Università di Padova – Dipartimento di Matematica "Tullio Levi-Civita"

Scuole di Dottorato in Matematica Pura e Computazionale

Seminario Dottorato 2021/22

Preface

This document offers an overview of the activity of Seminario Dottorato 2021/22.

Our "Seminario Dottorato" (Graduate Seminar) has a double purpose. At one hand, the speakers — usually Ph.D. students or post-docs, but sometimes also senior researchers — are invited to communicate their researches to a public of mathematically well-educated but not specialist people, by preserving both understandability and the flavour of a research report. At the same time, people in the audience enjoy a rare opportunity to get an accessible but also precise idea of what's going on in some mathematical research area that they might not know very well.

Due to the continuation of the COVID-19 pandemic emergency, all sessions of the seminar have been held in dual mode, both in presence and online. Once more we have observed that keeping up this activity has helped the PhD students to stay in contact among them and with the Department life.

Let us take this opportunity to warmly thank once again all the speakers for having held these interesting seminars and for their nice agreement to write down these notes to leave a concrete footstep of their participation.

We are also grateful to the collegues who helped us, through their advices and suggestions, in building an interesting and culturally complete program.

Padova, June 20th, 2022

Corrado Marastoni, Tiziano Vargiolu

Abstracts (from Seminario Dottorato's webpage)

Wednesday 6 October 2021

The Modularity Theorem and Fermat's Last Theorem

Luca MASTELLA (Padova, Dip. Mat.)

Fermat's Last Theorem (FLT) is one of the most important and challenging problems of the last centuries in Number Theory. Its complete proof rests on the Modularity Conjecture for semistable elliptic curves defined over $\mathbb{Q}$, proven to be true only in 1994 by A. Wiles.
The seminar will give an introduction to FLT, focusing on some historical aspects of its proof. In the second part we will give to the audience the statement of the Modularity theorem and introduce in an elementary way the arithmetic objects involved in it. We intend moreover to give a brief account of the implication Modularity $\Longrightarrow$ FLT.

Wednesday 17 November 2021

Embeddings of spaces with multiweighted derivatives and their applications

Zhanar KEULIMZHAYEVA (S. Seifullin Kazakh Agro Technical University)

The analysis of function spaces is particularly relevant in several areas of Mathematics such as differential and integral equations. Among function spaces, the weighted function spaces turn out to be suitable for the analysis of boundary value problems with various types of singularities.
This seminar presents a brief overview of with weight functional space of the Kudryavtsev type, called the space with multiweighted derivatives. In the first part of the talk, we will define space with multiweighted derivatives and consider the behavior of the function at the boundary of the investigated space. Following the ideas of L.D. Kudryavtsev, we will define the boundary values of a function and of its derivatives at the singular point. In addition, we will give necessary and sufficient conditions for weight functions in order that each function of the space is stabilized to some unique polynomial at zero, and will provide estimates for the rate of stabilization to a polynomial. Next, we will introduce functionals that depend on the boundary values at the singular point and that are equivalent to the norm of the space. In the very last part of the talk, we will introduce necessary and sufficient conditions for weight functions so that continuous and compact embeddings between spaces of multiweighted derivatives hold. Moreover, we will show conditions on the weight functions in order that the inequality of the Nikol'skii-Lizorkin-Kudryatsev type is valid.

Wednesday 1 December 2021

Mathematical Finance: a Tale of Stochastic Processes

Guillaume SZULDA (Padova, Dip. Mat.)

Financial markets are highly uncertain environments where the evolution of asset prices exhibits random fluctuations over time, in particular due to complex and unpredictable market mechanisms. In this regard, stochastic analysis, which is at the intersection between the theory of probability and functional analysis, plays a fundamental role in financial modeling.
Being aware of the non-specialist yet mathematically strong nature of the audience, I divide my talk into two major parts. The first part is mostly introductory, where I first give/recall elementary but indispensable notions of probability and stochastic calculus, then I illustrate the fundamentals of mathematical finance. I mention that throughout this part, I put the emphasis on the modeling aspects, most notably the extensive application of stochastic processes. In the second part, I present the topic of my doctoral research, i.e. Branching processes and multiple term structure modeling. I start by defining the multiple term structure framework and providing a construction of Continuous-state Branching processes with Immigration (CBI), which constitute a sophisticated class of stochastic processes. I carry on with examples of how CBI processes can be exploited for the modeling of financial markets where multiple term structures typically coexist. Finally, I propose some avenues of further developments.

Wednesday 15 December 2021

The critical node/edge detection problem on trees

Syed MD Omar FARUK (Padova, Dip. Mat.)

Critical node or edge detection problems are a family of optimization problems defined on graphs, where one is required to remove a limited number of nodes and/or edges in order to minimize some measure of the connectivity of the residual graph. Problems of this type are important from a practical point of view because of their relevance in a number of practical applications.
We start this seminar by giving the definitions of the critical node/edge detection problem (CNDP/ CEDP) and some connectivity metrics with an example. After that, we present a dynamic programming approach for solving the CNDP on trees when the node weights are all equal to one and all connections between pairs of nodes have unit cost. Then, we will move to consider the CEDP on trees and similarly deal with the case with unit costs and unit edge weights. Finally, we will present dynamic programming algorithms for the ?mixed? case, in which nodes and edges can be simultaneously removed from the graph.

Wednesday 19 January 2022

Optimizing smooth objectives on convex sets without projections

Damiano ZEFFIRO (Padova, Dip. Mat.)

The well known gradient descent method for smooth unconstrained optimization can be extended in a straightforward way to problems with convex constraints by using projections. However, in many cases there are more effective ways to generate feasible descent directions. One of the most popular alternatives to the projected gradient method is the Frank-Wolfe method, characterized by a linear minimization subproblem replacing the projection subproblem.
In this seminar, after a brief review of the above mentioned methods, some examples of sets commonly used in optimization where linear minimization is cheaper than projection will be discussed. Then, variants to improve the convergence rate of the Frank-Wolfe method will be presented, together with a general framework to study such variants. Finally, an algorithm for fast cluster detection in networks based on a Frank-Wolfe variant will be described.

Wednesday 2 February 2022

Modular curves and Heegner points

Daniele TROLETTI (Padova, Dip. Mat.)

One the the main open conjecture is the one due to Birch and Swinnerton-Dyer about elliptic curves. There are many attempts to prove it but they were able to prove only some special cases, like the rank 1 case proven by Kolyvagin using the Heegner points method.
This seminar will give an introduction on the basis required to define the Heegner points, such as elliptic and modular curves. After that we are going to define Heegner points and show some results achieved using them.

Wednesday 23 February 2022

Beyond Nash Equilibria in Mean Field Games

Ofelia BONESINI (Padova, Dip. Mat.)

The concept of Nash Equilibrium is the most important (and famous) notion in Game Theory.
Assuming that the audience is not familiar with the topic, we will first warm up with an introduction to recall all the basic definitions and results. Then, we will focus on two extensions: Correlated Equilibria and Mean Field Games. Finally, we will gather things together to see how the definition of a Correlated solution can be formulated and its validity checked in the mean field context. Time permitting, I will mention some results of my research.

Wednesday 2 March 2022

Optimal control problems: existence of minimizers, necessary conditions, and gap phenomena

Giovanni FUSCO (Padova, Dip. Mat.)

By optimal control problem we mean the minimization of a functional over arcs that satisfy certain constraints (dynamics, control, endpoint and state constraints).
After a brief introduction on the subject, we will discuss the notion of closure of trajectories associated with a controlled differential equation, so that to present an existence theorem for optimal control problems. Then, we will announce the Pontryagin's Maximum Principle, that is, the most known set of necessary conditions that has to be fulfilled by a minimizer. Afterwards, we will introduce the most common extensions for the optimal control problems which do not admit minimizers and we will analyze the properness of such extensions. In particular, we will deal with the issue of gap phenomena between an optimal control problem and an its extension and we will prove a link between this occurrence and a topological property of the trajectories which is usually called isolation. Finally, we will establish that isolated trajectories satisfy the Maximum Principle in abnormal form, i.e. there exists at least a set of multipliers with cost multiplier equal to zero. We will conclude with some examples that illustrate the outcomes.

Wednesday 13 April 2022

The Graph p-Laplacian Eigenvalue Problem

Piero DEIDDA (Padova, Dip. Mat.)

In the last years graphs have been used to model a large variety of phenomena, such as the classical transportation networks, the social network interactions, the chemical reactions, ecological interactions. In the study of the structure of a graph, two classical problems are the shortest path problem and the optimal partition or cut problem. Interestingly both of these problems can be related to the p-Laplacian eigenvalue problem and in particular with the two degenerate cases of $p = 1$ and $p = \infty$. The p-Laplacian eigenvalue problem historically arose from the study of the classical Poincare' constant and can be addressed as the study of the critical points of the so called "Rayleigh quotient". It is one of the most classical example of non-linear eigenvalue problem and has been widely studied both in the continuous and discrete setting. We provide a short summary of the main results in this field, highlighting the differences from the linear ($p = 2$) case and the topological information provided by the limit cases $p = 1$ and $p = \infty$. We will show at the end some original numerical schemes for the computation of the p-eigenpairs.

Wednesday 11 May 2022

Kolmogorov-Arnold-Moser (KAM) stability and its application in the planetary n-body problem

Rita MASTROIANNI (Padova, Dip. Mat.)

The study of exoplanetary systems with two or more planets in orbits with non-zero mutual inclination is an interesting topic of Hamiltonian dynamics, in view of the many applications related to the astronomical discovery, in the last 20 years, of several such systems. The present report discusses the mathematical context of the theory of the long term stability for nearly Keplerian perturbed n-body systems, following the so-called Kolmogorov-Arnold-Moser (KAM) Theorem. The KAM Theorem is a cornerstone of canonical perturbation theory: it allows to conjugate, through a convergent sequence of canonical trasformations, particular solutions of the "pertubed" dynamical system to the invariant dynamics on a torus. We provide a short summary of classical results of perturbation theory. We also briefly present some recent progress on the construction of the Kolmogorov normal form for 'isochronous systems'. Finally, we explain in an introductory manner, how the above concepts can be implemented in exoplanetary systems with a 3D-orbital architecture.

Wednesday 25 May 2022

Chaotic dynamical systems and applications to the solar system dynamics

Mattia ROSSI (Padova, Dip. Mat.)

Dynamical systems are an essential tool to model physical phenomena in applied sciences whose state changes over time according to either differential or discrete difference equations. In this context two concepts are in opposition: "order", or "integrability", versus "chaos". Integrable systems, for which all the solutions can be explicitly analytically determined, are special and represent only a crude approximation of the real dynamics. On the other hand, more accurate models are usually represented by non-integrable differential equations, whose solutions exhibit a highly sensitive dependence on initial conditions, termed as chaotic.
In this talk we discuss some of the main geometric and topological properties of deterministic chaos in connection with orbital stability of small objects in our solar system. After a short recap of the theory of non-linear dynamical systems, we present a modern approach of detecting and quantifying chaotic behaviors using finite time chaos indicators, a numerical strategy capable to capture the dynamical structure of the phase space. In the second part of the seminar, we introduce the restricted N-body problem in Hamiltonian mechanics and implement the above technique to discriminate between the realms of regular and chaotic motions of asteroids.

Wednesday 15 June 2022

Introduction to sub-Riemannian geometry

Alessandro SOCIONOVO (Padova, Dip. Mat.)

We discuss the general notions of sub-Riemannian geometry. In particular, we study in parallel a toy sub-Riemannian model taken from real life and the general sub-Riemannian setting. In the final part of this talk, we introduce the open problem of the regularity of sub-Riemannian geodesics.

The Modularity Theorem and Fermat's Last Theorem

Luca Mastella (*)

Abstract. Fermat's Last Theorem (FLT) is one of the most important and challenging problems of the last centuries in Number Theory. Its complete proof rest on the Modularity Conjecture for semistable elliptic curves defined over $\mathbb{Q}$, proven to be true only in 1994 by A. Wiles. The seminar will give an introduction to FLT, focusing on some historical aspects of its proof. In the second part we will give to the audience the statement of the Modularity theorem and introduce in an elementary way the arithmetic objects involved in it. We intend moreover to give a brief account of the implication Modularity $\Longrightarrow$ FLT.

1 Historical Remarks

It is well known at least since the ancient Greek time, and even before, that for the sides of a right triangle the equation

$$(P) \qquad a^2 + b^2 = c^2$$

holds, where a, b are the length of the legs and c is the length of the hypothenuse, that is a theorem called after Pythagoras. In number theory we are interested to this kind of equations, namely of polynomial equations with integral coefficient (called *Diophantine equations*) and in their integral solutions. The triples $(a, b, c) \in \mathbb{Z}^3$ that solve the equation (P) are called in particular *Phytagorean triples* and any such a solution correspond to a right triangle with sides of integral length. There are some trivial solutions of this equation, namely the triples $(0, 0, 0), (\pm 1, 0, \pm 1), (0, \pm 1, \pm 1)$, but with few calculations one can easily show that there are also a huge amount of nontrivial ones, e.g. $(3, 4, 5), (5, 12, 13), (7, 24, 25)$, etc. Naturally it comes up the question: how many are them? Are they a finite number or an infinite one? In the first case can we list them all? Or, in the latter, can we parametrize them in some way?

First of all observe that whenever (a, b, c) is an integral solution and $x \in \mathbb{Z}$, then (ax, bx, cx) is an integral solution, too. Therefore we have an infinite set of pythagorean

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: luca.mastella@math.unipd.it . Seminar held on 6 October 2021.

triples, but one can show that even the "primitive" ones (meaning that a, b and c have no common factor) are an infinite set and are of the form

$$\big(\pm(m^2-n^2), \pm 2mn, \pm(m^2+n^2)\big), \big(\pm 2mn, \pm(m^2-n^2), \pm(m^2+n^2)\big)$$

where $m, n \in \mathbb{Z}$, relatively primes and not both odd. This characterization easily follows from the "prime decomposition" in the ring of Gauss integers $\mathbb{Z}[i]$. (see [4, Ch. 1])

Once we know how all the pythagorean triples are made it is natural to consider equations of the form

$$(F_n) \qquad x^n + y^n = z^n,$$

for $n > 2$. They obviously still have the trivial solutions $(0,0,0), (\pm 1, 0, \pm 1), (0, \pm 1, \pm 1)$, but if one try to find some nontrivial ones he could have really an hard time. In fact it is now a theorem that there are none.

Fermat's Last Theorem (FLT) Let $n > 2$, if $(a, b, c) \in \mathbb{Z}^3$ is a solution of (F_n), then $abc = 0$.

Fermat's Last Theorem is one of the most important and challenging problems in Number Theory. Since it has been formulated, many of the most important mathematicians faced with it and a lot of beautiful mathematics was built in the attempt of proving it. The statement of the theorem was an intuition of the French judge Pierre Fermat (1608-1665), that wrote it (~ 1630) without proof on the margin on a copy of the *Arithmetica* of Diophantus, as he was used to do. But, conversely to his other theorems, that were proved (or sometimes disproved) by the mathematicians of that time, this theorem resisted to any attempt of proving it, despite the claim of Fermat of having a beautiful proof in his hands.

Special cases were proved thanks to the work of the best mathematicians of the last centuries (the most relevant is the work of Kummer), but the general case were fully proved only in 1994 by Andrew Wiles (with the help of a former student of him, Richard Taylor, with whom he filled a missing point in its proof) with some of the most advanced tools of modern Number Theory and Arithmetic Geometry.

We resume the main steps of the history of this proof.

