Open problems

Let us now present some open problems in one space dimension related to this work.

Does the small-time global null controllability hold for Eq.  $(E_\gamma)$  with  $1 < \gamma \leq 3/2$ ? Our entire control strategy is based on using the hyperbolicity of the evolution equation to dissipate the initial condition at the cost of a boundary layer. Then, for this non-small boundary layer to disappear, we use the fact that the moment  $\int (1-x)|y|$  is small. A generalisation of this idea was

used in [CMS20] where the preparation of the dissipation of boundary layer plays a crucial role. In our case, the return method with  $\vartheta$  ensures the dissipation in the case  $\gamma > 3/2$  (the limiting step is Section VII.4). In the linear case  $\gamma = 1$ , there is no boundary layer. As a consequence, we believe that our method cannot be extended to  $[1, 3/2]$ . Another approach would be to use some highly oscillating controls to ensure a better preparation of the boundary residue.

We can also ask if the small-time global null controllability holds for more general flux functions ( $f(u)_x$  instead of  $(|u|^\gamma)_x$ ). The extension of our proof to strictly convex viscous conservation laws should be possible. For more general functions, a precise study of the solutions of Eq. (VII.1) would be necessary.

Another interesting direction would be to consider a dispersive model like the Korteweg–De Vries equation. With the help of two boundary controls and a uniform in space internal control, Chapouly proved in [Cha09a] that the small-time global null controllability holds. To the best of the author knowledge, whether the small-time global null controllability holds without the use of the right Dirichlet boundary control (and possibly with the help of a right control on the derivative) remains an open question.

## VII.A Parabolic regularity estimates for the heat equation

We recall here a well-known result on the regularity of the heat equation reproduced for example in [Léa12, Appendix 4.1].

Let  $\mathcal{H}^m = D((-\Delta)^{m/2})$ ,  $m \geq 0$ , be the domain of the fractional Dirichlet Laplacian on  $L^2(0, 1)$ , and  $\mathcal{H}^{-m}$  the dual of  $\mathcal{H}^m$  with pivot space  $L^2(0, 1)$ . In particular  $\mathcal{H}^1 = H_0^1(0, 1)$ ,  $\mathcal{H}^0 = L^2(0, 1)$  and  $\mathcal{H}^{-1} = H^{-1}(0, 1)$ . Let  $y$  be a classical solution of

$$\begin{cases} y_t - y_{xx} = f(t, x) & \text{on } (0, T) \times (0, 1), \\ y(t, 0) = 0 & \text{on } (0, T), \\ y(t, 1) = 0 & \text{on } (0, T), \\ y(0, \cdot) = y_0 & \text{on } (0, 1), \end{cases} \quad (\text{VII.87})$$

with  $m \in \mathbb{R}$ ,  $u_0 \in \mathcal{H}^m(0, 1)$  and  $f \in L^2(0, T; \mathcal{H}^{m-1}(0, 1))$ . Then,

$$y \in \mathcal{C}^0(0, T; \mathcal{H}^m(0, 1)) \cap L^2(0, T; \mathcal{H}^{m+1}(0, 1)) \cap H^1(0, T; \mathcal{H}^{m-1}(0, 1)), \quad (\text{VII.88})$$

and, for  $t \leq T$ ,

$$\|y(t)\|_{\mathcal{H}^m}^2 + \int_0^t \|y(s)\|_{\mathcal{H}^{m+1}}^2 ds + \int_0^t \|y_t(s)\|_{\mathcal{H}^{m-1}}^2 ds = \|y_0\|_{\mathcal{H}^m}^2 + \int_0^t \|f(s)\|_{\mathcal{H}^{m-1}}^2 ds. \quad (\text{VII.89})$$



# List of Figures

I.1	Simulation de la trajectoire d'une particule dans un champ magnétique axisymétrique	4
I.2	Simulation de la trajectoire d'une particule dans un champ magnétique avec une composante poloïdale.	5
I.3	Schéma d'un tokamak, la torsion des lignes de champ magnétique est produite grâce à un courant électrique circulant à l'intérieur le plasma.	5
I.4	Schéma de Wendelstein 7-X, stellarator du Max-Planck Institut für Plasmaphysik achevé en 2015. Les bobines sont en bleue, le plasma en jaune et une ligne de champ magnétique est représentée en vert. L'image provient du Max-Planck Institut für Plasmaphysik	6
I.5	Courant surfacique optimal pour la CWS utilisée pour la conception de NCSX. La CWS présente une symétrie discrète sous la rotation d'angle $\frac{2\pi}{3}$ selon l'axe $Oz$ .	11
I.6	Courant surfacique optimal sur une surface optimisée par <b>Stellacode</b> . L'échelle montre une réduction d'un facteur trois de l'intensité maximale du courant avec la situation initiale.	11
I.7	Une CWS obtenue numériquement dans une optimisation ne pénalisant pas la courbure.	12
I.8	Implications entre les différents types de contrôlabilité d'ensemble.	24
I.9	Reproduction de la Figure V.1. Un exemple de chemin adiabatique (en bleu) permettant l'inversion robuste de populations pour $\alpha \in [\alpha_0, \alpha_1]$	26
I.10	Un exemple de contrôle complexe de type <i>chirp</i> . L'amplitude augmente puis diminue alors que la fréquence baisse au cours du temps.	27
II.1	CWS (blue and white) and plasma surface (orange) of the National Compact Stellarator Experiment (NCSX) designed by the Princeton Plasma Physics Laboratory. There is a three-folds discrete symmetry in the design.	41
II.2	This figure illustrates the difference between $U_r(S_\infty)$ filled in grey, $U_r(S_n)$ (resp., $A_r(S_n)$ ) delimited by the blue dashed (resp., purple dotted) curves. The black arrows represent the field $\nabla b_{V_\infty}$ and the red ones represent $\nabla b_{V_n}$ . Note that both $V_n$ and $V_\infty$ are on the right of the figure.	49
II.3	$p_{S_n}(x)$ is obtained by taking the intersection of the flow of $\nabla b_{V_\infty}$ and $S_n$ . Whereas the standard projector (in the sense of shortest distance) on $S_n$ is obtained by using the flow of $\nabla b_{V_n}$ .	49
II.4	Main pattern of the optimal CWS for the DP simulation with $\lambda = 2.5e^{-16}$ , top and bottom spikes have been truncated.	66
II.5	Main pattern of the CWS (blue and white) for the DPR simulation with $\lambda = 2.5e^{-16}$ .	66
II.6	Main pattern of the optimal CWS (blue and white) for the DPR simulation with $\lambda = 2.5e^{-16}$ .	66