- (1885) G. Frey suggest that the existence of a nontrivial integral solution of the FLT equation might contradict the Modularity conjecture of Taniyama, Shimura and Weil;
- (1985-6) J. P. Serre formulated its *Epsilon Conjecture* and proved that this together with the Modularity Conjecture would imply FLT;
- (1986) K. Ribet proved Serre's *Epsilon Conjecture*, reducing the proof of FLT to the Modularity Conjecture for semistable elliptic curves;
- (1993-4) A. Wiles gave a (later found incomplete) proof of the Modularity Conjecture for semistable elliptic curves;
- (1994-5) A. Wiles and R. Taylor filled the gap into Wiles proof.

This seminar is indeed meant to be an introduction to the Modularity conjecture and to give an idea of the reason why it implies FLT.

2 Elliptic curves

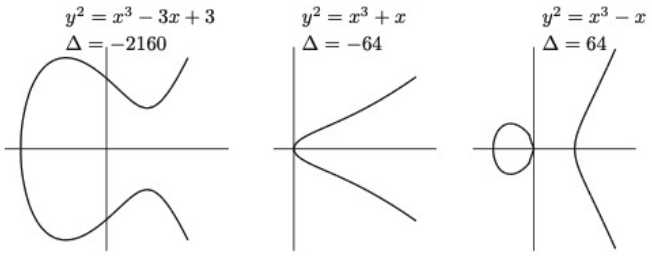
Despite the name *Number Theory*, or *Arithmetic*, the focus of the modern discipline is no more the integer numbers in themselves, but it had enriched with a lot of other tools coming from Algebra, Geometry, Analysis, etc. One of the most important objects studied in modern Number Theory are Elliptic curves. A good reference for the arithmetic theory of elliptic curves is [7].

Definition An Elliptic curve E is the plane complex curve whose points are the solution of the cubic equation

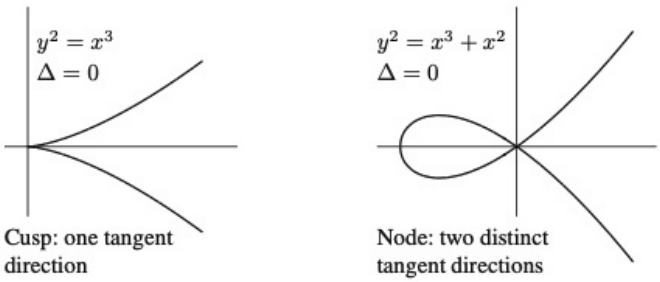
$$y^2 = 4x^3 - g_2 x - g_3,$$

with $g_2, g_3 \in \mathbb{C}$ and $\Delta = g_2^3 - 27g_3^2 \neq 0$.

Such an equation is called the *Weierstrass equation* of the elliptic curve. In a more geometric language an Elliptic Curve is a smooth complex algebraic cubic curve (see Figure 1). The feature that makes elliptic curves more significant for Number Theory than other algebraic plane curves is that their points give rise to an abelian group.



(a) Three elliptic curves.



(b) Two cubic singular curves.

Figure 1. An example of the real points of some elliptic curves (a) and some cubic curves that are not elliptic (b). Note that the elliptic curves are smooth, while the others have a singular point. Indeed, there is a classification theorem of cubic complex plane curves up to linear change of coordinates: either they are elliptic curves, or they have exactly one singular point that can be a note or a cusp. The two figures are taken from [7] (Figures 3.1 and 3.2).

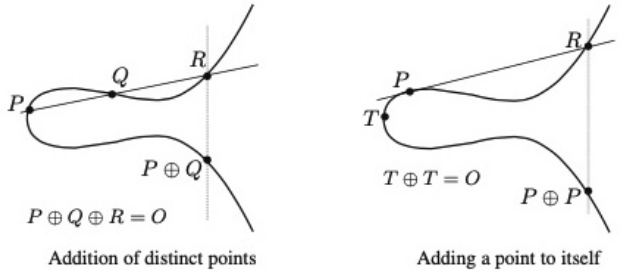


Figure 2. Geometric description of the composition law on the elliptic curve E. On the left we see the sum of two points $P \neq Q$: in this case $P \oplus Q$ is obtained reflecting w.r.t the x-axis the point R, namely the third point in which the line through P and Q meets E. On the right side we see the sum $P \oplus P$: it is the reflection of the other point in which the tangent line at P meets E. This figure is taken from [7] (Figure 3.3).

In order to define it we need to compactyfy the curve with a *point at infinity* (that in the sense on projective geometry is the point of infinity of the y-axis), that we denote by ∞. The set of complex (projective) points of E is defined to be

$$E(\mathbb{C}) = \{\, (x,y) \in \mathbb{C}^2 \colon y^2 = 4x^3 - g_2 x - g_3, \text{ s.t. } x, y \in \mathbb{C} \,\} \cup \{\, \infty \,\}\,.$$

Proposition *The composition law described in geometric terms in Figure 2 makes $E(\mathbb{C})$ into an abelian group, whose zero is ∞ and that, in coordinates, is defined algebraically (i.e. by rational functions, in other words by quotients of polynomials).*

A complex projective algebraic variety (i.e. the zero locus into the complex n-th projective space of a set of homogeneus polynomials) whose points form a group (necessarly abelian) whose composition law is algebraically defined is called an *abelian variety* over $\mathbb{C}$. Therefore the previous proposition says that elliptic curves are abelian varieties. It turns out that all abelian varieties (over $\mathbb{C}$) of dimension 1 are elliptic curves.

We begin to do Number Theory when the equation of an elliptic curve has only rational coefficients (ore more generally number fields, i.e. finite degree extensions of $\mathbb{Q}$). The next definition give us a little bit of language on that.

Definition If K is any subfield of $\mathbb{C}$ and $g_2, g_3 \in K$ we say that E is defined over K. If E is defined over K and $K \subseteq L \subseteq \mathbb{C}$ is an intermediate field we define

$$E(L) = \{\, (x,y) \in \mathbb{C}^2 \colon y^2 = 4x^3 - g_2 x - g_3, \text{ s.t. } x, y \in L \,\} \cup \{\, \infty \,\} = E(\mathbb{C}) \cap (L^2 \cup \{\, \infty \,\})$$

the set of L-rational point.

In particular one see, by the expression of the sum in terms of coordinates, that $E(L)$ is a subgroup of $E(\mathbb{C})$, namely if we start with two points with coordinate belonging to L, then the coordinates of their sum still belong to L.

The next theorem gives the structure of the group of $(\mathbb{Q})$-rational points of E: it gives an example of how one get important arithmetic invariants working with elliptic curves.

Theorem *Let E be an elliptic curve defined over $\mathbb{Q}$, then $E(\mathbb{Q})$ is a finitely generated abelian group, hence isomorphic to $E(\mathbb{Q})_{\text{tors}} \times \mathbb{Z}^r$, where $E(\mathbb{Q})_{\text{tors}}$ denotes the $\mathbb{Q}$-rational torsion points (i.e. points of finite order, namely points $P \in E(\mathbb{Q})$ such that $nP = \infty$ for some $n \in \mathbb{N}_{>0}$), and $r \in \mathbb{N}_{\geq 0}$ is called the (algebraic) rank of E.*

The rank of an elliptic curve in particular is an important arithmetic invariant and is the subject of important conjectures, as e.g. the Birch and Swinnerton-Dyer conjecture (one of the Millennium problems) who relates the algebraic rank of an elliptic curve with the so called analytic rank, that is the order of vanishing of the L-function attached to it (it is a complex analytic function that one can attach to the an elliptic curve) at $s = 1$. In fact, there is no known finite time algorithm to compute the algebraic rank of a general elliptic curve.

3 Modular Curves

An alternative way to describe an elliptic curve is to see it as Riemann surface, i.e. as a "complex differential manifold of dimension 1". A reference for this material is [1].

Definition A Riemann surface is a topological space X such that locally at any point $x \in X$ there is an homeomorphism $\varphi\colon U \to V$ of an open neighborhood U of x with an open subset V of $\mathbb{C}$ and such that the *local charts* are compatible, in the sense that the *change of coordinates* $\psi \circ \varphi^{-1}\colon \varphi(U \cap U') \to \psi(U \cap U')$ are biholomorphic maps (holomorphic and bijective). Here $\varphi\colon U \to V$ and $\psi\colon U' \to V'$ are two local charts such that $U \cap U' \neq \emptyset$. The set of local charts of a Riemann surface is called a *complex atlas*.

Example The most trivial example of Riemann surface is of course $\mathbb{C}$ itself, with the identity as the unique chart; another easy example is the so called *Riemann Sphere*, that is a sphere covered by two local charts: the two stereographic projections, one defined on the whole sphere without the North pole, the other on the sphere without the South one. Note that the latter example is a compact Riemann surface, that is topologically the compactification with one point of the first one.

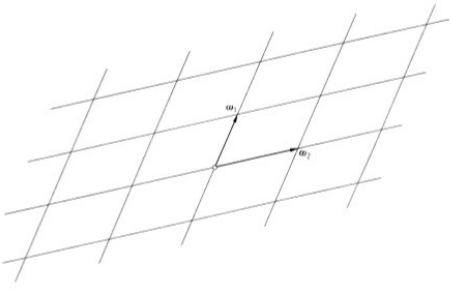


Figure 3. A lattice in the complex plane. This picture has been taken from [3] (Figure 1.1).

(a) (b)

Figure 4. A fundamental parallelogram (a): up to the identification of the opposite sides it represent the quotient space $\mathbb{C}/\Lambda$. Once identified the opposite sides it becomes a topological torus (b). The picture (a) has been taken from [3] (Figure 1.2) and (b) from Wikipedia.

Definition-Proposition A lattice in $\mathbb{C}$ is a free subgroup of rank 2, i.e. a subset of the form $\mathbb{Z}\,\omega_1 \oplus \mathbb{Z}\,\omega_2$, where $\omega_1, \omega_2 \in \mathbb{C}$, such that $\mathrm{Im}(\omega_2/\omega_1) > 0$ (see Figure 3). The quotient group

$$\mathbb{C}/\Lambda = \{\, x + \Lambda \colon x \in \mathbb{C} \,\}$$

endowed with the quotient topology can be endowed with a complex atlas. The resulting Riemann surface is called a complex torus (it is indeed topologically a torus, as Figure 4 explains).

The important result is that, using the Weierstrass $\wp$ fuction, any complex torus can be embedded into the complex projective plane in a biholomorphic way as an elliptic curve and all elliptic curves are obtained in this way, hence as Riemann surface the elliptic curves and complex tori are the same thing. Therefore classify elliptic curves up to biholomorphism is the same thing as classify complex tori. Let us do it.

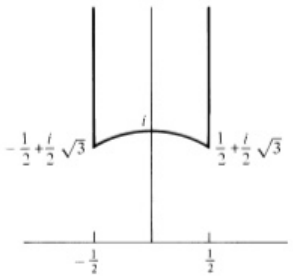


Figure 5. A fundamental domain for $\mathrm{SL}_2(\mathbb{Z})$. In formulas it is $D = \{\, \tau \in \mathbb{H} \colon |\tau| \geq 1, |\mathrm{Im}(\tau)| \leq 1/2 \,\}$. This picture has been taken from [2] (Figure III.1).

Proposition *Two complex tori $\mathbb{C}/\Lambda$ and $\mathbb{C}/\Lambda'$ are biholomorphic if and only if Λ and Λ' are homotetic, that is $\Lambda' = \varepsilon\Lambda$ for some $\varepsilon \in \mathbb{C}^\times$.*

In particular, writing $\tau = \omega_2/\omega_1$ for a lattice $\Lambda = \mathbb{Z}\omega_1 \oplus \mathbb{Z}\omega_2$, any complex torus is biholomorphic to a torus $E_\tau = \mathbb{C}/\Lambda_\tau$, where $\Lambda_\tau = \mathbb{Z} \oplus \mathbb{Z}\tau$ and $\mathrm{Im}(\tau) > 0$. Moreover:

Corollary *E_τ is biholomorphic to $E_{\tau'}$ if and only if there are $a, b, c, d \in \mathbb{Z}$ with $ad - bc = 1$ such that $\tau' = \frac{a\tau+b}{c\tau+d}$.*

With a more sophisticate language this means that classify all elliptic curves up to biholomorphism is equivalent to classify the numbers τ that belong to the complex upper halfplane

$$\mathbb{H} = \{\, \tau \in \mathbb{C} \colon \mathrm{Im}(\tau) > 0 \,\}$$

up to the (left) action of the group

$$\mathrm{SL}_2(\mathbb{Z}) = \left\{ \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M_2(\mathbb{Z}) \colon \det(\gamma) = 1 \right\}$$

given by the formula

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot \tau = \frac{a\tau + b}{c\tau + d}.$$

We will denote the set of orbits by $Y(1) = \mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$. One can prove that $Y(1)$ has the structure of a Riemann surface, that we call an *(open) Modular Curve*. We can see how it works topologically: Figure 5 represents a so called *fundamental domain*, i.e. a connected subset of the complex plane that (outside the borders) is in bijection with the orbits of the action. Giving to $Y(1)$ the quotient topology is therefore equivalent to glue the left and right borders of the fundamental domain.

Other (open) modular curves can be defined as the quotient $\Gamma\backslash\mathbb{H}$ of $\mathbb{H}$ by the induced action of some particular a subgroups Γ of $\mathrm{SL}_2(\mathbb{Z})$, called *congruence subgroups*. We are in particular interested in congruence subgroups of the form

$$\Gamma_0(N) = \left\{ \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) \colon c \equiv 0 \mod N \right\}, \quad N \in \mathbb{N}_{>0}.$$

The modular curve $Y_0(N) = \Gamma_0(N)\backslash\mathbb{H}$ can be interpreted as a classifying space, too: it parametrizes the couples (E, C), where E is an elliptic curve and C is a cyclic subgroup of E of order N.

Another important features of modular curves is that they can be compactified with a finite number of points, called *cusps*, in a canonical way, giving rise to a compact Riemann surface, that we call a *closed Modular Curve*. We will denote the compactifications by the letter X, so e.g. $X_0(N)$ is the compactification of $Y_0(N)$. The importance of $X_0(N)$ is that it can be embedded (biholomorphically) into the complex projective plane as an algebraic curve defined over $\mathbb{Q}$ (i.e. it has an equation with all rational coefficients). Hence in the rest we will treat $X_0(N)$ as such a curve.

4 Modularity

We can now state the definition of modularity for an elliptic curve. Heuristically an elliptic curve is modular if it is a quotient of a modular curve.

Definition Let E an elliptic curve defined over $\mathbb{Q}$ of conductor N_E. We say that E is *modular* if there is a surjective morphism $\pi \colon X_0(N_E) \to E$ of algebraic curves defined over $\mathbb{Q}$.

The integer N_E occurring above, the conductor of E, is another arithmetic invariant attached to an elliptic curve, that encodes its *reduction type*.

Modularity Theorem Any elliptic curve E defined over $\mathbb{Q}$ is modular.

The FLT is a consequence of this theorem, that was conjectured by Shimura, Taniyama and Weil between 1957 and 1967. It was proved for semistable elliptic curves by Wiles and Taylor in 1993-1995 (and this case is enough for FLT), and it was established in general only in 2001 in a joint paper by Breuil, Conrad, Diamond, Taylor and Richard.

4.1 Tate Module

To explain how the Modularity theorem for semistable elliptic curves implies FLT we need to introduce another important arithmetic object attached to an elliptic curve E, the so called Tate Module. For simplicity let E be defined over $\mathbb{Q}$. In this paragraph we assume that the reader has some familiarity with l-adic numbers and the notion of Galois group.

Let l a rational prime, $r \in \mathbb{N}_{>0}$. Then the set of l^r-torsion points

$$E[l^r](\mathbb{C}) = \{\, P \in E(\mathbb{C}) \colon l^r P = \infty \,\}$$

is a finite abelian group isomorphic to $\mathbb{Z}/l^r\mathbb{Z} \times \mathbb{Z}/l^r\mathbb{Z}$, as it can be easily seen from the complex torus description of E. Therefore the inverse limit

$$T_l E = \varprojlim_r E[l^r](\mathbb{C}) \cong \mathbb{Z}_l \times \mathbb{Z}_l$$

has a natural structure of $\mathbb{Z}_l$ module.

Since E is defined over $\mathbb{Q}$, then the torsion points have coordinate in $\bar{\mathbb{Q}}$ and hence the absolute Galois group of $\mathbb{Q}$ (i.e. the group of field automorphism of $\bar{\mathbb{Q}}$ fixing $\mathbb{Q}$) acts on $E[l^r](\mathbb{C})$ and therefore on $T_l E$ (where $\sigma \in \mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$ acts on a point $P \in E(\bar{\mathbb{Q}})$ coordinatewise).

Defining $V_l E = T_l E \otimes \mathbb{Q}_l \cong \mathbb{Q}_l \times \mathbb{Q}_l$, and letting $\mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$ act on $V_l E$ by $\sigma \otimes 1$, we obtain an l-adic representation of $\mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$.

Definition Let K be a finite extension of $\mathbb{Q}_l$, an l-adic Galois representation is a K-vector space V endowed with a continuous action of $\mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$. Equivalently, by the choice of a K-basis of V, we may see the representation V as a continuous group homomorphism

$$\rho \colon \mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}) \longrightarrow \mathrm{GL}(V) \cong \mathrm{GL}_n(K)$$

where $n = \dim_K V$ and $\rho(\sigma)x = \sigma \cdot x$, for $\sigma \in \mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$, $x \in V$.

The first example of an l-adic Galois representation is, as we said above, the Tate module of an elliptic curve. Note that, if we denote by

$$\rho_{E,l} \colon \mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_2(\mathbb{Q}_l)$$

the attached homomorphism, the very definition of Tate module shows that $\mathrm{Im}(\rho_{E,l}) \subseteq \mathrm{GL}_2(\mathbb{Z}_l)$ and therefore we may see it as

$$\rho_{E,l} \colon \mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_2(\mathbb{Z}_l)$$

and reducing the coefficients of the matrices modulo the maximal ideal of $\mathbb{Z}_l$ we get the residue representation

$$\bar{\rho}_{E,l}\colon \mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_2(\mathbb{F}_l),$$

that can be tought as an $\mathbb{F}_p$-vector space together with a continuous action of the absolute Galois group of $\mathbb{Q}$.

In modern Number Theory l-adic Galois representation are extremely important: we may attach l-adic representations to a lot of arithmetic objects and many arithmetic properties of them can be read by their attached Galois representations. For instance the modularity theorem can be expressed using Galois representations.

Proposition *An elliptic curve E is modular if and only if there exist a modular newform $f \in S_2\big(\Gamma_0(N_E)\big)$ of level N_E such that for some prime l the attached representation $\rho_{f,l}$ is equivalent to $\rho_{E,l}$.*

We don't introduce modular forms as in this context they are only an auxiliary tool, the interested reader can find a nice introduction to them in [1]. We remark only that we may attach a field K_f, finite extension of $\mathbb{Q}_l$, to any modular form f as in the proposition and a Galois representation

$$\rho_{f,l}\colon \mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_2(K_f)$$

with image in $\mathrm{GL}_2(\mathcal{O}_f)$, where $\mathcal{O}_f$ is the ring of integers of K_f. It still make sense therefore to consider the residual representation

$$\bar{\rho}_{f,l}\colon \mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_2(\kappa_f),$$

where by κ_f we denote the residue field of $\mathcal{O}_f$ (that is a finite field, finite extension of $\mathbb{F}_l$).

4.2 Modularity $\Longrightarrow$ FLT

We now describe how this theorem implies FTL. Fist of all note that it is enough to prove it for $n = p$ an odd prime, as if $n = mp$ and (F_n) has a solution (a, b, c), then (a^m, b^m, c^m) is a solution of (F_p). Suppose by contraddiction that (F_p) has a nontrivial primitive (i.e. a, b, c have no common factor) solution, then also the equation

$$(F'_p) \qquad a^p + b^p + c^p = 0$$

has a nontrivial primitive solution. Let $(a, b, c) \in \mathbb{Z}^3$ be such a solution. Without loss of generality we may assume that $a \equiv -1 \mod 4$ and b is even. Indeed we know that exactly one of a, b and c must be even, as if two of them were even the third would be even, too, but then they would have 2 as common factor. Thus at least two of them must be odd, and therefore the third one is forced to be even; permuting a, b and c fix the even one to be b. Now since a is odd, either $a \equiv 1 \mod 4$, or $a \equiv -1$ but in the latter case we may replace the triple (a, b, c) with $(-a, -b, -c)$, that is still a nontrivial primitive solution of (F'_p), and hence we may assume that $a \equiv 1 \mod 4$. We attach to a triple as above a particular elliptic curve, known as *Frey curve*,

$$E_{(a^p,b^p,c^p)}\colon y^2 = x(x - a^p)(x + b^p)$$

that can be proven to be semistable (and therefore modular by Wiles' Theorem) and to have discriminant and conductor

$$\Delta = 2^{-8}(abc)^{2p}, \qquad N = \prod_{l \text{ prime},\, l|abc} l.$$

By modularity there exist a newform f of level N such that $\rho_{E_{(a^p,b^p,c^p)}}$ is equivalent to $\rho_{f,l}$. But, since it can be proven that $\bar{\rho}_{E_{(a^p,b^p,c^p)}}$ has some special properties (it is absolutely irreducible, unramified outside $2l$ and flat at l), then $\bar{\rho}_{f,l}$ satisfies the hypotesis of the *epsilon conjecture*, proven by Ribet (see [6]); hence the residual representations $\bar{\rho}_{E_{(a^p,b^p,c^p)}}$ and $\bar{\rho}_{g,l}$ are equivalent, for a suitable newform $g \in S_2(\Gamma_0(2))$. This fact leads to a contraddiction, finally proving Fermat's Last Theorem, as the dimension of $S_2(\Gamma_0(2))$ is known to be 0, thus such a g does not exist.

5 Further readings on FLT

For an hystorical introduction to FLT see the first chapter of [5]. The rest of the book discuss in detail what was known until that Frey suggested, few years after the publication of this book, a possible link between Fermat's Last Theorem and the Shimura-Taniyama-Weil conjecture on modularity.

For a short and elementary reading that gives the flavor of Kummer's approach to FLT in a special case, see the first chapter of [4].

For a brief account of Wiles proof of FLT see [8]. The book that contains this article is an introduction to the main tools used into the proof, all discussed separately for their own sake in different articles by authors expert in that particular subject.

References

[1] F. Diamond and J. Shurman, "A First Course in Modular Forms". Springer-Verlag, 2005.

[2] N. Koblitz, "Introduction to Elliptic Curves and Modular Forms, 2nd edition". Springer-Verlag, 1993.

[3] S. Lang, "Elliptic Functions". Springer-Verlag, 1987.

[4] D. Marcus, "Number Fields, 2nd edition". Springer-Verlag, 2018.

[5] P. Ribenboim, "13 Lectures on Fermat's Last Theorem". Springer-Verlag, 1979.

[6] K.A. Ribet, *On modular representations of* $\mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$ *arising from modular forms.* Inventiones Mathematicae 100 (1990), 431–476.

[7] J.H. Silverman, "The Arithmetic of Elliptic Curves, 2nd edition". Springer-Verlag, 2009.

[8] G. Steavens, *An overview of the proof of Fermat's Last Theorem.* in: Modular Forms and Fermat's Last Theorem, ed. by G. Cornell, J.H. Silverman and G. Steavens, Springer-Verlag, 1997.

Embeddings of spaces with multiweighted derivatives and their applications

Zhanar Keulimzhayeva (*)

Abstract. The analysis of function spaces is particularly relevant in several areas of Mathematics such as differential and integral equations. Among function spaces, the weighted ones turn out to be suitable for the analysis of boundary value problems with various types of singularities. This seminar presents a brief overview of with weight functional space of the Kudryavtsev type, called the space with multiweighted derivatives. In the first part of the talk, we will define space with multiweighted derivatives and consider the behavior of the function at the boundary of the investigated space. Following the ideas of L.D. Kudryavtsev, we will define the boundary values of a function and of its derivatives at the singular point. In addition, we will give necessary and sufficient conditions for weight functions in order that each function of the space is stabilized to some unique polynomial at zero, and will provide estimates for the rate of stabilization to a polynomial. Next, we will introduce functionals that depend on the boundary values at the singular point and that are equivalent to the norm of the space. In the very last part of the talk, we will introduce necessary and sufficient conditions for weight functions so that continuous and compact embeddings between spaces of multiweighted derivatives hold. Moreover, we will show conditions on the weight functions in order that the inequality of the Nikolâskii-Lizorkin-Kudryatsev type is valid.

Introduction

The idea to study function spaces with the purpose to apply them to different problems concerning differential equations appeared in papers by S. L. Sobolev in the thirties.

The Sobolev space $\mathbf{W_p^m(\Omega)}$. We define a functional $\|\cdot\|_{m,p}$, where m is a positive integer and $1 \leq p \leq \infty$, as follows:

$$\|u\|_{m,p} = \left(\sum_{0 \leq \alpha \leq m} \|D^\alpha u\|_p^p \right)^{\frac{1}{p}}, \quad p \in [1, \infty), \tag{0.1}$$

(*)Department of Higher Mathematics, S. Seifullin Kazakh Agro Technical University - 010000 Republic of Kazakhstan - Nur-Sultan, st. Y. Altynsarin, 2. E-mail: zh.keulimzhayeva@mail.ru . Seminar held on 16 November 2021.

$$\|u\|_{m,\infty} = \max_{0\le|\alpha|\le m} \|D^\alpha u\|_\infty, \ \ p=\infty, \tag{0.2}$$

where $D^\alpha u$ is the weak (or distributional) partial derivative.

$$W_p^m(\Omega) \equiv \{u \in L^p(\Omega) : D^\alpha u \in L^p(\Omega) \ for \ 0 \le |\alpha| \le m\}, \ \Omega \in \mathbb{R}^n.$$

Equipped with the appropriate norm (1) or (2) this is called **Sobolev space** over Ω.

From this time the theory of Sobolev spaces has been developed to be a very powerful instrument for solving boundary value problems of differential equations.

In view of various types of questions connected to differential equations a great number of new function spaces has appeared and this, in turn, has implied the appearance of new directions in Analysis such as the theory of generalized functions, the embedding theorems and embedding compactness of function spaces the problem of traces and the variational methods for studying boundary value problems.

In particular, the developed methods have contributed crucially to the development of the theory of regular differential equations. Singular differential equations are less studied than regular differential equations.

These reasons implied that such notions as "weight" and "weighted space" were introduced.

Since weight functions, in fact, can take care of some problems connected to singularities, they began to be used widely. Moreover, they became an essential and natural part of the theory of function spaces with their applications in the theory of singular differential equations.

The first result of embedding theorem type for weighted spaces was obtained in the paper [1] by V.I. Kondrashev in 1938. A systematic study of weighted spaces was started at the end of the fifties in the papers by L. D. Kudryavtsev (see e. g. [2]). In these papers, L. D. Kudryavtsev built his theory on the standard variational method for solving elliptic boundary value problems.

In work (see [3] L.D. Kudryaveev considered a space $L^n_{p,\gamma} = L^n_{p,\gamma}(1,+\infty)$ of functions $f : (1,+\infty) \to \mathbb{R}$, which on the interval $(1,+\infty)$ have a generalized derivative of n:th order with the finite semi-norm

$$\|f\|_{L^n_{p,\gamma}} = \|x^\gamma f^{(n)}\|_p,$$

where $\gamma \in \mathbb{R}$, $1 \le p \le \infty$ and n is a natural number.

It was shown that if $\gamma < n - \frac{1}{p}$, then for functions from this class we have that neither the function f nor its derivatives $f^{(k)}$, $k = 1, 2, \ldots, n-1$, in general, have limit values when $x \to +\infty$. If $\gamma > n - \frac{1}{p}$, then the $(n-1)$:th order derivative has always a limit value when $x \to +\infty$, but the limit values of less order derivatives are, in general, infinite.

Therefore, at infinity we have such a singularity that can not be handled even with a weight. This leads to that it is necessary to formulate boundary conditions at infinity in a different manner. Hence, L.D. Kudryavtsev proposed to interpret boundary conditions for

differential equations as coefficients of a polynomial, to which a solution becomes stable in the sense explained below. For the case $\gamma > n - \frac{1}{p}$ L. D. Kudryavtsev proved the existence of a unique polynomial $P_{n-1} = a_0 + a_1 x + \cdots + a_{n-1}x^{n-1}$ such that

$$\lim_{x\to\infty} [f(x) - P_{n-1}(x)]^{(k)} = 0,\ k = 0, 1, \ldots, n-1.$$

This stability condition means that the function f and its derivatives up to $(n-1)$:th order go to the polynomial P_{n-1} and to its corresponding derivatives at infinity. Consequently, the coefficients of this polynomial can be considered as "boundary values" at infinity of this function.

In 1991/1992, at a scientific seminar under his supervision, R. Oinarov expanded the idea of L. D. Kudryavtsev and set the task of studying a multiweighted space $W^n_{p,\bar{\rho}}$, showing ways to solve boundary value problems at singular finite and infinite points in simple differential equations. Later this space was called the multiweighted derivatives space.

At the beginning, when the weights were power functions, that is $\rho_i = t^{\alpha_i},\ i = 1, 2, ..., n,$ numerous results were obtained by A.O. Bayarystanov, B.L.Baidel'dinov, A. A. Kalybay, Z. T. Abdykalykova [4]-[6].

For example, B. L. Baidel'dinov [4] showed how to represent a boundary value problem in the space $W^n_{p,\bar{\alpha}}$ depending on the singularity of a simple differential equation of symmetric order n.

Recently, A. A. Kalybay [7] showed that the generalized Cauchy problem is well-posed for a simple differential equation with a singularity of the n:th order, regardless of the singularity of the equation at zero. This result was achieved due to the properties of space $W^n_{p,\bar{\alpha}}$.

These works show that the space $W^n_{p,\bar{\rho}}$ has big potential.

Therefore, the study of this space $W^n_{p,\bar{\rho}}$ in general, is an important urgent problem.

For general weights $\bar{\rho} = \{\rho_1, \rho_2, ..., \rho_n\}$, we can say that space $W^n_{p,\bar{\rho}}$ has not been studied.

1 The space with multiweighted derivatives

1.1 Definition of a space with multiweighted derivatives

Let $I = (0, 1)$, n - be a natural number, $\rho_i : I \to \mathbb{R}$, $i = 1, 2, ..., n$, be positive functions integrable on $I_\delta = [\delta, 1]$, $\forall \delta \in I$, such that

$$\rho_i^{-1} \equiv \frac{1}{\rho_i} \in L_1(I_\delta),\ i = 1, ..., n-1, \quad \rho_n^{-1} \in L_{p'}(I_\delta),\ 1 < p' < \infty. \tag{1.1}$$

For a function $f : I \to \mathbb{R}$ we assume that

$$D^0_{\bar{\rho}} f(x) \equiv f(x), D^k_{\bar{\rho}} f(x) = \rho_k(x) \frac{d}{dx} D^{k-1}_{\bar{\rho}} f(x), \quad x \in I, \quad k = 1, ..., n.$$

Suppose that the functions $D^k_{\bar{\rho}} f(x)$, $k = 0, 1, ..., n-1$ are absolutely continuous on the interval $[\delta, 1]$, $\forall \delta \in I$; then $D^n_{\bar{\rho}} f(x)$ exists for almost every $x \in I$. We call this operation $D^k_{\bar{\rho}} f$ the $\bar{\rho}$- multiweighted derivative of f of order k, $k = 1, ..., n$.