II.7	History of convergence for the implemented optimization algorithm. From left to right, evolution of the costs (left), distance and perimeter constraints (middle) and the curvature constraint (right) along the optimization process. From top to bottom: Table II.1 ( $\lambda = 2.5e^{-16}$ ) configurations DPR and DP, then table II.2 ( $\lambda = 5e^{-19}$ ) configurations DPR and DP. . . . .	67
IV.1	Average norm of $\mathbf{B}$ as a function of the distance $\varepsilon$ from the surface $S$ , for two different grids, more coarse (red) or fine (blue). $h_{min}$ and $h_{max}$ refer to the smallest and largest mesh size (the poloidal $\times$ toroidal mesh being non-uniform in real space). The plot guides the selection of $\varepsilon$ : an excessively small value, $\varepsilon \lesssim h$ , results in the numerical artifact of a diverging field. . . . .	97
IV.2	Plot of the local cost $f_e$ as a function of the local force $w$ . Note that $f_e$ diverges at $c_1$ and vanishes in $[0, c_0]$ . In other words, the force is non-linearly optimized: small values are permitted, intermediate ones are increasingly, non-linearly penalized, and large ones are forbidden. . . . .	100
IV.3	Convergence of $\mathbf{L}_\varepsilon$ toward $\mathbf{L}$ for NCSX, for two different grids. Convergence stops when $\varepsilon \lesssim h$ , due to a numerical error in $\mathbf{B}$ (Fig. IV.1). . . . .	101
IV.4	NCSX LI383 plasma surface in orange-yellow and CWS in green-white, for a half period. The triangular mesh is only used for rendering; the actual calculations were carried out on a $64 \times 64$ , poloidal $\times$ toroidal mesh. . . . .	102
IV.5	(a) Trade-off between plasma shape accuracy and Laplace force metrics (as defined in Eq. IV.26) for different weightings in Eq. IV.25 and different numbers of harmonics, and thus of Degrees of Freedom (DOF). Such trade-off, expected when optimizing a linear combination of $\chi_B^2$ and $\chi_F^2$ (symbols), is also observed in the minimization of $\chi_B^2$ and $\chi_j^2$ (curves). (b) Similar trade-off between maximum field and maximum Laplace force (Eq. IV.27). . . . .	106
IV.6	Trade-offs between: (a) plasma shape accuracy and Laplace force metrics (Eq. IV.26) and (b) maximum field and maximum Laplace force (Eq. IV.27). Unlike Fig. IV.5, all simulations here used 624 DOF. Circle symbols correspond to the four cases discussed in Sec. IV.4 and presented in Fig. IV.7. As expected, the $C_e$ cases (red and blue) fall between the penalized and forbidden forces $c_0$ and $c_1$ defined in Fig. IV.2, marked here by vertical dotted lines. . . . .	107

IV.7	Results of minimizing Eq. IV.25 for NCSX, for four different cases (four different choices of weights in the equation, as summarized by Table IV.28. Each column refers to a different case; its title is color-coded like the corresponding data-point in Fig. IV.6. From top to bottom the three rows refer respectively to the results for simultaneous (1) current regularization (if any), (2) field accuracy and (3) force-minimization (if any). Case 4 (last column) demonstrates that it is possible to simultaneously optimize these three competing objectives without excessively penalizing any of them with respect to established codes. On the contrary, case 4 actually exhibits higher field accuracy and lower peak forces compared to REG-COIL (first column). Shown in the legends are the Root Mean Square (RMS) surface-averages and local maxima of the quantities plotted, as well as the $H^1$ norm of $\mathbf{j}$ and $C_e$ force metric (Eq. IV.27). The quantities actually minimized are marked in purple. 624 DOF are used for $\mathbf{j}$ in every simulation. It is well-known from [Lan17] and Fig. IV.5 that a higher number of DOF will lower all individual metrics $\chi_B^2$ , $\chi_j^2$ , $\chi_{\nabla j}^2$ , $\chi_F^2$ and find better compromises among them. Correspondingly, all contours presented here will improve, for all 4 cases, and by the same proportion. However, this will require more computational resources, and is left as future work. . . . .	114
IV.8	Tangential and normal components of the Laplace forces of the simulations in Fig. IV.7. . . . .	115
IV.9	The Laplace forces from the last column of Fig. IV.7. The unit for the pressure is Pascal. The triangular mesh is only used for rendering; the actual calculations were carried out on a $64 \times 64$ , poloidal $\times$ toroidal mesh. . . . .	116
IV.10	Examples of current filamentation for case 1 and 4. . . . .	116
V.1	An adiabatic path as the one applied in Proposition V.2. . . . .	123
V.2	Comparison of the real-valued and complex-valued chirp scheme of the first point of Remark V.5 with $E = 0.75$ , $\alpha = 0.25$ , $\varepsilon_1 = 1$ , $v_0 = -0.5$ , $v_1 = 0.5$ . Notice that the assumptions of Theorem V.3 are not satisfied. . . . .	124
V.3	$E = 0.75$ , $\alpha = 0.25$ , $\varepsilon_1 = 1$ , $\varepsilon_2 = 0.1$ , $v_0 = -0.5$ , $v_1 = 0.5$ . Assumption $4E + 3\Delta'(s) - 2\alpha < 0$ is satisfied if and only if $\alpha < 0$ . . . . .	126
V.4	Taking $v_0 = -0.5$ , $v_1 = 0.5$ , $E = 0$ , and $\alpha = 0$ , we observe that the fidelity does not converge to 1 as $(\varepsilon_1, \varepsilon_2) \rightarrow 0$ in the regime $\varepsilon_1 \ll \varepsilon_2$ . The plot corresponds to the choice $\varepsilon_1 = \varepsilon_2^2$ . . . . .	131
V.5	Eigendirection corresponding to the negative eigenvalue of $\tilde{H}_{\text{slow}}$ as a function of $(u, \Delta') \in \mathbb{R}^2$ , for $\varepsilon_1 = 0.01$ and $\alpha = 0$ . . . . .	139
V.6	Eigendirection corresponding to the negative eigenvalue of $\tilde{H}_{\text{slow}}$ as a function of $(u, \Delta') \in \mathbb{R}^2$ , for $\varepsilon_1 = 1$ and $\alpha = 0$ . . . . .	139
V.7	Log of the distance from $\psi_{\varepsilon_1, \varepsilon_2}^0(\frac{1}{\varepsilon_1 \varepsilon_2})$ to the orbit of $(1, 0)$ . . . . .	141
V.8	Population transfer as a function of $\alpha$ for $E = 1$ , $\varepsilon_1 = 0.5$ and $\varepsilon_2 = 0.1$ . . . . .	142
V.9	$\varepsilon_1 = 0.5$ , $\varepsilon_2 = 0.1$ and $\alpha = 0$ . In thick line are the trajectories corresponding to the equivalent 1st order RWA system. . . . .	142
V.10	$\varepsilon_1 \varepsilon_2 = 0.05$ , $\alpha = 0$ . In thick line are the trajectories corresponding to the equivalent 1st order RWA system and in dotted line the theoretical AA trajectories. . . . .	143

VI.1	Schematic description of the time scale invariance of the optimal control in the chattering process. The optimal solution of the quantum control problem described in this chapter is plotted in the top panel. Near the final time $t_f$ , the control switches infinitely many times with an asymptotic invariant structure by time dilation as represented on the two lower panels (successive zooms around $t = t_f$ given by the different plot colors). . . . .	146
VI.2	Plot of the optimal trajectory from $\mathbf{X}(0) = (0, 1, 0)$ to $\mathbf{X}(t_f) =  3\rangle = (0, 0, 1)$ on the sphere $x_1^2 + x_2^2 + x_3^2 = 1$ . The switching curves are plotted in red. The parameter $\Delta$ is set to 10. . . . .	150
VI.3	Time evolution of the control $u(t)$ (bold black line) and of the switching function $\Phi(t)$ (blue line) for the optimal trajectory of Fig. VI.2. The control switches from $\pm 1$ to $\mp 1$ when the switching function changes sign. . . . .	150
VI.4	Switching curves corresponding to three different values of $\Delta$ . This parameter is set respectively to 6, 10 and 14 for . . . . .	151
VI.5	Comparison between the optimal solution obtained by our numerical integration of the PMP (bottom right panel) and a numerical control designed by a direct approach for different time steps $N$ . The control time is fixed to 2.59 in the numerical optimization. . . . .	152
VI.6	Schematic representation of the quantum system with the coupling $\Delta$ and $u(t)$ between the states $ 1\rangle$ and $ 2\rangle$ and $ 2\rangle$ and $ 3\rangle$ (red arrows). The black and blue dots indicate respectively the initial and the target states. The black arrow represents the relaxation process. . . . .	153
VI.7	(top)Optimal trajectory for the Fuller model (black solid line). The switching curves are plotted in red. (bottom) Zoom of the top panel near the origin . . . .	156
VI.8	Plot onto the $(x_1, x_2)$ - plane of the optimal trajectory of the quantum control problem(solid black line). The dashed line depicts the solution of the linear approximation. The red and blue curves represent respectively the switching curves for the quantum and Fuller systems. The bottom panel is a zoom of the top one. . . . .	157
VI.9	Evolution of the distance to the target (top) and of the cost $\mathcal{C}$ (bottom) as a function of the number of time steps (crosses). In the bottom panel, the horizontal solid line (in red) indicates the efficiency of the optimal solution. The solid black line is just to guide the lecture in the two panels. . . . .	161
VII.1	Steady states $\vartheta$ with $\theta = 2$ and $\theta = 5$ for different values of $\gamma$ . . . . .	168
VII.2	Simulation of the dissipation of the boundary layer residue for different values $\gamma$ . . . . .	177

# Bibliography

- [Aam+19] E. Aamari, J. Kim, F. Chazal, B. Michel, A. Rinaldo, and L. Wasserman. “Estimating the reach of a manifold”. In: *Electronic Journal of Statistics* 13.1 (2019).
- [AMR88] R. Abraham, J. E. Marsden, and T. Ratiu. *Manifolds, Tensor Analysis, and Applications*. 2nd ed. Applied Mathematical Sciences. New York: Springer-Verlag, 1988.
- [AB05] R. Adami and U. Boscain. “Controllability of the Schrödinger equation via intersection of eigenvalues”. In: *Proceedings of the 44th IEEE conference on decision and control*. 2005.
- [ABS16] A. Agrachev, Y. Baryshnikov, and A. Sarychev. “Ensemble controllability by Lie algebraic methods”. In: *ESAIM. Control, Optimisation and Calculus of Variations* 22.4 (2016).
- [AS04] A. A. Agrachev and Y. L. Sachkov. *Control theory from the geometric viewpoint*. Vol. 87. Encyclopaedia of mathematical sciences. Berlin: Springer-Verlag, 2004.
- [Alo+18] A. Alonso-Rodríguez, J. Camaño, R. Rodríguez, A. Valli, and P. Venegas. “Finite element approximation of the spectrum of the curl operator in a multiply connected domain”. In: *Foundations of Computational Mathematics* 18.6 (2018).
- [Ash+07] S. Ashhab, J. R. Johansson, A. M. Zagoskin, and F. Nori. “Two-level systems driven by large-amplitude fields”. In: *Physical Review A: Atomic, Molecular, and Optical Physics* 75.6 (2007).
- [Aub63] J.-P. Aubin. “Un théorème de compacité”. In: *Comptes Rendus Hebdomadaires des Séances de l’Académie des Sciences* 256 (1963).
- [ABS19] N. Augier, U. Boscain, and M. Sigalotti. “On the compatibility between the adiabatic and the rotating wave approximations in quantum control”. In: *Proceedings of the 58th IEEE conference on decision and control*. 2019.
- [ABS18] N. Augier, U. Boscain, and M. Sigalotti. “Adiabatic ensemble control of a continuum of quantum systems”. In: *SIAM Journal on Control and Optimization* 56.6 (2018).
- [ABS20] N. Augier, U. Boscain, and M. Sigalotti. “Semi-conical eigenvalue intersections and the ensemble controllability problem for quantum systems”. In: *Mathematical Control and Related Fields* 10.4 (2020).
- [ABS22] N. Augier, U. Boscain, and M. Sigalotti. “Effective adiabatic control of a decoupled Hamiltonian obtained by rotating wave approximation”. In: *Automatica* 136 (2022).
- [BMS82] J. M. Ball, J. E. Marsden, and M. Slemrod. “Controllability for distributed bilinear systems”. In: *SIAM Journal on Control and Optimization* 20.4 (1982).
- [BBS20] M. Bank, M. Ben-Artzi, and M. Schonbek. “Viscous conservation laws in 1D with measure initial data”. In: *Quarterly of Applied Mathematics* 79.1 (2020).