Let $W^n_{p,\bar\rho}(I)$ be the set of all functions that have $\bar\rho$- multiweighted derivatives up to order n, $n \geq 1$, inclusive on the interval I. On the set $W^n_{p,\bar\rho}(I)$, consider the functional

$$\|f\|_{W^n_{p,\bar\rho}(I)} = \left\|D^n_{\bar\rho} f\right\|_{p,I} + \sum_{i=0}^{n-1} \left|D^i_{\bar\rho} f(1)\right|, \tag{1.2}$$

where $\|\cdot\|_{p,I}$ is the standard norm of the space $L_p(I)$. This functional is well defined and provides a norm on $W^n_{p,\bar\rho}(I)$.

Theorem 1.1 *Let $1 < p < \infty$. The space $W^n_{p,\bar\rho}(I)$ is a Banach space.*

In the case of $\rho_i \equiv 1,\ i = 1, 2, ..., n-1$, and $\rho_n \equiv \varphi$, it turns into the Kudryavtsev space $W^n_{p,\varphi}(I)$ with the norm

$$\|f\|_{W^n_{p,\varphi}(I)} = \left\|\varphi f^{(n)}\right\|_{p,I} + \sum_{i=0}^{n-1} \left|f^i(1)\right|,$$

which were well studied in [8] and [9] in connection with boundary value problems for degenerate elliptic equations.

And, in the case of $\rho_i(t) = t^{\alpha_i},\ i = 1, 2, ..., n,\ \alpha_i \in \mathbb{R}$, the space $W^n_{p,\bar\alpha}(I)$ was studied in details (see, for example, [10], [11], [12]).

1.2 The space $W^n_{p,\bar\rho}(I)$ and its properties

For $0 \leq s \leq x < 1$ and $i, j = 0, 1, ..., n-1$, we define the following functions $K_{j,i+1}$ and $\bar K_{j,i+1}$:

$$K_{j,i+1}(x,s) = (-1)^{j-i} \int_s^x \rho_j^{-1}(t_j) \int_s^{t_j} \rho_{j-1}^{-1}(t_{j-1}) ... \int_s^{t_{i+2}} \rho_{i+1}^{-1}(t_{i+1}) dt_{i+1} dt_{i+2} ... dt_j$$

and

$$\bar K_{j,i+1}(x,s) = \int_s^x \rho_j^{-1}(t_j) \int_{t_j}^x \rho_{j-1}^{-1}(t_{j-1}) ... \int_{t_{i+2}}^x \rho_{i+1}^{-1}(t_{i+1}) dt_{i+1} dt_{i+2} ... dt_j$$

for $j > i$, $K_{j,i+1}(x,s) \equiv \bar K_{j,i+1}(x,s) \equiv 1$ for $j = i$ and $K_{j,i+1}(x,s) \equiv \bar K_{j,i+1}(x,s) \equiv 0$ for $j < i$.

Let us note that the system of functions $\{K_{i,1}(1,t),\ t \in I\}_{i=0}^{n-1}$ is a system of linearly independent solutions of the homogeneous equation

$$D^n_{\bar\rho} f(t) = 0, \quad t \in I.$$

In [13] L. D. Kudryavtsev introduced the concept of stabilization of a function at infinity belong to $W^n_{p,\varphi}(I)$ to an algebraic polynomial of the $(n-1)$th order. The coefficients of this polynomial can be considered as "boundary values" at infinity of this function.

Let us consider the polynomials with respect to the system of functions $\{K_{i,1}(1,t)\}_{i=0}^{n-1}$, $P_n(t) \equiv P_n(\bar\rho, t) = \sum\limits_{i=0}^{n-1} a_i K_{i,1}(1,t)$, $t \in I$, where $a_i,\ i = 0, 1, ..., n-1$, are real numbers.

Definition 1.2 We say that a function $f \in W^n_{p,\bar{\rho}}(I)$ becomes stable at zero to the polynomial $P_n(t) \equiv P_n(t,f)$, if

$$\lim_{t\to 0^+} D^i_{\bar{\rho}}[f(t) - P_n(t,f)] = 0, \quad i = 0,1,...,n-1. \tag{1.3}$$

From (1.3) it follows that if $f \in W^n_{p,\bar{\rho}}(I)$ becomes stable at $t = 0$ to the polynomial $P_n(t,f)$, then the coefficients of the polynomial $P_n(t,f)$ can be consequently defined from the relations

$$a_i(f) = \lim_{t\to 0^+}[D^i_{\bar{\rho}} f(t) - \sum_{j=i+1}^{n-1} a_j(f) K_{j,i+1}(1,t)],\ i = n-1, n-2, ..., 0. \tag{1.4}$$

The values $a_i(f),\ i = 0,1,...,n-1$, can be interpreted as "boundary values" of the function $f \in W^n_{p,\bar{\rho}}(I)$ and of its weighted derivatives at zero.

Theorem 1.3 *Let $1 < p < \infty$. Each function $f \in W^n_{p,\bar{\rho}}(I)$ becomes stable at zero to the polynomial $P_n(t,f)$ if and only if the following integrals*

$$\int_0^t \bar{K}^{p'}_{n-1,i+1}(t,s)\rho_n^{-p'}(s)ds, \quad i = 0,1,...,n-1, \quad t \in I \tag{1.5}$$

converge. Moreover, the following identities

$$D^i_{\bar{\rho}} f(t) = \sum_{j=i}^{n-1} a_j(f) K_{j,i+1}(1,t) + \int_0^t \bar{K}_{n-1,i+1}(t,s)\rho_n^{-1}(s) D^n_{\bar{\rho}} f(s)ds \tag{1.6}$$

hold for $i = 0,1,...,n-1$.

1.3 Estimates of the rate of convergence of a function to a polynomial

We now set $r_{n,i}(t) = \left\|\rho_n^{-1}(\cdot)\bar{K}_{n-1,i+1}(t,\cdot)\right\|^{-1}_{p',(0,t)}$.

Theorem 1.4 *Let $1 < p < \infty$ and assume that (1.5) holds. Then for any $f \in W^n_{p,\bar{\rho}}(I)$ there exists a unique polynomial $P_n(t)$ such that*

$$\text{(i)} \quad \sup_{0<t<x} \left|r_{n,i}(t) D^i_{\bar{\rho}}\left[f(t) - P_n(t)\right]\right| \le \left\|D^n_{\bar{\rho}} f\right\|_{p,(0,x)}, \quad i = 0,1,...,n-1; \tag{1.7}$$

(ii) *if for $0 \le i \le n-1$ there exists a locally integrable function $u_i(\cdot) \ge 0$ such that*

$$\sup_{0<x<1} \int_x^1 u_i^p(t) \left(\int_0^x \rho_n^{-p'}(s)\bar{K}^{p'}_{n-1,i+1}(x,s)ds\right)^{p-1} dt < \infty, \tag{1.8}$$

then the estimate

$$\left\|u_i D^i_{\bar{\rho}}\left[f - P_n(t)\right]\right\|_{p,I} \le C \left\|D^n_{\bar{\rho}} f\right\|_{p,I} \tag{1.9}$$

holds and in particular, if (1.8) holds for $u_i = \rho_i^{-1}$, *then*

$$\left\| \frac{d}{dt} D_{\bar{\rho}}^{i-1} \left[f - P_n(t) \right] \right\|_{p,I} \leq C \left\| D_{\bar{\rho}}^{n} f \right\|_{p,I}. \tag{1.10}$$

1.4 Equivalent norms of the space $W_{p,\bar{\rho}}^n(I)$

Let N_1, N_0 be two subsets of $N = \{0, 1, ..., n-1\}$ such that $N_1 \bigcap N_0 = \emptyset$ and $N_1 \cup N_0 = N$. Let $N_1 = \{i_1, i_2, ..., i_k\}$, $N_0 = \{j_1, j_2, ..., j_m\}$.

Theorem 1.5 *Let* $1 < p < \infty$ *and assume that (1.5) holds.*

i) *if* $N_1 = \emptyset$ *and* $N_0 = N$, *then the functional*

$$\|f\|_{W_{p,\bar{\rho}}^n(I)}^1 = \left\| D_{\bar{\rho}}^{n} f \right\|_{p,I} + \sum_{i=0}^{n-1} |a_i(f)| \tag{1.11}$$

is equivalent to the norm $\|f\|_{W_{p,\bar{\rho}}^n(I)}$;

ii) *if* $N_1 \neq \emptyset$ *and* $N_0 \neq \emptyset$, *then the functional*

$$\|f\|_{W_{p,\bar{\rho}}^n(I)}^2 = \left\| D_{\bar{\rho}}^{n} f \right\|_{p,I} + \sum_{\mu=1}^{k} |D_{\bar{\rho}}^{i_\mu} f(1)| + \sum_{\lambda=1}^{m} |a_{j_\lambda}(f)| \tag{1.12}$$

is equivalent to the norm $\|f\|_{W_{p,\bar{\rho}}^n(I)}$.

2 Embeddings between spaces with multiweighted derivatives and their applications

2.1 Embeddings between spaces with multiweighted derivatives

Along with the space $W_{p,\bar{\rho}}^n(I)$ we will consider the space $W_{q,\bar{\rho}}^k(\tau_k, I)$ with the norm

$$\|f\|_{W_{q,\bar{\rho}}^k(\tau_k,I)} = \left\| D_{\bar{\rho},\tau_k}^{k} f \right\|_{q,I} + \sum_{i=1}^{k-1} |D_{\bar{\rho}}^{i} f(1)|,$$

where $1 \leq k < n-1$ and $D_{\bar{\rho},\tau_k}^{k} f(x) \equiv \tau_k(x) \frac{d}{dx} D_{\bar{\rho}}^{k-1} f(x),\ x \in I$.

We will investigate the embeddings

$$W_{p,\bar{\rho}}^n(I) \hookrightarrow W_{q,\bar{\rho}}^k(\tau_k, I), \tag{2.1}$$

that is, the fulfillment of the inequality

$$\|f\|_{W_{q,\bar{\rho}}^k(\tau_k,I)} \leq C \|f\|_{W_{p,\bar{\rho}}^n(I)}, \quad \forall f \in W_{p,\bar{\rho}}^n(I). \tag{2.2}$$

The best constant C, for which (2.2) holds, is called the operator norm of the embedding $E : W_{p,\bar{\rho}}^n(I) \mapsto W_{q,\bar{\rho}}^k(\tau_k, I)$, and is denoted by $\|E\|$ i.e., we set $C = \|E\|$.

Now, let us find the conditions under which the continuous compact embedding (2.1) takes place.

We assume that

$$B_1 = \max_{k\le j\le n-1} \left\|\tau_k(\cdot)\rho_k^{-1}(\cdot)K_{j,k+1}(1,\cdot)\right\|_{q,I},$$

$$B_2(z) = \left(\int_z^1 \rho_n^{-p'}(x)\left(\int_0^z K_{n-1,k+1}^q(x,s)\tau_k^q(s)\rho_k^{-q}(s)ds\right)^{\frac{p'}{q}} dx\right)^{\frac{1}{p'}},$$

$$B_2 = \sup_{0<z<1} B_2(z),\ B = \max\{B_1, B_2, 1\}.$$

Theorem 2.1 *Let be $1 < p \le q < \infty$, $1 \le k < n-1$. Then the embedding (2.1)*

(i) *is continuous if and only if $B < \infty$, and $\|E\| \approx B$, where $\|E\|$- is the norm of the embedding operator (2.1);*

(ii) *is compact if and only if $B < \infty$ and*

$$\lim_{z\to 0} B_2(z) = 0.$$

2.2 Inequality of the Nikolskii-Lizorkin-Kudryavtsev type in the space $W^n_{p,\bar{\rho}}(I)$

S. I. Nikolski and P.I. Lizorkin [14] established the inequality

$$\|f\|_p \le C\left(\sum_{i=0}^{k-1}\left|f^{(i)}(0)\right| + \sum_{j=0}^{m-1}\left|f^{(j)}(T)\right| + \left\|f^{(n)}\right\|_{p,\alpha,\beta}\right),$$

for functions $f \in W^n_{p,\alpha,\beta}[0,T]$, $0 < T < \infty$, where $k+m=n,\ 1 < p < \infty$,

$$\|f\|_{p,\alpha,\beta} = \left(\int_0^T \left|t^\alpha (T-t)^\beta f^{(n)}(t)\right|^p dt\right)^{\frac{1}{p}},$$

$\|f\|_p = \|f\|_{p,0,0}$. The constant $C > 0$ does not depend on the function f and the indicators α and β satisfy the "weak degeneracy" condition ($\max\{\alpha,\beta\} < 1/p'$) that ensures the existence of finite limiting values.

In the paper [15], L. D. Kudryavtsev established an analogue of this inequality for the case of an unbounded interval $(1,+\infty)$:

$$\left|f^{(s)}(t)\right| \le \left(\sum_{\nu=1}^{k}\left|f^{(i_\nu)}(1)\right| + \sum_{\mu=1}^{m}\left|a_{j_\mu}\right| + \left\|f^{(n)}\right\|_{p,\alpha}\right) t^{n-s-1}, \tag{2.3}$$

where $0 \le s < n$ and the indices $0 \le i_1 < i_2 < ... < i_k \le n-1$ and $0 \le j_1 < j_2 < ... < j_m \le n-1$, $k+m=n$, and the indices $0 \le \bar{j}_1 < \bar{j}_2 < ... < \bar{j}_k \le n-1$ are additional to indices $j_1, j_2, ..., j_m$ and satisfy the Polya condition

$$i_1 \le \bar{j}_1,\ i_2 \le \bar{j}_2,\ ..., i_k \le \bar{j}_k. \tag{2.4}$$

The numbers $a_{j_1}, a_{j_2}, ..., a_{j_m}$ are found from the condition of the existence of a polynomial $P_{n-1}(t) = \sum\limits_{i=0}^{n-1} a_i t^i$ such that $\lim\limits_{t\to\infty} [f(t) - P_{n-1}(t)]^{(k)} = 0$, for all $k = 0, 1, ..., n-1$ and $\|f\|_{p,\alpha} = \left(\int\limits_1^\infty t^\alpha |f(t)|^p \, dt \right)^{\frac{1}{p}}$, $\alpha \in \mathbb{R}$.

In the space $W^n_{p,\bar{\rho}}(I)$, when $I = (0,1)$ and $\rho_i = t^{\alpha_i}$, $\alpha_i \in R$, $i = 1, 2, ..., n$, analogues of the Nikolskii – Lizorkin – Kudryavtsev inequality were proved in [4], [16] and [17].

Let $N = \{0, 1, ..., n-1\}$. Let $N_1 \subset N$ and $N_0 \subset N$ be such that $N_1 \bigcap N_0 = \emptyset$ and $N_1 \cup N_0 = N$.

If $N_1 \neq \emptyset$ and $N_0 \neq \emptyset$, then we set $N_1 = \{i_1, i_2, ..., i_k\}$, $N_0 = \{j_1, j_2, ..., j_m\}$ and $k + m = n$.

Assume that $0 < t_0 \le 1$. Consider a polynomial by the system $\{K_{i,1}(t_0, t)\}_{i=0}^{n-1}$: the function $P_n(t_0, t) = \sum\limits_{i=0}^{n-1} a_i K_{i,1}(t_0, t)$ where a_i, $i = 0, 1, ..., n-1$, are real numbers.