- [BCR10a] K. Beauchard, J.-M. Coron, and P. Rouchon. “Controllability issues for continuous-spectrum systems and ensemble controllability of Bloch equations”. In: *Communications in Mathematical Physics* 296.2 (2010).
- [Bek96] D. Bekiranov. “The initial-value problem for the generalized Burgers’ equation”. In: *Differential and Integral Equations* 9.6 (1996).
- [Ber74] L. D. Berkovitz. “Lower semicontinuity of integral functionals”. In: *Transactions of the American Mathematical Society* 192 (1974).
- [BGM11] F. Bonnans, V. Grelard, and P. Martinon. *Bocop, the optimal control solver, open source toolbox for optimal control problems*. 2011.
- [BS12] B. Bonnard and D. Sugny. *Optimal control with applications in space and quantum dynamics*. AIMS series on applied mathematics. American Institute of Mathematical Sciences, 2012.
- [Bor00] V. F. Borisov. “Fuller’s phenomenon: Review”. In: *Journal of Mathematical Sciences* 100 (2000).
- [Bos+12a] U. Boscain, M. Caponigro, T. Chambrion, and M. Sigalotti. “A weak spectral condition for the controllability of the bilinear Schrödinger equation with application to the control of a rotating planar molecule”. In: *Communications in Mathematical Physics* 311.2 (2012).
- [BC03] U. Boscain and Y. Chitour. “Time optimal synthesis for a  $so(3)$ -left-invariant control system on a sphere”. In: *42nd IEEE International Conference on Decision and Control (IEEE Cat. No.03CH37475)*. 42nd IEEE International Conference on Decision and Control (IEEE Cat. No.03CH37475). Vol. 3. 2003.
- [BSS21] U. Boscain, M. Sigalotti, and D. Sugny. “Introduction to the Pontryagin maximum principle for quantum optimal control”. In: *PRX Quantum* 2.3 (2021).
- [Bos+12b] U. Boscain, F. Chittaro, P. Mason, and M. Sigalotti. “Adiabatic control of the Schrödinger equation via conical intersections of the eigenvalues”. In: *Institute of Electrical and Electronics Engineers. Transactions on Automatic Control* 57.8 (2012).
- [Bos+15] U. Boscain, J.-P. Gauthier, F. Rossi, and M. Sigalotti. “Approximate controllability, exact controllability, and conical eigenvalue intersections for quantum mechanical systems”. In: *Communications in Mathematical Physics* 333.3 (2015).
- [BM06] U. Boscain and P. Mason. “Time minimal trajectories for a spin  $1/2$  particle in a magnetic field”. In: *Journal of Mathematical Physics* 47.6 (2006).
- [BP04] U. Boscain and B. Piccoli. *Optimal Syntheses for Control Systems on 2-D Manifolds*. Vol. 43. Mathématiques & Applications (Berlin). Springer-Verlag, Berlin, 2004.
- [BPS21] U. Boscain, E. Pozzoli, and M. Sigalotti. “Classical and quantum controllability of a rotating symmetric molecule”. In: *SIAM Journal on Control and Optimization* 59.1 (2021).
- [BCC19] N. Boussaïd, M. Caponigro, and T. Chambrion. “On the Ball-Marsden-Slemrod obstruction for bilinear control systems”. In: *2019 IEEE 58th conference on decision and control (CDC)*. 2019.
- [Bre11] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011.
- [BCR10b] C. Brif, R. Chakrabarti, and H. Rabitz. “Control of quantum phenomena: past, present and future”. In: *New Journal of Physics* 12.7 (2010).

- [BHH75] A. E. Bryson, Y.-C. Ho, and Y. C. Ho. *Applied Optimal Control: Optimization, Estimation, and Control*. Hemisphere Publishing Corporation, 1975. 504 pp.
- [Bur48] J. M. Burgers. “A Mathematical Model Illustrating the Theory of Turbulence”. In: *Advances in Applied Mechanics*. Ed. by R. Von Mises and T. Von Kármán. Vol. 1. Elsevier, 1948.
- [CDG01] J. Cantarella, D. DeTurck, and H. Gluck. “The Biot–Savart operator for application to knot theory, fluid dynamics, and plasma physics”. In: *Journal of Mathematical Physics* 42.2 (2001).
- [CDG02] J. Cantarella, D. DeTurck, and H. Gluck. “Vector calculus and the topology of domains in 3-Space”. In: *The American Mathematical Monthly* 109.5 (2002).
- [Cao+10] X. Cao, J. Q. You, H. Zheng, A. G. Kofman, and F. Nori. “Dynamics and quantum Zeno effect for a qubit in either a low- or high-frequency bath beyond the rotating-wave approximation”. In: *Physical Review A: Atomic, Molecular, and Optical Physics* 82.2 (2010).
- [Cap+18] M. Caponigro, R. Ghezzi, B. Piccoli, and E. Trélat. “Regularization of chattering phenomena via bounded variation controls”. In: *IEEE Transactions on Automatic Control* 63.7 (2018).
- [CL95] E. A. Carlen and M. Loss. “Optimal smoothing and decay estimates for viscously damped conservation laws, with applications to the 2-D Navier-Stokes equation”. In: *Duke Mathematical Journal* 81.1 (1995).
- [Cha09a] M. Chapouly. “Global controllability of a nonlinear Korteweg-De Vries equation”. In: *Communications in Contemporary Mathematics* 11.03 (2009).
- [Cha09b] M. Chapouly. “Global controllability of nonviscous and viscous Burgers-type equations”. In: *SIAM Journal on Control and Optimization* 48.3 (2009).
- [Cha09c] M. Chapouly. “On the global null controllability of a Navier–Stokes system with Navier slip boundary conditions”. In: *Journal of Differential Equations* 247.7 (2009).
- [Che75] D. Chenaïs. “On the existence of a solution in a domain identification problem”. In: *Journal of Mathematical Analysis and Applications* 52.2 (1975).
- [Col51] J. D. Cole. “On a quasi-linear parabolic equation occurring in aerodynamics”. In: *Quarterly of Applied Mathematics* 9.3 (1951).
- [Cop+13] B. Coppi et al. “New developments, plasma physics regimes and issues for the Ignitor experiment”. In: *Nuclear Fusion* 53.10 (2013).
- [Cor93] J. Coron. “Contrôlabilité exacte frontière de l’équation d’Euler des fluides parfaits incompressibles bidimensionnels”. In: *c. R. Acad. Sci. Paris* 317 (1993).
- [CF96] J. Coron and A. Fursikov. “Global exact controllability of the 2D Navier-Stokes equations on a manifold without boundary”. In: *Russian Journal of Mathematical Physics* 4 (1996).
- [Cor92] J.-M. Coron. “Global asymptotic stabilization for controllable systems without drift”. In: *Mathematics of Control, Signals, and Systems* 5 (1992).
- [Cor09] J.-M. Coron. *Control and Nonlinearity*. Vol. 136. Mathematical Surveys and Monographs. Providence, Rhode Island: American Mathematical Society, 2009.
- [CMS20] J.-M. Coron, F. Marbach, and F. Sueur. “Small-time global exact controllability of the Navier-Stokes equation with Navier slip-with-friction boundary conditions”. In: *Journal of the European Mathematical Society* 22.5 (2020).