Lemma 2.2 *Let $1 < p < \infty$ and assume that (1.5) holds.*

i) *if $N_1 = \emptyset$ and $N_0 = N$, then the functional*

$$\|f\|^{(1)}_{W^n_{p,\bar{\rho}}(I)} = \left\| D^n_{\bar{\rho}} f \right\|_{p,I} + \sum_{i=0}^{n-1} |a_i(t_0, f)|, \tag{2.5}$$

is equivalent to the norm $\|f\|_{W^n_{p,\bar{\rho}}(I)}$ of the space $W^n_{p,\bar{\rho}}(I)$;

ii) *if $N_1 \neq \emptyset$ and $N_0 \neq \emptyset$, then the functional*

$$\|f\|^{(2)}_{W^n_{p,\bar{\rho}}(I)} = \left\| D^n_{\bar{\rho}} f \right\|_{p,I} + \sum_{\mu=1}^{k} |D^{i_\mu}_{\bar{\rho}} f(1)| + \sum_{\lambda=1}^{m} |a_{j_\lambda}(t_0, f)|, \tag{2.6}$$

is equivalent to the norm $\|f\|_{W^n_{p,\bar{\rho}}(I)}$ of the space $W^n_{p,\bar{\rho}}(I)$.

Now, the main goal is to establish an analog of the Nikolski - Lizorkin - Kudryavtsev inequality in the space $W^n_{p,\bar{\rho}}(I)$ in the general case: $N_1 = \{i_1, i_2, ..., i_k\} \subset N$, $N_0 = \{j_1, j_2, ..., j_m\} \subset N$ and $N_1 \neq \emptyset$, $N_0 \neq \emptyset$.

In the general case we do not assume that $N_1 \bigcap N_0 = \emptyset$ and $N_1 \cup N_0 = N$. Therefore, it can be $N_1 \bigcap N_0 \neq \emptyset$ and $N_1 \cup N_0 \neq N$.

For this, we first prove for the general case, that (2.6) is equivalent to the norm $\|f\|_{W^n_{p,\bar{\rho}}(I)}$ of space $W^n_{p,\bar{\rho}}(I)$.

Let $1 > t_0 > y_0 > 0$. Consider the following matrices:

$$AK(t_0,y_0)=\begin{pmatrix} 1 & K_{1,1}(t_0,y_0) & ... & ... & K_{n-1,1}(t_0,y_0) \\ 0 & 1 & K_{2,2}(t_0,y_0) & ... & K_{n-1,2}(t_0,y_0) \\ 0 & 0 & 0 & ... & ... \\ 0 & 0 & ... & 1 & K_{n-1,n-2}(t_0,y_0) \\ 0 & 0 & 0 & ... & 1 \end{pmatrix},$$

$$AK(1,y_0)=\begin{pmatrix} 1 & K_{1,1}(1,y_0) & ... & ... & K_{n-1,1}(1,y_0) \\ 0 & 1 & K_{2,2}(1,y_0) & ... & K_{n-1,2}(1,y_0) \\ 0 & 0 & 0 & ... & ... \\ 0 & 0 & ... & 1 & K_{n-1,n-2}(1,y_0) \\ 0 & 0 & 0 & ... & 1 \end{pmatrix},$$

In matrix $AK(t_0,y_0)$ we choose rows with numbers $i=i_1,i_2,...,i_k$, and in matrix $AK(1,y_0)$ we choose row numbers $j=j_1,j_2,...,j_m$, where $0\le i_1<i_2<...<i_k\le n-1$, $0\le j_1<j_2<...<j_m\le n-1$ and $k+m=n$.

$$D\begin{vmatrix} i_1 & i_2 & ... & i_k \\ j_1 & j_2 & ... & j_m \end{vmatrix} \tag{2.7}$$

the determinant of the matrix of order n, the first k rows of which are rows of the matrix $AK(t_0,y_0)$ with indices $i_1,i_2,...,i_k$, and the next $m=n-k$ rows are rows of the matrix $AK(1,y_0)$ with indices $j_1,j_2,...,j_m$.

Theorem 2.3 *Let $1<p<\infty$ and assume that (1.5) and general case holds. If $k+m=n$ and*

$$D\begin{vmatrix} i_1 & i_2 & ... & i_k \\ j_1 & j_2 & ... & j_m \end{vmatrix}\neq 0, \tag{2.8}$$

then functional $\|f\|^{(2)}_{W^n_{p,\bar\rho}(I)}=\left\|D^n_{\bar\rho}f\right\|_{p,I}+\sum\limits_{\mu=1}^{k}|D^{i_\mu}_{\bar\rho}f(1)|+\sum\limits_{\lambda=1}^{m}|a_{j_\lambda}(t_0,f)|$ is equivalent to the norm $\|f\|_{W^n_{p,\bar\rho}(I)}$ of space $W^n_{p,\bar\rho}(I)$.

Remark 2.4 It is known [18] that if $P_n(t)$- is an algebraic polynomial of degree $n-1$, then the condition (2.8) is satisfied if and only if the indices $\{i_\mu\}_{\mu=1}^{k}$ and $\{j_\nu\}_{\nu=1}^{m}$ satisfy the Polya condition (2.4). However, the author does not know similar conditions for system $\left\{\bar K_{j,i+1}(1,y_0)\right\}_{i,j=0}^{n-1}$.

Now from Theorems 2.1 and 2.3 we obtain inequality of the Nikolskii - Lizorkin - Kudryavtsev type in the space $W^n_{p,\bar\rho}(I)$

Theorem 2.5 *Let the conditions of Theorem 2.3 be satisfied and $B<\infty$. Let the indices $0\le i_1<i_2<...<i_k\le n-1$ and $0\le j_1<j_2<...<j_m\le n-1$ satisfy (2.8). Then for*

any $f \in W^n_{p,\bar{\rho}}(I)$ *and for all* $i = 0, 1, ..., n-1$, *inequality*

$$\left\|D^i_{\bar{\rho},\tau_i} f\right\|_{p,I} \leq CB \left(\left\|D^n_{\bar{\rho}} f\right\|_{p,I} + \sum_{\mu=1}^{k} |D^{i_\mu}_{\bar{\rho}} f(1)| + \sum_{\nu=1}^{m} |a_{j_\nu}(t_0, f)| \right)$$

holds.

References

[1] V.I. Kondrashov, *On one estimate for family of functions satisfying some integral inequalities.* Dokl. Akad. Nauk SSSR, 4-5 (18) (1938), 253–254 (in Russian) .

[2] L.D. Kudryavtsev, *Direct and inverse embedding theorems. Applications to the solution of elliptic equations by variational methods.* Trudy Mat. Inst. Steklov, 55, Acad. Sci. USSR, Moscow, 1959, 3–182, 184 pp (in Russian).

[3] L.D. Kudryavtsev,, *On norms in weighted spaces of functions given on infinite intervals.* Anal. Math. 12 (1986), No. 4, 269–282.

[4] L.A. Baydeldinov, "The theory of multiweighted spaces and its application to boundary value problems for singular differential equations". Doctoral dissertation. Almaty, 1998, 273 pp.

[5] A.A. Kalybay,, "A new development of Nikol'skii -Lizorkin and Hardy type inequalities with applications". PhD Thesis, Lulea University of Technology, 2006, 145 pp.

[6] Z.T. Abdikalikova, "Some new results concerning boundedness and compactness for embeddings between spaces of the multiweighted derivatives". PhD Thesis, Lulea University of Technology, 2009, 97 pp.

[7] A.A. Kalybay, *Boundary value conditions for linear differential equations with power degenerations.* Bound Value Probl, V. 2020, 110 (2020).

[8] L.D. Kudryavtsev, "Selected Works. Volume II. Chapter I. Functional spaces.". Differential equations. Fizmatlit, Moscow, 2008 (in Russian).

[9] L.D. Kudryavtsev, "Selected Works. Volume II. Chapter II. Functional spaces.". Differential equations. Fizmatlit, Moscow, 2008 (in Russian).

[10] L.A. Baydeldinov,, "The theory of multiweighted spaces and its application to boundary value problems for singular differential equations". Doctoral dissertation. Almaty, 1998, 273 pp.

[11] A.A. Kalybay, "A new development of Nikol'skii-Lizorkin and Hardy type inequalities with applications". PhD Thesis, Lulea University of Technology, 2006, 145 pp.

[12] Z.T. Abdikalikova, "Some new results concerning boundedness and compactness for embeddings between spaces of the multiweighted derivatives". PhD Thesis, Lulea University of Technology, 2009, 97 pp.

[13] L.D. Kudryavtsev, *On norms in weighted spaces of functions given on infinite intervals.* Anal. Math. 12 (1986), No. 4, 269–282.

[14] S.M. Nikol'skii and N.I. Lizorkin, *On some inequalities for functions from weighted classes and boundary value problems with strong degeneracy on the boundary.* Dokl. 1964, T. 154, No. 3, 512–515 (in Russian).

[15] L.D. Kudryavtsev, *On norms in weighted spaces of functions given on infinite intervals.* Anal. Math. 12 (1986), No. 4, 269–282.

[16] A.A. Kalybay, *Generalized multiparameter weighted Nikol'skii-Lizorkin inequality.* Dokl. 2003, T. 391, No. 6, 727–733 (in Russian).

[17] A.A. Kalybay and R. Oinarov, *Some properties of spaces with multiweighted derivatives.* Progress in Analysis. Proc. 3rd Int. ISAAC Congress. Singapore. World Scientific 2003, 1–13.

[18] L.D. Kudryavtsev, *Variational problems with different numbers of boundary conditions.* Tr. Steklov Mathematical Institute of the USSR, 1990, T. 192, 85–104 (in Russian).

Mathematical Finance: a Tale of Stochastic Processes

Guillaume Szulda [(*)]

Abstract. Financial markets are highly uncertain environments where the evolution of asset prices exhibits random fluctuations over time, in particular due to complex and unpredictable market mechanisms. In this regard, *stochastic analysis*, which is at the intersection between the theory of probability and functional analysis, plays a fundamental role in financial modeling.
This short document constitutes a detailed version of the presentation that I gave at the doctoral seminar of the University of Padua on December 10, 2021. It is divided into two major sections. The first one is mostly introductory, where I first give/recall elementary but indispensable notions of probability and stochastic calculus, then I illustrate the fundamentals of mathematical finance, most notably the extensive application of *stochastic processes*.
In the second section, I provide an example of a more technical modeling framework, which constitutes a simplified part of my doctoral research (see [Szu21]). This example is devoted to the modeling of the post-crisis interest rate market, where multiple term structures typically coexist. More specifically, we make use of a flow of *Continuous-state Branching processes with Immigration* (CBI processes), which form a sophisticated class of stochastic processes.

1 Stochastic calculus for financial modeling

A financial market is a place where agents trade in financial assets, which can be stocks, bonds, currencies, commodities, etc. Broadly speaking, assets prices exhibit two peculiar stylized facts. First, they are functions of time, which naturally involves notions of *functional analysis*. Then, they fluctuate randomly over time, which also uses the theory of *probability*.

In mathematical finance, the purpose of any financial model consists in setting a mathematical formulation, exploiting *stochastic analysis* that combines functional analysis and probability, which seeks to fully explain the random fluctuations of asset prices over time.

To proceed, we commonly start by fixing a *stochastic basis*, which is mathematically denoted by $\big(\Omega, \mathcal{F}, \mathbb{F} = (\mathcal{F}_t)_{t\geq 0}, \mathbb{P}\big)$. This corresponds to the combination of a *probability space* $\big(\Omega, \mathcal{F}, \mathbb{P}\big)$ and a *filtration* $\mathbb{F} = (\mathcal{F}_t)_{t\geq 0}$ (see [JS03] for further details). Let us first

[(*)]Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy.
E-mail: szulda.guillaume@gmail.com . Seminar held on 2 December 2021.

define the former, which represents, in mathematical terms, the presence of randomness in the financial market.

Definition 1.1 A *probability space* is a triplet $\big(\Omega, \mathcal{F}, \mathbb{P}\big)$ where

- Ω is the set of all possible outcomes;
- $\mathcal{F}$ is a σ-algebra on Ω;
- $\mathbb{P}$ is a probability measure on $\mathcal{F}$, i.e. a function $\mathbb{P} : \mathcal{F} \to [0, 1]$.

In financial terms, $\{\omega\} \subset \Omega$ stands for one possible scenario that can occur in the market, more specifically, one possible evolution of the different asset prices.

As mentioned above, asset prices evolve with respect to time in a financial market. In view of taking into account the time evolution of the financial market in the modeling, let us now define the notion of *filtration*. Without loss of generality, we typically fix a time horizon $T > 0$, which then restricts the modeling of the financial market to the segment $[0, T]$.

Definition 1.2 A *filtration* $\mathbb{F} = (\mathcal{F}_t)_{t\geq 0}$ is an increasing family of σ-algebras in the sense of inclusion, meaning that $\mathcal{F}_s \subseteq \mathcal{F}_t$, for all $0 \leq s \leq t \leq T$.

As a financial interpretation, $\mathbb{F}$ illustrates the fact that information is revealed progressively by the market as time elapses.

Let us henceforth introduce a fundamental object of stochastic analysis and mathematical finance, which can be deemed as the analogue of the classic function in functional analysis.

Definition 1.3 A *stochastic process* $X = (X_t)_{0\leq t\leq T}$ is a family of random variables $X_t : \Omega \to \mathbb{R}$ indexed by $0 \leq t \leq T$.

For each $\omega \in \Omega$, $t \to X_t(\omega)$ is a trajectory (or sample path) of X, and is one possible evolution of X associated to the scenario $\omega \in \Omega$. So as to ensure consistency between $\mathbb{F}$ and X, we usually suppose that X is *adapted* to $\mathbb{F}$, meaning that X_t is $\mathcal{F}_t$-measurable for all $t \leq T$.

A canonical example of a stochastic process is *Brownian motion*, defined as follows.

Definition 1.4 A stochastic process $W = (W_t)_{t\leq T}$ is said to be a *Brownian motion* or *Wiener process* if

(a) $W_0 = 0$;

(b) W is almost surely continuous, i.e. for each $\omega \in \Omega$, $t \to W_t(\omega)$ is continuous;

(c) W has independent increments, i.e. for all $0 \leq s_1 \leq t_1 \leq s_2 \leq t_2$, $W_{t_2} - W_{s_2}$ and $W_{t_1} - W_{s_1}$ are independent;

(d) For all $0 \leq s \leq t$, $W_t - W_s \sim \mathcal{N}(0, t - s)$, where $\mathcal{N}(m, \sigma^2)$ is the Gaussian law of mean m and variance σ^2.

A peculiar feature of Brownian motion is that for every single $\omega \in \Omega$, the function $t \to W_t(\omega)$ is continuous everywhere, however the latter is not differentiable anywhere, typically referred to as "a continuous function which consists entirely of corners" (refer to [Bjo09, Section 4.5]).

Let us at present consider a simplified financial market in which agents can trade in one single risky asset whose price is represented by the stochastic processes $S = (S_t)_{t \leq T}$ up to the time horizon T. The very starting point in view of modeling the stochastic behavior of $S = (S_t)_{t \leq T}$ stems from the very well-known deterministic setting, more particularly the first-order ordinary differential equation thus formulated:

$$\frac{\mathrm{d}S_t}{\mathrm{d}t} = \lambda(t, S_t),$$

which is defined as the limit of the following difference quotient as $\Delta t \to 0$:

$$\frac{S_{t+\Delta t} - S_t}{\Delta t} = \lambda(t, S_t).$$

The natural intuition here is to add a Gaussian disturbance term of variance Δt, whose scale is ruled by a volatility term $\sigma(t, S_t)$. In doing so, we write

$$S_{t+\Delta t} - S_t = \lambda(t, S_t)\,\Delta t + \sigma(t, S_t)\,\sqrt{\Delta t}\,Z,$$

where $Z \sim \mathcal{N}(0, 1)$. We can thus make use of Brownian motion previously defined, as follows:

$$S_{t+\Delta t} - S_t = \lambda(t, S_t)\,\Delta t + \sigma(t, S_t)\,\big(W_{t+\Delta t} - W_t\big).$$

At this point, we are mathematically tempted, taking inspiration from the deterministic world, to divide by Δt and make it tend to zero. Unfortunately, we cannot proceed in this way owing to the non-differentiability of Brownian motion, meaning that the following derivative taken in the standard sense:

$$\frac{\mathrm{d}W_t}{\mathrm{d}t} = \lim_{\Delta t \to 0} \frac{W_{t+\Delta t} - W_t}{\Delta t},$$

does not exist anywhere. In this regard, the interested reader is re-invited to remember the quotation from [Bjo09, Section 4.5]: "a continuous function which consists entirely of corners".