- [Cre+20] A. J. Creely et al. “Overview of the SPARC tokamak”. In: *Journal of Plasma Physics* 86.5 (2020).
- [DA107] D. D’Alessandro. *Introduction to Quantum Control and Dynamics*. New York: Chapman and Hall/CRC, 2007. 360 pp.
- [Dal18] J. Dalphin. “Uniform ball property and existence of optimal shapes for a wide class of geometric functionals”. In: *Interfaces and Free Boundaries* 20.2 (2018).
- [Dal20] J. Dalphin. “Existence of optimal shapes under a uniform ball condition for geometric functionals involving boundary value problems”. In: *ESAIM: Control, Optimisation and Calculus of Variations* 26 (2020).
- [Das16] Dask Development Team. *Dask: Library for dynamic task scheduling*. manual. 2016.
- [DZ11] M. C. Delfour and J. -P. Zolésio. *Shapes and geometries*. Second. Vol. 22. Advances in Design and Control. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011.
- [Del00] M. C. Delfour. “Tangential differential calculus and functional analysis on a  $C^{1,1}$  submanifold”. In: *Differential geometric methods in the control of partial differential equations*. Vol. 268. Contemp. Math. Amer. Math. Soc., Providence, RI, 2000.
- [DH98] R. Dewar and S. Hudson. “Stellarator symmetry”. In: *Physica D: Nonlinear Phenomena* 112.1-2 (1998).
- [EGP18] A. Enciso, M. Á. García-Ferrero, and D. Peralta-Salas. “The Biot-Savart operator of a bounded domain”. In: *Journal de Mathématiques Pures et Appliquées* 119 (2018).
- [EVZ93] M. Escobedo, J. L. Vazquez, and E. Zuazua. “Asymptotic behaviour and source-type solutions for a diffusion-convection equation”. In: *Archive for Rational Mechanics and Analysis* 124.1 (1993).
- [EG92] L. C. Evans and R. F. Gariepy. *Measure theory and fine properties of functions*. Studies in advanced mathematics. CRC Press, Boca Raton, FL, 1992.
- [Fed69] H. Federer. *Geometric measure theory*. Die grundlehren der mathematischen wissenschaften, band 153. Springer-Verlag New York Inc., 1969.
- [FL99] E. Feireisl and P. Laurençot. “The  $L^1$  -stability of constant states of degenerate convection–diffusion equations”. In: *Asymptotic Analysis* 19 (3-4 1999).
- [FG07] E. Fernández-Cara and S. Guerrero. “Null controllability of the Burgers system with distributed controls”. In: *Systems & Control Letters* 56.5 (2007).
- [FZ00] E. Fernández-Cara and E. Zuazua. “Null and approximate controllability for weakly blowing up semilinear heat equations”. In: *Annales de l’Institut Henri Poincaré C, Analyse non linéaire* 17.5 (2000).
- [Ful60] A. T. Fuller. “Relay control systems optimized for various performance criteria”. In: *IFAC Proceedings Volumes*. 1st International IFAC Congress on Automatic and Remote Control, Moscow, USSR, 1960 1.1 (1960).
- [FI99] A. V. Fursikov and O. Y. Imanuvilov. “Exact controllability of the Navier-Stokes and Boussinesq equations”. In: *Russian Mathematical Surveys* 54.3 (1999).
- [FI95] A. V. Fursikov and O. Y. Imanuvilov. “On controllability of certain systems simulating a fluid flow”. In: *Flow Control*. Ed. by M. D. Gunzburger. The IMA Volumes in Mathematics and its Applications. New York, NY: Springer, 1995.
- [FI96] A. V. Fursikov and O. Imanuvilov. *Controllability of evolution equations*. Lecture note series 34. Seoul National University, 1996.

- [GD02] M. Garwood and L. DelaBarre. “The return of the frequency sweep: Designing adiabatic pulses for contemporary NMR”. In: *Journal of magnetic resonance (San Diego, Calif.: 1997)* 153 (2002).
- [Gat+17] D. Gates et al. “Recent advances in stellarator optimization”. In: *Nuclear Fusion* 57.12 (2017).
- [Gla+15] S. J. Glaser et al. “Training Schrödinger’s cat: quantum optimal control. Strategic report on current status, visions and goals for research in Europe”. In: *European Physical Journal D* 69, 279 (2015).
- [Gla97] O. Glass. “Contrôlabilité exacte frontière de l’équation d’Euler des fluides parfaits incompressibles en dimension 3”. In: *Comptes Rendus de l’Académie des Sciences - Series I - Mathematics* 325.9 (1997).
- [GD03] A. Granas and J. Dugundji. *Fixed point theory*. Springer monographs in mathematics. Springer-Verlag, New York, 2003.
- [Gri85] P. Grisvard. *Elliptic problems in nonsmooth domains*. Vol. 24. Monographs and studies in mathematics. Pitman (Advanced Publishing Program), Boston, MA, 1985.
- [GI07] S. Guerrero and O. Y. Imanuvilov. “Remarks on global controllability for the Burgers equation with two control forces”. In: *Annales de l’Institut Henri Poincaré C, Analyse non linéaire* 24.6 (2007).
- [GIP06] S. Guerrero, O. Y. Imanuvilov, and J.-P. Puel. “Remarks on global approximate controllability for the 2-D Navier–Stokes system with Dirichlet boundary conditions”. In: *Comptes Rendus Mathématique* 343.9 (2006).
- [GIP12] S. Guerrero, O. Y. Imanuvilov, and J.-P. Puel. “A result concerning the global approximate controllability of the Navier–Stokes system in dimension 3”. In: *Journal de Mathématiques Pures et Appliquées* 98.6 (2012).
- [GY13] B.-Z. Guo and D.-H. Yang. “On convergence of boundary Hausdorff measure and application to a boundary shape optimization problem”. In: *SIAM Journal on Control and Optimization* 51.1 (2013).
- [Hah50] E. L. Hahn. “Spin echoes”. In: *Physical Review* 80.4 (1950).
- [HZ16] Y. Han and M. Zhu. “Hardy–Littlewood–Sobolev inequalities on compact riemannian manifolds and applications”. In: *Journal of Differential Equations* 260.1 (2016).
- [Har+20] C. R. Harris et al. “Array programming with NumPy”. In: *Nature* 585.7825 (2020).
- [Heb00] E. Hebey. *Nonlinear Analysis on Manifolds: Sobolev Spaces and Inequalities: Sobolev Spaces and Inequalities*. Courant lecture notes in mathematics. Courant Institute of Mathematical Sciences, 2000.
- [Hel14] P. Helander. “Theory of plasma confinement in non-axisymmetric magnetic fields.” In: *Reports on progress in physics. Physical Society* (2014).
- [HS05] P. Helander and D. J. Sigmar. *Collisional Transport in Magnetized Plasmas*. Cambridge University Press, 2005.
- [HP18] A. Henrot and M. Pierre. *Shape variation and optimization: a geometrical analysis*. Vol. 28. EMS Tracts in Mathematics. Zürich, Switzerland: European Math. Soc (EMS), 2018.
- [HB98] S. P. Hirshman and J. Breslau. “Explicit spectrally optimized Fourier series for nested magnetic surfaces”. In: *Physics of Plasmas* 5.7 (1998).