A possible way out is to directly consider sums and transform them into integrals. To do so, let us introduce a partition of $[0, T]$ into subintervals $[t_k, t_{k+1}]$ for $k = 0, \ldots, n-1$ with $t_k := k\Delta$ and $\Delta := \frac{T}{n}$. By summing over k, we obtain

$$S_T = S_0 + \sum_{k=0}^{n-1} \lambda(t_k, S_{t_k})\Delta + \sum_{k=0}^{n-1} \sigma(t_k, S_{t_k})\Delta W_{t_k}.$$

We now make Δt tend to zero, yielding on one hand

$$\sum_{k=0}^{n-1} \lambda(t_k, S_{t_k})\Delta \underset{\Delta \to 0}{\longrightarrow} \int_0^T \lambda(s, S_s)\,\mathrm{d}s$$

which is a classic Riemann integral defined for each $\omega \in \Omega$, i.e. pathwise, and on the other hand

$$\sum_{k=0}^{n-1} \sigma(t_k, S_{t_k}) \Delta W_{t_k} \underset{\Delta \to 0}{\longrightarrow} \int_0^T \sigma(s, S_s) \, \mathrm{d}W_s$$

which exists in $\mathbb{L}^2$ but not pathwise, and refers to the *Itô integral* (or even *stochastic integral*), and was introduced by Kiyoshi Itô (1915–2008) in 1944.

We finally model the asset price $S = (S_t)_{t \leq T}$ by the following stochastic integral equation:

$$S_t = S_0 + \int_0^t \lambda(s, S_s) \, \mathrm{d}s + \int_0^t \sigma(s, S_s) \, \mathrm{d}W_s,$$

which we typically rewrite as a stochastic differential equation, often referred to as a *diffusion*:

$$\mathrm{d}S_t = \lambda(t, S_t) \, \mathrm{d}t + \sigma(t, S_t) \, \mathrm{d}W_t.$$

Example 1.5 As seminal examples, we present two financial models built upon a diffusion:

- The *Bachelier model* (1900):

$$\mathrm{d}S_t = \sigma \, \mathrm{d}W_t \quad \Longleftrightarrow \quad S_t = S_0 + \sigma \, W_t;$$

- First Samuelson (1965), then Black & Scholes (1973) and Merton (1973), gave rise to the well-known *Black–Scholes model* (also called *geometric Brownian motion*):

$$\mathrm{d}S_t = S_t \, \lambda \, \mathrm{d}t + S_t \, \sigma \, \mathrm{d}W_t \quad \Longleftrightarrow \quad S_t = S_0 \, e^{\left(\lambda - \frac{\sigma^2}{2}\right) t + \sigma \, W_t}.$$

2 Example of a more technical modeling framework

In this section, we investigate an example of a more technical financial model. The latter is in fact part of my doctoral research (see [Szu21]), nonetheless we have simplified it in such a way that it only relies on the ingredients defined in the previous section.

From now on, we place ourselves within a very particular financial market, namely the post-crisis interest rate market. Since the 2007–2009 financial crisis, the interest rate market has been characterized by a segmentation into multiple yield curves. Simply speaking, a yield curve refers to the graph of a specific interest rate with respect to the maturity.

On one hand, we have the Overnight Indexed Swaps (OIS) rates denoted by $T \mapsto F(T, T, \delta)$. These are determined as geometric average of overnight rates and represent the most widely adopted proxies nowadays for risk-free rates, where $F(T, T, \delta)$ is the OIS spot rate for the period $[T, T + \delta]$ with $\delta > 0$.

On the other hand, we have the interbank offered rates. These are the rates at which primary financial institutions can borrow money from each other for some period of time (referred to as *tenor*). They are denoted by $T \mapsto L^\delta(T, T, \delta)$, for every tenor δ of a generic set $\mathcal{D} := \{\delta_1, \ldots, \delta_m\}$ with $0 < \delta_1 < \ldots < \delta_m$ for some $m \in \mathbb{N}$, where $L^\delta(T, T, \delta)$ is the spot interbank offered rate for the period $[T, T + \delta]$.

In the post-crisis environment, interbank and OIS rates, together with interbank rates of different tenors, demonstrate a very distinct behavior. This phenomenon is especially due to the stronger presence of numerous risk sources (see for instance [MU08, GKP11, CD13, GSS17]), and is reflected by the presence of tenor-dependent spreads between different yield curves.

There exist in the literature different ways to express these spreads by means of mathematical terms. A possibility is to define them as follows. For every tenor $\delta \in \mathcal{D}$, consider the stochastic process $S^\delta = (S^\delta_t)_{t\geq 0}$ given by

$$S^\delta_t := \frac{1+\delta L^\delta(t,t+\delta)}{1+\delta F(t,t+\delta)}, \qquad \text{for all } t \geq 0.$$

The processes $(S^\delta)_{\delta\in\mathcal{D}}$ thus given correspond to the *multiplicative spreads* between (normalized) spot interbank offered rates and (normalized) OIS spot rates. The idea of modeling multiple yield curve markets via multiplicative spreads is initially due to [Hen14], and has been pursued by [CFG16, CFG19, EGG20, FGGS20] among others.

Nevertheless, modeling these quantities poses a real challenge insofar that they exhibit peculiar empirical features that are intrinsically elaborate to capture. A brief inspection of Figure 1 directly provides their most basic empirical features, which, in particular, we seek to reproduce in the present simplified modeling framework (we refer the interested reader to [FSG21] for the complete list of their empirical features):

- The spreads are typically greater than one, which derives from the fact that interbank rates are riskier than OIS rates;
- They are non-decreasing with respect to the tenor, i.e. S^δ is non-decreasing over $\mathcal{D}$ as a function of δ. This is due to the fact that the longer the length of the loan is, the higher the risk becomes.

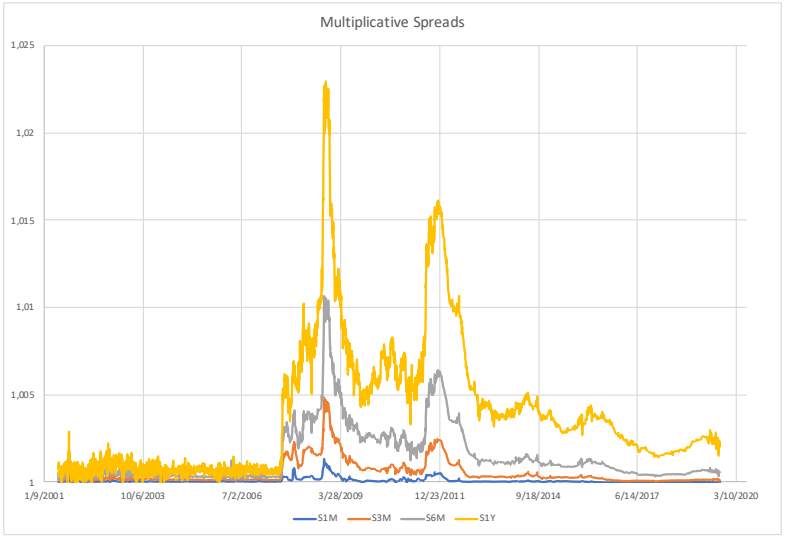


Figure 1. Euribor-OIS spreads from 06/2001 to 09/2019 (Source: Bloomberg).

The approach that we now adopt for modeling multiple yield curves as discussed above can be performed in three steps as follows. We especially recall that it corresponds to a simplified version of the more technical modeling framework that can be found in [FSG21], to which we refer the interested reader for further details.

Step 1: We start by defining a special type of a stochastic process, which we denote by $X = (X_t)_{t\geq 0}$. The latter is totally characterized by the following diffusion as in Section 1:

$$\mathrm{d}X_t = \alpha\left(\beta - X_t\right)\mathrm{d}t + \sigma\sqrt{X_t}\,\mathrm{d}W_t,$$

which we can also rewrite in its integral form:

$$X_t = X_0 + \int_0^t \alpha\left(\beta - X_s\right)\mathrm{d}s + \int_0^t \sigma\sqrt{X_s}\,\mathrm{d}W_s.$$

The stochastic process $X = (X_t)_{t\geq 0}$ is typically referred to as a *Cox–Ingersoll–Ross process* (CIR process), first introduced by [CIR85]. Such a process is commonly employed in stochastic volatility modeling (see e.g. [Hes93]) or in interest rate modeling (see e.g. [Fil01]), owing to its peculiar features. Last, but by no means not least, it represents the simplest form in the literature of *Continuous-state Branching processes with Immigration* (CBI processes), which are a very sophisticated class of stochastic processes that we are going to exploit shortly (we also refer the interested reader to [KW71] and [Li11, Li20] for full details on CBI processes).

Step 2: This constitutes the delicate step of the construction of our modeling framework for multiple yield curves, since it resorts to tools that are not usually adopted in mathematical finance. Let us first assume that the stochastic basis $\big(\Omega, \mathcal{F}, \mathbb{F} = (\mathcal{F}_t)_{t\geq 0}, \mathbb{P}\big)$, which we fixed in Section 1, supports a *white noise* $B(\mathrm{d}t, \mathrm{d}u)$ on $\mathbb{R}^2_+$ of intensity $\mathrm{d}t\,\mathrm{d}u$, and which we briefly recall the definition below (refer to [Wal86]).

Definition 2.1 $B(\mathrm{d}t, \mathrm{d}u)$ is said to be a *white noise* on $\mathbb{R}^2_+$ of intensity $\lambda(\mathrm{d}t, \mathrm{d}u)$ if this is a *Gaussian random measure* in the following sense:

- For any $A \in \mathcal{B}\big(\mathbb{R}^2_+\big)$ with $\lambda(A) < +\infty$, $B(A) \sim \mathcal{N}\big(0, \lambda(A)\big)$;
- If $A_1, \ldots, A_n$ are pairwise disjoint, then $B(A_1), \ldots, B(A_n)$ are mutually independent.

Therefore, by exploiting the characterization of the CIR process $X = (X_t)_{t\geq 0}$ as a CBI process, we can thus rewrite its defining diffusion by means of the white noise B as follows:

$$\mathrm{d}X_t = \alpha\left(\beta - X_t\right)\mathrm{d}t + \int_0^{X_t} \sigma\, B(\mathrm{d}t, \mathrm{d}u).$$

or in the stochastic integral form given by

$$X_t = X_0 + \int_0^t \alpha\left(\beta - X_s\right)\mathrm{d}s + \int_0^t \int_0^{X_s} \sigma\, B(\mathrm{d}s, \mathrm{d}u).$$

This corresponds to the *Dawson–Li representation* of $X = (X_t)_{t\geq 0}$ in the sense of [DL12], which yields a weakly equivalent CIR process, however not almost surely (pathwise) equivalent. More specifically, the peculiarity of such a representation lies in its comparison property (see e.g. [Li20, Theorem 8.4]), which is the starting point in the construction of a financial model for multiple yield curves driven by a *flow of CBI processes*.

Step 3: This final step consists in formulating our modeling framework for multiple yield curves. To this purpose, let $X_0, \beta : \mathcal{D} \to \mathbb{R}_+$ be both non-decreasing and deterministic functions on $\mathcal{D}$ (i.e. the set of tenors). Hence, for every $\delta \in \mathcal{D}$, define the stochastic process $X(\delta) = (X_t(\delta))_{t\geq 0}$ by means of the following stochastic integral equation:

$$X_t(\delta) = X_0(\delta) + \int_0^t \alpha \Big(\beta(\delta) - X_s(\delta)\Big)\, \mathrm{d}s + \int_0^t \int_0^{X_s(\delta)} \sigma\, B(\mathrm{d}s, \mathrm{d}u).$$

$\big\{X(\delta) : \delta \in \mathcal{D}\big\}$ then represents a simple instance of a *flow of CBI processes* in the sense of [DL12]. Consequently, our model configuration for multiple yield curves, which takes the multiplicative spreads $(S^\delta)_{\delta\in\mathcal{D}}$ as principal modeling quantities, is given by

$$\log\, S_t^\delta = X_t(\delta), \qquad \text{for every } \delta \in \mathcal{D}.$$

By definition, we have $X_t(\delta) \geq 0$ almost surely, for all $t \geq 0$ and for every $\delta \in \mathcal{D}$, we then automatically obtain that all multiplicative spreads $(S^\delta)_{\delta\in\mathcal{D}}$ are greater than one almost surely. As far as their non-decreasing behavior with respect to the tenor is now concerned, we can write the following proposition.

Proposition 2.2 *It holds* $S_t^{\delta_i} \leq S_t^{\delta_{i+1}}$ *almost surely, for all* $t \geq 0$ *and every* $1 \leq i \leq m-1$.

Proof. Since both functions X_0 and β are non-decreasing on $\mathcal{D}$, [DL12, Theorem 3.2] immediately implies that for every $1 \leq i \leq m-1$, we have $\mathbb{P}\big(X_t(\delta_i) \leq X_t(\delta_{i+1}), \forall t \geq 0\big) = 1$. The claim then follows from the model configuration as formulated above. □

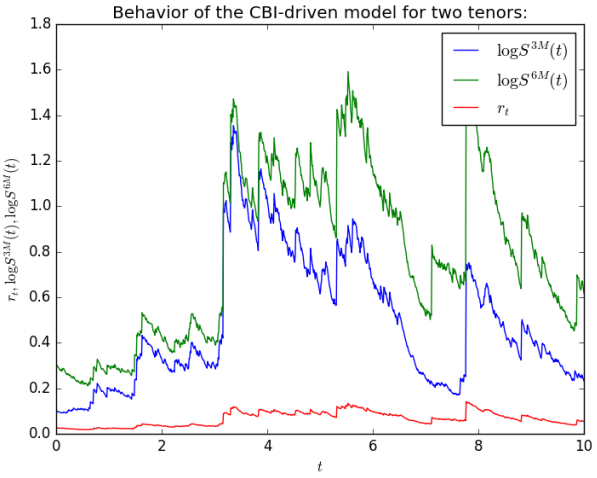


Figure 2. One sample path of the multiplicative spreads $(S^\delta)_{\delta\in\mathcal{D}}$ for two tenors (3M in blue and 6M in green). Compare with the empirical features of Figure 1.

In conclusion, we report in Figure 2 a sample path of the model configuration previously defined, which has been generated by a more sophisticated financial model including a short-rate model (in red) as well as jumps (refer to [FSG21] for full details).

References

[Bjo09] T. Bjork, "Arbitrage Theory in Continuous Time; 3rd ed". Oxford University Press, 2009.

[CD13] S. Crépey and R. Douady, *Lois: credit and liquidity.* Risk Magazine, pages 82–86, June 2013.

[CFG16] C. Cuchiero, C. Fontana, and A. Gnoatto, *A general HJM framework for multiple yield curve modeling.* Finance and Stochastics, 20(2): 267–320, 2016.

[CFG19] C. Cuchiero, C. Fontana, and A. Gnoatto, *Affine multiple yield curve models.* Mathematical Finance, 29(2): 568–611, 2019.

[CIR85] J.C. Cox, J.E. Ingersoll, and S.A. Ross, *A theory of the term structure of interest rates.* Econometrica, 53(2): 385–407, 1985.

[DL12] D.A. Dawson and Z. Li, *Stochastic equations, flows and measure-valued processes.* The Annals of Probability, 40(2): 813–857, 2012.