- [Hop50] E. Hopf. “The partial differential equation  $u_t + uu_x = \mu u_{xx}$ ”. In: *Communications on Pure and Applied Mathematics* 3.3 (1950).
- [Hun07] J. D. Hunter. “Matplotlib: A 2D graphics environment”. In: *Computing in Science & Engineering* 9.3 (2007).
- [ILT06] R. Illner, H. Lange, and H. Teismann. “Limitations on the control of Schrödinger equations”. In: *ESAIM. Control, Optimisation and Calculus of Variations* 12.4 (2006).
- [Ima+02] S. Imagawa, A. Sagara, K. Watanabe, T. Satow, and O. Motojima. “Magnetic field and force of helical coils for force free helical reactor (FFHR)”. In: *J. Plasma Fusion Res. SERIES* 5 (2002).
- [IPW19] L.-M. Imbert-Gerard, E. Paul, and A. Wright. “An introduction to symmetries in stellarators”. 2019.
- [02] *ITER technical basis*. ITER EDA documentation series 24. Vienna: International Atomic Energy Agency, 2002.
- [Jo+17] H. Jo, H.-g. Lee, S. Guérin, and J. Ahn. “Robust two-level system control by a detuned and chirped laser pulse”. In: *Physical Review A* 96.3 (2017).
- [Jos17] J. Jost. *Riemannian geometry and geometric analysis*. seventh. Universitext. Springer, Cham, 2017.
- [Kal63] R. E. Kalman. “Mathematical Description of Linear Dynamical Systems”. In: *Journal of the Society for Industrial and Applied Mathematics Series A Control* 1.2 (1963).
- [Koc+22] C. P. Koch et al. “Quantum optimal control in quantum technologies. Strategic report on current status, visions and goals for research in Europe”. 2022.
- [Kup90] I. Kupka. “The ubiquity of Fuller’s phenomenon”. In: *Nonlinear Controllability and Optimal Control*. Monograph Textbooks Pure and Applied Mathematics 133. 1990.
- [LSU68] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Uralbceva. *Linear and quasilinear equations of parabolic type*. Translations of mathematical monographs, vol. 23. American Mathematical Society, Providence, R.I., 1968.
- [LPS15] S. K. Lam, A. Pitrou, and S. Seibert. “Numba: a LLVM-based python JIT compiler”. In: *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC* (Austin, Texas). LLVM ’15. New York, NY, USA: Association for Computing Machinery, 2015.
- [Lan17] M. Landreman. “An improved current potential method for fast computation of stellarator coil shapes”. In: *Nuclear Fusion* 57.4 (2017).
- [Léa12] M. Léautaud. “Uniform controllability of scalar conservation laws in the vanishing viscosity limit”. In: *SIAM Journal on Control and Optimization* 50.3 (2012).
- [Lee12] J. M. Lee. *Introduction to Smooth Manifolds*. Vol. 218. Graduate Texts in Mathematics. Springer New York, 2012.
- [LSR11] Z. Leghtas, A. Sarlette, and P. Rouchon. “Adiabatic passage and ensemble control of quantum systems”. In: *Journal of Physics B* 44.15 (2011).
- [Lei+22] M. Leibscher et al. “Full quantum control of enantiomer-selective state transfer in chiral molecules despite degeneracy”. In: *Communications Physics* 5.1 (1 2022).
- [LK09] J.-S. Li and N. Khaneja. “Ensemble control of Bloch equations”. In: *IEEE Transactions on Automatic Control* 54.3 (2009).

- [LZZ19] J. Li, B.-Y. Zhang, and Z. Zhang. “Well-posedness of the generalized Burgers equation on a finite interval”. In: *Applicable Analysis* 98.16 (2019).
- [LK06] J.-S. Li and N. Khaneja. “Control of inhomogeneous quantum ensembles”. In: *Physical Review A: Atomic, Molecular, and Optical Physics* 73.3 (2006).
- [Lin76] G. Lindblad. “On the generators of quantum dynamical semigroups”. In: *Communications in Mathematical Physics* 48.2 (1976).
- [Lio91] J. L. Lions. “Exact Controllability for Distributed Systems. Some Trends and Some Problems”. In: *Applied and Industrial Mathematics: Venice - 1, 1989*. Ed. by R. Spigler. Mathematics and Its Applications. Dordrecht: Springer Netherlands, 1991.
- [Lio69] J.-L. Lions. *Quelques méthodes de résolution des problèmes aux limites non linéaires*. Paris: Dunod, 1969.
- [LM72] J.-L. Lions and E. Magenes. *Non-homogeneous boundary value problems and applications. Vol. II*. Die grundlehren der mathematischen wissenschaften, band 182. Springer-Verlag, New York-Heidelberg, 1972.
- [Mac+19] U. A. Maciel Neto, P. S. Pereira da Silva, K. Beauchard, and P. Rouchon. “ $H^1$ -Control of an ensemble of half-spin systems replacing Rabi pulses by adiabatic following”. In: *Proceedings of the 58th IEEE conference on decision and control*. 2019.
- [MK01] V. Malinovsky and J. Krause. “General theory of population transfer by adiabatic rapid passage with intense, chirped laser pulses”. In: *The European Physical Journal D: Atomic, Molecular, Optical and Plasma Physics* 1450 (2001).
- [Mar14] F. Marbach. “Small time global null controllability for a viscous Burgers’ equation despite the presence of a boundary layer”. In: *Journal de Mathématiques Pures et Appliquées* 102.2 (2014).
- [Mar+15] E. Marmar, S. Baek, H. Barnard, P. Bonoli, D. Brunner, J. Candy, et al. “Alcator C-Mod: research in support of ITER and steps beyond”. In: *Nucl. Fusion* 55 (2015).
- [Mea+02] D. Meade, S. Jardin, C. Kessel, J. Mandrekas, M. Ulrickson, et al. “FIRE, exploring the frontiers of burning plasma science”. In: *J. Plasma Fusion Res. SER.* 5 (2002).
- [Mer86] P. Merkel. “An integral equation technique for the exterior and interior neumann problem in toroidal regions”. In: *Journal of Computational Physics* 66.1 (1986).
- [Mer87] P. Merkel. “Solution of stellarator boundary value problems with external currents”. In: *Nuclear Fusion* 27.5 (1987).
- [MR04] M. Mirrahimi and P. Rouchon. “Controllability of quantum harmonic oscillators”. In: *Automatic Control, IEEE Transactions on* 49 (2004).
- [Mir+14] M. Mirrahimi et al. “Dynamically protected cat-qubits: a new paradigm for universal quantum computation”. In: *New Journal of Physics* 16.4 (2014).
- [MS77] B. Misra and E. C. G. Sudarshan. “The Zeno’s paradox in quantum theory”. In: *Journal of Mathematical Physics* 18.4 (1977).
- [Mit13] M. H. Mittleman. *Introduction to the theory of laser-atom interactions*. Springer Science & Business Media, 2013.
- [MS76a] F. Murat and J. Simon. “Sur le contrôle par un domaine géométrique”. In: *Publication du Laboratoire d’Analyse Numérique de l’Université Paris 6* 189 (1976).
- [MS76b] F. Murat and J. Simon. *Étude de problèmes d’optimal design*. Vol. 41. Lecture notes in computer science. Berlin: Springer-Verlag, 1976.

- [Mur70a] J. D. Murray. “On the Gunn effect and other physical examples of perturbed conservation equations”. In: *Journal of Fluid Mechanics* 44.2 (1970).
- [Mur70b] J. D. Murray. “Perturbation effects on the decay of discontinuous solutions of nonlinear first order wave equations”. In: *SIAM Journal on Applied Mathematics* 19.2 (1970).
- [Pau20] E. Paul. *Adjoint methods for stellarator shape optimization and sensitivity analysis*. 2020. arXiv: 2005.07633 [physics.plasm-ph].
- [Pau+18] E. J. Paul, M. Landreman, A. Bader, and W. Dorland. “An adjoint method for gradient-based optimization of stellarator coil shapes”. In: *Nuclear Fusion* 58.7 (2018).
- [Pau+19] E. J. Paul, I. G. Abel, M. Landreman, and W. Dorland. “An adjoint method for neoclassical stellarator optimization”. In: *Journal of Plasma Physics* 85.5 (2019).
- [Pom+01] N. Pomphrey et al. “Innovations in compact stellarator coil design”. In: *Nuclear Fusion* 41.3 (2001).
- [Pon+74] L. Pontriaguine, V. Boltianski, R. Gamkrélidzé, and E. Michtchenko. *Théorie mathématique des processus optimaux*. Ed. by É. Mir. 1974.
- [Poz22] E. Pozzoli. “Classical and quantum controllability of a rotating asymmetric molecule”. In: *Applied Mathematics & Optimization* 85.1 (2022).
- [PRS22a] Y. Privat, R. Robin, and M. Sigalotti. *Existence of surfaces optimizing geometric and PDE shape functionals under reach constraint*. 2022. arXiv: 2206.04357 [math].
- [PRS22b] Y. Privat, R. Robin, and M. Sigalotti. “Optimal shape of stellarators for magnetic confinement fusion”. In: *Journal de Mathématiques Pures et Appliquées* 163 (2022).
- [PS07] P. Pucci and J. Serrin. *The Maximum Principle*. Red. by H. Brezis et al. Vol. 73. Progress in Nonlinear Differential Equations and Their Applications. Basel: Birkhäuser, 2007.
- [Puc+15] G. Pucella et al. “Overview of the FTU results”. In: *Nuclear Fusion* 55.10 (2015).
- [Que+18] V. Queral, F. Volpe, D. Spong, S. Cabrera, and F. Tabarés. “Initial exploration of high-field pulsed Stellarator approach to ignition experiments”. In: *Journal of Fusion Energy* 37 (2018).
- [RV11] P. Ramachandran and G. Varoquaux. “Mayavi: 3D visualization of scientific data”. In: *Computing in Science & Engineering* 13.2 (2011).
- [Rob+22a] R. Robin, U. Boscain, M. Sigalotti, and D. Sugny. *Chattering phenomenon in quantum optimal control*. 2022. arXiv: 2206.13868 [quant-ph].
- [Rob22] R. Robin. *Small-time global null controllability of generalized Burgers’ equations*. 2022. arXiv: 2206.05931 [math].
- [Rob+22b] R. Robin, N. Augier, U. Boscain, and M. Sigalotti. “Ensemble qubit controllability with a single control via adiabatic and rotating wave approximations”. In: *Journal of Differential Equations* 318 (2022).
- [RV22] R. Robin and F. A. Volpe. “Minimization of magnetic forces on stellarator coils”. In: *Nuclear Fusion* 62.8 (2022).
- [Rou02] P. Rouchon. “On the control of quantum oscillators”. In: *Technical report, Centre Automatique et Systèmes, Ecole des Mines de Paris* (2002).

- [Rou08] P. Rouchon. “Quantum systems and control 1”. In: *Revue Africaine de la Recherche en Informatique et Mathématiques Appliquées* Volume 9, Conference in Honor of Claude Lobry (2008).
- [SN87] P. L. Sachdev and K. R. C. Nair. “Generalized Burgers equations and Euler–Painlevé transcendents. II”. In: *Journal of Mathematical Physics* 28.5 (1987).
- [SIN10] A. Sagara, Y. Igitkhanov, and F. Najmabadi. “Review of stellarator/heliotron design issues towards MFE DEMO”. In: *Fusion Engineering and Design* 85.7 (2010).
- [SVM07] J. A. Sanders, F. Verhulst, and J. Murdock. *Averaging Methods in Nonlinear Dynamical Systems*. 2nd ed. Applied Mathematical Sciences. New York: Springer-Verlag, 2007.
- [SL09] H. Schättler and U. Ledzewicz. “Singular controls and chattering arcs in optimal control problems arising in biomedicine”. In: *Control and Cybernetics* 38 (2009).
- [SL12] H. Schättler and U. Ledzewicz. *Geometric Optimal Control*. Vol. 38. Interdisciplinary Applied Mathematics. New York, NY: Springer New York, 2012.
- [SEB13] F. Schauer, K. Egorov, and V. Bykov. “HELIAS 5-B magnet system structure and maintenance concept”. In: *Fusion Engineering and Design* 88.9 (2013).
- [Sch+14] J. Scheuer et al. “Precise qubit control beyond the rotating wave approximation”. In: *New Journal of Physics* 16.9 (2014).
- [Sha12] P. Shapiro Moshe; Brumer. *Quantum control of molecular processes*. John Wiley & Sons, Ltd, 2012.
- [Sho08] B. Shore. “Coherent manipulations of atoms using laser light”. In: *Acta Physica Slovaca* 58 (2008).
- [Sho11] B. W. Shore. *Manipulating quantum structures using laser pulses*. Cambridge: Cambridge University Press, 2011.
- [Sim+11] C.-M. Simon et al. “Robust quantum dot exciton generation via adiabatic passage with frequency-swept optical pulses”. In: *Physical Review Letters* 106.16 (2011).
- [SG18] D. G. A. Smith and J. Gray. “opt\_einsum - A Python package for optimizing contraction order for einsum-like expressions”. In: *Journal of Open Source Software* 3.26 (2018).
- [Ste13] A. Stern. “ $L^p$  change of variables inequalities on manifolds”. In: *Mathematical Inequalities & Applications* 1 (2013).
- [Str+04] D. Strickler et al. “Development of a robust quasi-poloidal compact Stellarator”. In: *Fusion Science and Technology* 45 (2004).
- [SBH02] D. J. Strickler, L. A. Berry, and S. P. Hirshman. “Designing coils for compact stellarators”. In: *Fusion Science and Technology* 41.2 (2002).
- [SW99] X. Sun and M. J. Ward. “Metastability for a generalized Burgers equation with applications to propagating flame fronts”. In: *European Journal of Applied Mathematics* 10.1 (1999).
- [Tel83] N. Teleman. “The index of signature operators on Lipschitz manifolds”. In: *Publications Mathématiques de l’IHÉS* 58 (1983).
- [Teu03] S. Teufel. *Adiabatic perturbation theory in quantum dynamics*. Vol. 1821. Lecture notes in mathematics. Berlin: Springer-Verlag, 2003.