[EGG20] E. Eberlein, C. Gerhart, and Z. Grbac, *Multiple curve Lévy forward price model allowing for negative interest rates.* Mathematical Finance, 30(1): 167–195, 2020.

[FGGS20] C. Fontana, Z. Grbac, S. Gümbel, and T. Schmidt, *Term structure modeling for multiple curves with stochastic discontinuities.* Finance and Stochastics, 24: 465–511, 2020.

[FGS21] C. Fontana, A. Gnoatto, and G. Szulda, *Multiple yield curve modeling with CBI processes.* Mathematics and Financial Economics, 15(2): 579–610, 2021.

[Fil01] D. Filipovíc, *A general characterization of one factor affine term structure models.* Finance and Stochastics, 5(3): 389–412, 2001.

[GKP11] D. Gefang, G. Koop, and S.M. Potter, *Understanding liquidity and credit risks in the financial crisis.* Journal of Empirical Finance, 18(5): 903–914, 2011.

[GSS17] J. Gallitschke, S. Seifried, and F.T. Seifried, *Interbank interest rates: Funding liquidity risk and XIBOR basis spreads.* Journal of Banking and Finance, 78: 142–152, 2017.

[Hen14] M. Henrard, "Interest Rate Modelling in the Multi-Curve Framework". Palgrave Macmillan, 2014.

[Hes93] L. Heston, *A closed-form solution for options with stochastic volatility with applications to bond and currency options.* Review of Financial Studies, 6(2): 327–343, 1993.

[JS03] J. Jacod and A. Shiryaev, "Limit Theorems for Stochastic Processes; 2nd ed". Springer, Berlin–Heidelberg–New York, 2003.

[KW71] K. Kawazu and S. Watanabe, *Branching processes with immigration and related limit theorems.* Theory of Probability and its Applications, 16(1): 36–54, 1971.

[Li11] Z. Li, "Measure-Valued Branching Markov Processes". Springer, Berlin–Heidelberg, 2011.

[Li20] Z. Li, *Continuous-state branching processes with immigration.* In Y. Jiao, editor, From Probability to Finance - Lecture Notes of BICMR Summer School on Financial Mathematics, pages 1–70. Springer, Singapore, 2020.

[MU08] F.L. Michaud and C. Upper, *What drives interbank rates? Evidence from the Libor panel.* BIS Quarterly Review, March 2008.

[Szu21] G. Szulda, "Branching Processes and Multiple Term Structure Modeling". PhD thesis, Université de Paris, 2021.

[Wal86] J.B. Walsh, *An introduction to stochastic partial differential equations.* In P.L. Hennequin, editor, École d'Été de Probabilités de Saint Flour XIV - 1984, pages 265–439. Springer, Berlin–Heidelberg, 1986..

The critical node/edge detection problem on trees

Syed MD Omar Faruk [(*)]

Abstract. Critical node or edge detection problems are a family of optimization problems defined on graphs, where one is required to remove a limited number of nodes and/or edges in order to minimize some measure of the connectivity of the residual graph. Problems of this type are important from a practical point of view because of their relevance in a number of practical applications. We start this seminar by giving the definitions of the critical node/edge detection problem (CNDP/CEDP) and some connectivity metrics with an example. After that, we present a dynamic programming approach for solving the CNDP on trees when the node weights are all equal to one and all connections between pairs of nodes have unit cost. Then, we will move to consider the CEDP on trees and similarly deal with the case with unit costs and unit edge weights. Finally, we will present dynamic programming algorithms for the "mixed" case, in which nodes and edges can be simultaneously removed from the graph.

1 Introduction

Given an undirected graph $G(V, E)$ with $|V| = n$ nodes, the Critical Node Problem (CNP) calls for removing from G a subset of nodes $S \subseteq V$ in order to minimize some connectivity measure in the subgraph $G[V - S]$ induced by $V - S$, while a constraint on the size or "weight" of S has to be enforced. In a possible — and fairly general — formulation, a nonnegative connection cost c_{ij} is specified for each pair of distinct nodes $i, j \in V$, a weight $w_j \geq 0$ for each $j \in V$ and a bound K are given. Two nodes i, j are connected if a path exists between them. The optimal solution is required to

$$\begin{aligned} &\text{minimize } f(S) = \sum\{c_{ij} : i, j \text{ are connected in } G[V - S]\} \\ &\text{subject to } \sum_{j \in S} w_j \leq K. \end{aligned}$$

The problem has attracted some attention in recent years; especially the case where $c_{ij} = 1$ for all $i \neq j$, $w_j = 1$ for all $j \in V$ has been tackled in the literature. In such a case

[(*)]Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy.
Current address of the author: Department of Mathematics, Shahjalal University of Science and Technology, Sylhet, Bangladesh. E-mail: `omarfaruk-mat@sust.edu` . Seminar held on 15 December 2021.

the problem amounts to removing at most $K \leq n$ nodes, minimizing the number of pairs connected in the residual graph. Applications of CNP considered in the literature include: fragmentation of *terrorist networks*, where a fixed number of persons has to be identified in such a way that their removal will result in the minimum communication between the remaining individuals (see [4]); *network immunization*, where a graph representing contacts between people is given, only a given maximum number of persons can be vaccinated, and we aim at minimizing the propagation of the virus (see [1, 5]); *transportation networks*, where identifying critical nodes, i.e., nodes whose failure would highly compromise the efficiency of the transportation, is quite important for a correct allocation of the resources (see [3]); *telecommunication networks*, when we want to prevent the spread of a virus or find some way to reduce as much as possible the communication within the network (see [2]).

1.1 The problems that we study

According to the pairwise connectivity measure mentioned above, the Critical Node Detection Problem (CNDP) is formally stated as follows:

Problem 1 CNDP *Given an undirected graph $G = (V, E)$, a weight $w_j \geq 0$ for every $j \in V$, a connection cost $c_{ij} \geq 0$ for all $i, j \in V$, and a weight limit $W \geq 0$, find $S \subseteq V$ such that the total weight of the nodes in S is at most W and the total cost of the pairs of nodes that are connected in $G - S$ is minimized.*

We are interested in some variants of CNDP (mainly on trees) in which nodes or edges or both nodes and edges can be removed. The variants that we analyze are the following:

- the Critical Edge Detection Problem (CEDP), which is formulated as CNDP, except that edges have to be removed instead of nodes;
- the Critical Node/Edge Detection Problem with a single weight limit (CNEDP), where a cumulative weight limit for the removal of nodes and edges is given.

These problems are formalized below:

Problem 2 CEDP *Given an undirected graph $G = (V, E)$, a weight $w_e \geq 0$ for every $e \in E$, a connection cost $c_{ij} \geq 0$ for all $i, j \in V$, and a weight limit $W \geq 0$, find $S \subseteq E$ such that $w(S) \leq W$ and $c(G - S)$ is minimized.*

Problem 3 CNEDP *Given an undirected graph $G = (V, E)$, a weight $w_v \geq 0$ for every $j \in V$, a weight $w_e \geq 0$ for every $e \in E$, a connection cost $c_{ij} \geq 0$ for all $i, j \in V$, and a weight limit $W \geq 0$, find $S \subseteq V \cup E$ such that $w(S) \leq W$ and $c(G - S)$ is minimized.*

2 The unit-costs, unit-weights case on trees

In this section we illustrate a polynomial algorithm for solving CNDP on trees when $c_{ij} = 1$ for all i, j and $w_j = 1$ for all $j \in V$. In this case the problem calls for minimizing the number of paths surviving in a tree $T(V, E)$ after having removed at most K nodes.

Given the tree $T(V, E)$ with $|V| = n$, let T_a be the subtree of T rooted at $a \in V$. If a is not a leaf of T, let $T_{a_1}, \ldots, T_{a_s}$ be the subtrees of T_a rooted at the children nodes $a_1, \ldots, a_s$ respectively, where s depends on a (see Figure 1). Let also $|T_a|$ be the number of nodes in T_a. In order to solve the problem by dynamic programming, we define the following functions.

$F_a(m, k)$ = minimum number of paths surviving in T_a when k nodes are removed from T_a and m nodes of T_a are still connected to a. Note that the number of nodes connected to some given $v \in V$ always counts v itself. Condition $m = 0$ indicates that a is removed from T_a. Furthermore, if it is not possible to remove k nodes from T_a so that m nodes of T_a are still connected to a, then we define $F_a(m, k) = \infty$.

$G_{a_i}(m, k)$ = minimum number of paths surviving in the subtree $T_{a_{i,s}} = a + T_{a_i} + \cdots + T_{a_s}$ when k nodes are removed from $T_{a_{i,s}}$ and m nodes of the subtree are still connected to a. As above, $m = 0$ indicates that a is removed from $T_{a_{i,s}}$ and $G_{a_i}(m, k) = \infty$ if it is not possible to remove k nodes from $T_{a_{i,s}}$ so that m nodes of $T_{a_{i,s}}$ are still connected to a.

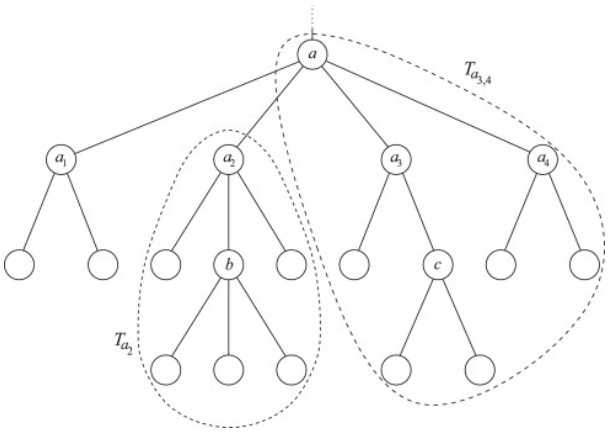


Figure 1 Example of a subtree T_a, where node a has four children (i.e. $s = 4$). The subtrees T_{a_2} and $T_{a_{3,4}}$ are shown.

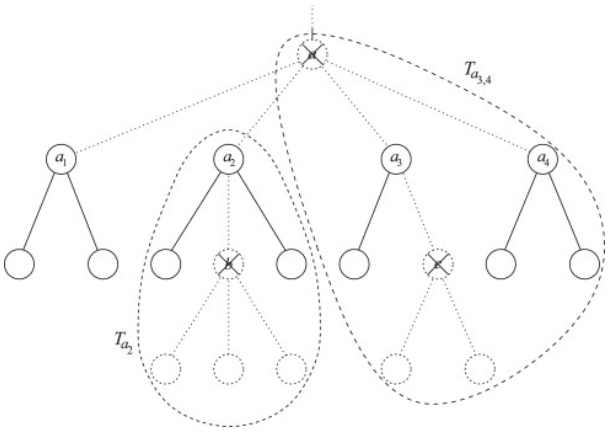


Figure 2 Application of recursion (2) to the subtree of Figure 1 for $m = 0$, $i = 2$ and $k = 3$.

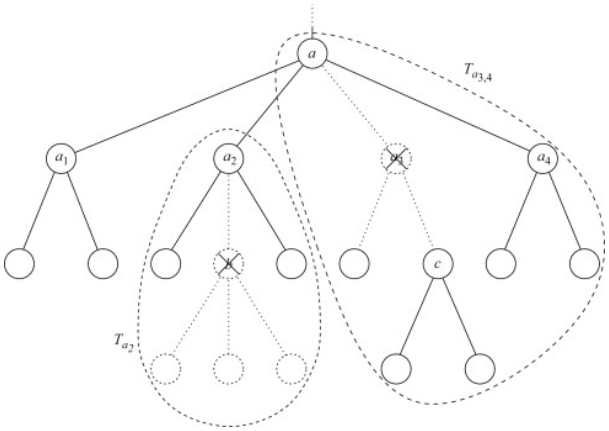


Figure 3 Application of recursion (2) to the subtree of Figure 1 for $m = 7$, $i = 2$ and $k = 2$.

The values for F and G can be computed by traversing the tree in postorder (from leaves to root), by means of the following relations:

(1) $F_a(m,k) = G_{a_1}(m,k)$ for any non-leaf node $a \in V$;

$$(2)\quad G_{a_i}(m,k) = \begin{cases} \min\{F_{a_i}(p,q) + G_{a_{i+1}}(0,k-q) : \\ \qquad p = 0,\dots,|T_{a_i}|,\ q = 0,\dots,k-1\} & \text{if } m = 0\ (a \in S), \\ \min\{F_{a_i}(p,q) + G_{a_{i+1}}(m-p,k-q) + p(m-p) : \\ \qquad p = 0,\dots,m-1,\ q = 0,\dots,k\} & \text{if } m > 0\ (a \notin S), \end{cases}$$

for any non-leaf node $a \in V$ and $i < s$;

the initial conditions on each leaf a and on each rightmost subtree T_{a_s} are the following:

$$(3)\quad F_a(m,k) = \begin{cases} 0 & \text{if } (m = 0, k = 1, \text{ i.e. } a \in S) \text{ or } (m = 1, k = 0, \text{ i.e. } a \notin S), \\ \infty & \text{otherwise,} \end{cases}$$

$$(4)\quad G_{a_s}(m,k) = \begin{cases} \infty & \text{if } m = k = 0, \\ \min\{F_{a_s}(p,k-1) : p = 0,\dots,|T_{a_s}|\} & \text{if } m = 0,\ k > 0\ (a \in S), \\ F_{a_s}(m-1,k) + (m-1) & \text{if } m > 0\ (a \notin S). \end{cases}$$

Equation (1) follows because $T_a = T_{a_{1,s}}$ for any non-leaf node $a \in V$.

Recursion (2) can be interpreted as follows.

- Consider first the case $m = 0$ (i.e., node a is removed from the subtree), which is illustrated in Figure 2 (where $i = 2$ and $k = 3$). Expression $F_{a_i}(p,q)$ gives the minimum number of paths that survive in T_{a_i} when q nodes are removed from T_{a_i}

and p nodes of T_{a_i} are still connected to a_i (for instance, in the example in Figure 2, for $p = 3$ and $q = 1$ we have $F_{a_2}(3, 1) = 3$: this is achieved by removing node b). Since q nodes have been removed from T_{a_i}, exactly $k - q$ nodes (including a) must be removed from $T_{a_{i+1,s}}$. The minimum number of paths that survive in $T_{a_{i+1,s}}$ when $k - q$ nodes (including a) are removed from $T_{a_{i+1,s}}$ is given by $G_{a_{i+1}}(0, k - q)$ (in the example, $G_{a_3}(0, 2) = 4$, which is achieved by removing nodes a and c). Thus the expression $F_{a_i}(p, q) + G_{a_{i+1}}(0, k - q)$ gives the minimum number of paths that survive in $T_{a_{i,s}}$ when q nodes are removed from T_{a_i} (and the other $k - q$ nodes are removed from $T_{a_{i+1,s}}$) and p nodes of T_{a_i} are still connected to a_i (this value is 7 in the example). By taking the minimum over $p = 0, \dots, |T_{a_i}|$ and $q = 0, \dots, k - 1$, we find the value of $G_{a_i}(0, k)$.