- [TGV11] B. T. Torosov, S. Guérin, and N. V. Vitanov. “High-fidelity adiabatic passage by composite sequences of chirped pulses”. In: *Physical Review Letters* 106.23 (2011).
- [VKL99] L. Viola, E. Knill, and S. Lloyd. “Dynamical decoupling of open quantum systems”. In: *Physical Review Letters* 82.12 (1999).
- [Vir+20] P. Virtanen et al. “SciPy 1.0: Fundamental algorithms for scientific computing in python”. In: *Nature Methods* 17 (2020).
- [Vit+01] N. V. Vitanov, M. Fleischhauer, B. W. Shore, and K. Bergmann. “Coherent manipulation of atoms and molecules by sequential laser pulses”. In: *Advances in Atomic Molecular and Optical Physics* 46 (2001).
- [Vit+17] N. V. Vitanov, A. A. Rangelov, B. W. Shore, and K. Bergmann. “Stimulated Raman adiabatic passage in physics, chemistry and beyond”. In: *Reviews of Modern Physics* 89.1 (2017).
- [WM93] H. M. Wiseman and G. J. Milburn. “Quantum theory of optical feedback via homodyne detection”. In: *Physical Review Letters* 70.5 (1993).
- [Wu+11] Y. Wu et al. “Population inversion in a single InGaAs quantum dot using the method of adiabatic rapid passage”. In: *Physical Review Letters* 106.6 (2011).
- [Zar+01] M. C. Zarnstorff et al. “Physics of the compact advanced stellarator NCSX”. In: *Plasma Physics and Controlled Fusion* 43 (12A 2001).
- [ZB94] M. I. Zelikin and V. F. Borisov. *Theory of Chattering Control: with applications to Astronautics, Robotics, Economics, and Engineering*. Systems & Control: Foundations & Applications. Birkhäuser Basel, 1994.
- [Zhu+18a] C. Zhu, S. R. Hudson, S. A. Lazerson, Y. Song, and Y. Wan. “Hessian matrix approach for determining error field sensitivity to coil deviations”. In: *Plasma Physics and Controlled Fusion* 60.5 (2018).
- [Zhu+18b] C. Zhu, S. R. Hudson, Y. Song, and Y. Wan. “Designing stellarator coils by a modified Newton method using FOCUS”. In: *Plasma Physics and Controlled Fusion* 60.6 (2018).
- [Zhu+18c] C. Zhu, S. R. Hudson, Y. Song, and Y. Wan. “New method to design stellarator coils without the winding surface”. In: *Nuclear Fusion* 58.1 (2018).
- [ZTC16] J. Zhu, E. Trélat, and M. Cerf. “Planar tilting maneuver of a spacecraft: singular arcs in the minimum time problem and chattering”. In: *Discrete and Continuous Dynamical Systems - Series B* 16.4 (2016).



## CONTROL AND OPTIMIZATION OF PHYSICAL SYSTEMS: QUANTUM DYNAMICS AND MAGNETIC CONFINEMENT IN STELLARATORS

### Abstract

Cette thèse porte sur l'optimisation et le contrôle de plusieurs systèmes physiques : elle est composée de trois parties.

La première partie est consacrée aux stellarators. Ce type de réacteur à fusion nucléaire pose de nombreux défis liés à l'optimisation. Nous nous sommes concentrés sur un problème inverse bien connu des physiciens, modélisant la conception optimale de bobines supraconductrices générant un champ magnétique donné. Nous avons conduit une étude théorique et numérique d'une extension de ce problème, portant sur une optimisation de forme. Nous avons ensuite développé une nouvelle méthode afin de prouver l'existence de formes optimales dans le cas de problèmes d'optimisation d'hypersurfaces. Nous avons enfin effectué l'étude et l'optimisation des forces de Laplace s'exerçant sur une densité surfacique de courant.

La deuxième partie porte ensuite sur l'étude du contrôle de systèmes quantiques de dimension finie. Nous avons étudié rigoureusement la combinaison de l'approximation de l'onde tournante avec l'approximation adiabatique. Dans un premier temps, nous avons obtenu la robustesse des méthodes de transfert de population sur les qubits. Cette dernière permet alors d'étendre des résultats de Li et Khaneja sur le contrôle d'ensemble des qubits en se restreignant à l'utilisation d'un seul contrôle. Nous présentons également une seconde contribution, consacrée à l'analyse d'un phénomène de *chattering* pour un problème de contrôle optimal d'un système quantique.

Enfin, la troisième partie est dédiée à la preuve d'un résultat de contrôlabilité à zéro en temps petit pour des équations de Burgers généralisées grâce à l'utilisation d'une couche limite.

**Keywords:** optimal control, quantum control, ensemble controlability, qubit, shape optimization, plasma physics, stellarator, Burgers equation, boundary layer

---

## CONTROL AND OPTIMIZATION OF PHYSICAL SYSTEMS: QUANTUM DYNAMICS AND MAGNETIC CONFINEMENT IN STELLARATORS

### Abstract

This PhD manuscript deals with the optimization and control of several physical systems. It is divided into three parts.

The first part is devoted to stellarators. This type of nuclear fusion reactor poses many challenges related to optimization. We focus on an inverse problem well known to physicists, modeling the optimal design of superconducting coils generating a given magnetic field. We conduct both a theoretical and a numerical study of an extension of this problem, involving shape optimization. Then, we develop a new method to prove the existence of optimal shapes in the case of hypersurface optimization problems. Finally, we study and optimize the Laplace forces acting on a current surface density.

The second part of this manuscript deals with the control of finite dimensional quantum systems. We rigorously study the combination of the rotating wave approximation with the adiabatic approximation. First, we obtain the robustness of a population transfer method on qubits. The latter then allows to extend results of Li and Khaneja on the ensemble control of qubits by restricting to the use of a single control. We also present a second contribution, devoted to the analysis of a chattering phenomenon for an optimal control problem of a quantum system.

Finally, the third part is dedicated to the proof of a small-time global null controllability result for generalized Burgers' equations using a boundary layer.

**Keywords:** optimal control, quantum control, ensemble controlability, qubit, shape optimization, plasma physics, stellarator, Burgers equation, boundary layer

---



Laboratoire Jacques-Louis Lions, Équipe Inria CAGE

Sorbonne Université – Campus Pierre et Marie Curie – 4 place Jussieu – 75005 Paris – France