- Consider now the case $m > 0$, which is illustrated in Figure 3 (where $m = 7$, $i = 2$ and $k = 2$). As above, expression $F_{a_i}(p, q)$ gives the minimum number of paths that survive in T_{a_i} when q nodes are removed from T_{a_i} and p nodes of T_{a_i} are still connected to a_i (for instance, in the example in Figure 3, for $p = 3$ and $q = 1$ we have $F_{a_2}(3, 1) = 3$: this is achieved by removing node b). Since q nodes have been removed from T_{a_i}, exactly $k - q$ nodes must be removed from $T_{a_{i+1,s}}$; and since p nodes of T_{a_i} are still connected to a_i and thus to a, exactly $m - p$ nodes of $T_{a_{i+1,s}}$ must remain connected to a. The minimum number of paths that survive in $T_{a_{i+1,s}}$ when $k - q$ nodes are removed from $T_{a_{i+1,s}}$ and $m - p$ nodes of $T_{a_{i+1,s}}$ are still connected to a is given by $G_{a_{i+1}}(m - p, k - q)$ (in the example, $G_{a_3}(4, 1) = 9$, which is achieved by removing node a_3). Thus the expression $F_{a_i}(p, q) + G_{a_{i+1}}(m - p, k - q)$ gives the minimum number of paths that survive in T_{a_i} or $T_{a_{i+1,s}}$ when q nodes are removed from T_{a_i} (and the other $k - q$ nodes are removed from $T_{a_{i+1,s}}$) and p nodes of T_{a_i} are still connected to a_i, while $m - p$ nodes of $T_{a_{i+1,s}}$ are still connected to a. Now we have to add the paths connecting nodes of T_{a_i} to nodes of $T_{a_{i+1,s}}$, i.e. $p(m - p)$ paths (12 paths in the example). This gives expression $F_{a_i}(p, q) + G_{a_{i+1}}(m - p, k - q) + p(m - p)$ of recursion (2) (whose value is 24 in the example). By taking the minimum over $p = 0, \dots, m - 1$ and $q = 0, \dots, k$, we find the value of $G_{a_i}(m, k)$.

Given a leaf $a \in V$, equation (3) says that

- if a is removed ($k = 1$), then no node ($m = 0$) and no path survive in T_a (which becomes empty);
- if a is not removed ($k = 0$), then node a survives ($m = 1$), and the number of paths is again 0.

For a justification of (4), recall that $T_{a_{s,s}} = a + T_{a_s}$.

Now, assuming that T is rooted at node 1, the optimal value for the problem is given by

$$\text{OPT} = \min\{F_1(m, K) : m = 0, \dots, n\}, \tag{5}$$

and the optimal solution is recovered by backtracking.

3 Solving CEDP on a tree with unit connection costs

In this part, we show how to solve CEDP on trees when $c_{ij} = 1$ for all i, j and $w_j = 1$ for all $j \in E$. In this scenario, the goal is to reduce the number of paths left in a tree $T(V, E)$ after removing at most K edges.

To derive a dynamic programming algorithm, we will calculate recursively the following values:

- $F_a(m, k)$ = minimum number of connections that still exists in the subtree T_a when k edges are removed from T_a and m nodes of T_a are still connected to the root a.
- $G_{a_i}(m, k)$ = minimum number of connections that still exists in the subtree $T_{a_{i,s}} = a + T_{a_i} + \cdots + T_{a_s}$ when k edges are removed from $T_{a_{i,s}}$ and m nodes of the subtree are still connected to a.

We remark that for both functions, the number of nodes connected to the root will never be 0 (i.e. $m > 0$) because we never remove the root as the root is a node and we can not remove a node in the edge deletion problem. Furthermore, whenever the conditions in one of the above definitions cannot be satisfied, we set the value of the function to infinity.

The values of F_a and G_{a_i} are calculated in this order:

- we determine F_a for every leaf a;
- for a non-leaf node a, assuming that the $F_{a'}$ and $G_{a'_i}$ have been already found for all $a' \in V(T_a)$, we calculate $G_{a_s}, G_{a_{s-1}}, \ldots, G_{a_1}$, and then F_a.

At the end of the recursion, we can return the optimal value of the problem, assuming that the tree T is rooted at node 1, which is

$$\mathrm{OPT} = \min\{F_1(m, K) : m = 0, \ldots, n\}.$$

As usual in dynamic programming, an optimal solution can be reconstructed by backtracking.

We now provide the explicit formulas and then a justification for each of them. For a non-leaf node $a \in V$, we have

$$F_a(m, k) = G_{a_1}(m, k), \tag{6}$$

while for every leaf a the formula is

$$F_a(m, k) = \begin{cases} 0 & \text{if } m = 1,\ k = 0, \\ \infty & \text{otherwise.} \end{cases} \tag{7}$$

For any non-leaf node $a \in V$ and $i < s$ (non-rightmost subtrees) we use the formula

$$(8)\quad G_{a_i}(m,k) =$$
$$\begin{cases} \min\{F_{a_i}(p,0) + G_{a_{i+1}}(m-p,0) + p(m-p) : p = 0,\dots,m\} & \text{if } k = 0, \\ \min\{\min\{F_{a_i}(p,q) + G_{a_{i+1}}(m,k-1-q) : p = 0,\dots,|V(T_{a_i})|, \\ \qquad q = 0,\dots,k-1\},\ \min\{F_{a_i}(p,q) + G_{a_{i+1}}(m-p,k-q) + p(m-p) : \\ \qquad p = 0,\dots,m,\ q = 0,\dots,k\}\} & \text{if } k > 0, \end{cases}$$

The initial conditions on each rightmost subtree T_{a_s} are calculated as follows::

$$(9)\qquad G_{a_s}(m,k) = \begin{cases} \infty & \text{if } m = 1,\ k = 0, \\ \min\{F_{a_s}(p,k-1) : p = 0,\dots,|V(T_{a_s})|\} & \text{if } m = 1,\ k > 0, \\ F_{a_s}(m-1,k) + (m-1) & \text{if } m > 1,\ k \geq 0. \end{cases}$$

We now give a justification for the above formulas. Equation (6) follows because $T_a = T_{a_{1,s}}$ for any non-leaf node $a \in V$.

Equation (7) handles the case of a one-node tree. Since $a \in V$ is a leaf, it is not possible to remove any edge ($k = 0$) and only a is connected to itself ($m = 1$) and the number of paths surviving in T_a is 0.

Recursion (8) can be interpreted as follows:

The case $k = 0$ means that we are not removing any edge from $T_{a_{i,s}} = T_{a_i} + T_{a_{i+1,s}}$. Since we have to keep everything, we are not allowed to remove anything from the subtrees T_{a_i} and $T_{a_{i+1,s}}$. If a_i is connected to p nodes of T_{a_i}, then a is connected to $m-p$ nodes in $T_{a_{i+1,s}}$ and the paths passing through a are exactly $p(m-p)$. Hence by definition of F and G the minimum number of paths that survive in $T_{a_{i,s}}$ when we are not removing anything will be $G_{a_i}(m,0) = \min_p\{F_{a_i}(p,0) + G_{a_{i+1}}(m-p,0) + p(m-p)\}$.

The case $k > 0$ means that we have to remove at least one edge. When the value of $G_{a_i}(m,k)$ is achieved from the expression $F_{a_i}(p,q) + G_{a_{i+1}}(m,k-1-q)$, we remove the edge e (which connects a to a_i). Expression $F_{a_i}(p,q)$ gives the minimum number of paths that survive in T_{a_i} when q edges are removed from T_{a_i} and p nodes of T_{a_i} are still connected to a_i. Since q edges have been removed from T_{a_i}, exactly $k-1-q$ edges must be removed from $T_{a_{i+1,s}}$. The minimum number of paths that survive in $T_{a_{i+1,s}}$ when $k-1-q$ edges are removed from $T_{a_{i+1,s}}$ is given by $G_{a_{i+1}}(m,k-1-q)$. Thus the expression $F_{a_i}(p,q)+G_{a_{i+1}}(m,k-1-q)$ gives the minimum number of paths that survive in $T_{a_{i,s}}$ when q edges are removed from T_{a_i} (and the other $k-1-q$ edges are removed from $T_{a_{i+1,s}}$) and p nodes of T_{a_i} are still connected to a_i. By taking the minimum over $p = 0,\dots,|V(T_{a_i})|$ and $q = 0,\dots,k-1$, we find the value of $G_{a_i}(m,k)$.

When the value of $G_{a_i}(m,k)$ is achieved from the expression $F_{a_i}(p,q) + G_{a_{i+1}}(m-p,k-q) + p(m-p)$, we are not removing the edge e. As above, expression $F_{a_i}(p,q)$ gives the minimum number of paths that survive in T_{a_i} when q edges are removed from T_{a_i} and p nodes of T_{a_i} are still connected to a_i. Since q edges have been removed from T_{a_i}, exactly

$k-q$ edges must be removed from $T_{a_{i+1,s}}$ and since p nodes of T_{a_i} are still connected to a_i and thus to a, exactly $m-p$ nodes of $T_{a_{i+1,s}}$ must remain connected to a. The minimum number of paths that survive in $T_{a_{i+1,s}}$ when $k-q$ edges are removed from $T_{a_{i+1,s}}$ and $m-p$ nodes of $T_{a_{i+1,s}}$ are still connected to a is given by $G_{a_{i+1}}(m-p,k-q)$. Thus the expression $F_{a_i}(p,q)+G_{a_{i+1}}(m-p,k-q)$ gives the minimum number of paths that survive in T_{a_i} or $T_{a_{i+1,s}}$ when q edges are removed from T_{a_i} (and the other $k-q$ edges are removed from $T_{a_{i+1,s}}$) and p nodes of T_{a_i} are still connected to a_i, while $m-p$ nodes of $T_{a_{i+1,s}}$ are still connected to a. Now we have to add the paths connecting nodes of T_{a_i} to nodes of $T_{a_{i+1,s}}$, i.e. $p(m-p)$ paths. This gives expression $F_{a_i}(p,q)+G_{a_{i+1}}(m-p,k-q)+p(m-p)$ of recursion (8). By taking the minimum over $p=0,\dots,m$ and $q=0,\dots,k$, we find the value of $G_{a_i}(m,k)$.

More specifically, if $k=0$, we have only one choice so that we have to keep the edge e (which connects a to a_i) because we are not removing any edge. While in the case $k>0$, we have two possibilities that both are possible i.e., we can choose if we want to remove the edge e or we want to keep it.

For a justification of (9), recall that $T_{a_{s,s}}=a+T_{a_s}$. If $m=1$ and $k>0$, then we have to remove the edge between a and a_s and the other $k-1$ edges we have to be removed from the subtree T_{a_s} and the number of connections that survive are those in the subtree T_{a_s}. On the other hand if $m>1$, then we can not remove the edge a to a_s and in this time we have to remove all the k edges inside the subtree T_{a_s}. Since m nodes are connected to a including a itself, in the subtree we will find the other $m-1$ nodes connected to a_s. Then we have to add all the connections of a to the nodes that are connected to a_s in the subtree.

4 Solving CNEDP on a tree with unit connection costs

The objective in this case is for minimizing the number of paths surviving in the tree $T(V,E)$ after having removed at most K_V nodes and K_E edges.

We use the same tree and subtree notation as in Section 3, and calculate recursively the following functions:

- $F_a(m,k_V,k_E)$ = minimum number of connections that still exists in the subtree T_a after k_V nodes and k_E edges have been removed from T_a and m nodes of T_a remains connected to the root a. Condition $m=0$ indicates that a is removed from T_a.
- $G_{a_i}(m,k_V,k_E)$ minimum number of connections that still exists in the subtree $T_{a_{i,s}}=a+T_{a_i}+\cdots+T_{a_s}$ after k_V nodes and k_E edges have been removed from $T_{a_{i,s}}$ and m nodes of the subtree are still connected to a. As above, $m=0$ indicates that a is removed from $T_{a_{i,s}}$.

We let the function values be infinity whenever the conditions cannot be satisfied. The values for F and G can be computed by traversing the tree in postorder (from leaves to root), by means of the following relations:

For every leaf a we have

$$F_a(m, k_V, k_E) = \begin{cases} 0 & \text{if } (m = 1,\ k_V = k_E = 0) \text{ or } (m = k_E = 0,\ k_V = 1), \\ \infty & \text{otherwise,} \end{cases} \tag{10}$$

while the formula for a non-leaf node a is

$$F_a(m, k_V, k_E) = G_{a_1}(m, k_V, k_E). \tag{11}$$

If a is a non-leaf node, we also have

$$G_{a_s}(m, k_V, k_E) = \tag{12}$$
$$\begin{cases} \min\{F_{a_s}(p, k_V - 1, k_E) : 0 \le p \le |V(T_{a_s})|\} & \text{if } m = 0, \\ \min\left\{F_{a_s}(0, k_V, k_E), \min\{F_{a_s}(p, k_V, k_E - 1) : 0 \le p \le |V(T_{a_s})|\}\right\} & \text{if } m = 1, \\ F_{a_s}(m - 1, k_V, k_E) + m - 1 & \text{if } m > 1, \end{cases}$$

while, for $i < s$,

$$G_{a_i}(m, k_V, k_E) = \tag{13}$$
$$\begin{cases} \min\{F_{a_i}(p, q_V, q_E) + G_{a_{i+1}}(0, k_V - q_V, k_E - q_E) : \\ \qquad 0 \le p \le |V(T_{a_i})|,\ 0 \le q_V \le k_V - 1,\ 0 \le q_E \le k_E\} & \text{if } m = 0, \\ \min\Big\{ \min\{F_{a_i}(p, q_V, q_E) + G_{a_{i+1}}(m, k_V - q_V, k_E - q_E - 1) : \\ \qquad\qquad 0 \le p \le |V(T_{a_i})|,\ 0 \le q_V \le k_V,\ 0 \le q_E \le k_E - 1\}, \\ \qquad \min\{F_{a_i}(p, q_V, q_E) + G_{a_{i+1}}(m - p, k_V - q_V, k_E - q_E) + p(m - p) : \\ \qquad\qquad 0 \le p \le m,\ 0 \le q_V \le k_V,\ 0 \le q_E \le k_E\}\Big\} & \text{if } m > 0. \end{cases}$$

The optimal value is calculated as follows if we denote by 1 the root node of the tree

$$\mathrm{OPT} = \min\{F_1(m, K_V, K_E) : m = 0, \dots, n\}. \tag{14}$$

Formulas (10) and (11) are immediate.

In (12) we assume that if $a \in S$ then $aa_s \notin S$: This is without loss of generality, as if $aa_s \in S$, we obtain the same objective value by removing the elements in $S \setminus \{aa_s\}$. The case $m = 0$ corresponds to $a \in S$, which leads to the formula on the first line. The case $m = 1$ occurs when $a_s \in S$ (first argument of the outer minimum on the second line) or $aa_s \in S$ (second argument of the outer minimum). Finally, $m > 1$ is the case in which $a, aa_s \notin S$.

Similar to (12), in (13) we assume that if $a \in S$ then $aa_i \notin S$. The first case ($m = 0$) corresponds to having $a \in S$. In this situation, we take the sum of the optimal values that we can obtain in each of the two subtrees T_{a_i} and $T_{a_{i+1,s}}$. For the second case ($m > 0$), in which $a \notin S$, we take the better of two possibilities, which correspond to the two arguments of the outer minimum. For the first possibility, which is when $aa_i \in S$, we take again the sum of the optimal values in each of the two subtrees T_{a_i} and $T_{a_{i+1,s}}$. For the second possibility ($aa_i \notin S$), we have to add the connections between the two subtrees.

References

[1] Cohen R., Ben Avraham D., Havlin S., *Efficient immunization strategies for computer networks and populations.* Physical Review Letters, 91: 247901–247905 (2003).

[2] Commander C.W., Pardalos P.M., Ryabchenko V., Uryasev S., Zrazhevsky G., *The wireless network jamming problem.* Journal of Combinatorial Optimization, 14: 481–498 (2007).

[3] Elefteriadou L., "Highway capacity". In Kutz M. (ed.) *Handbook of transportation engineering*, New York, McGraw-Hill, chapter 8 (2004).

[4] Krebs V., "Uncloaking terrorist networks". First Monday, 7 (2002).

[5] Zhou T., Fu Z.-Q., Wang B.-H., *Epidemic dynamics on complex networks.* Progress in Natural Science, 16: 452–457 (2006).

Optimizing smooth objectives on convex sets without projections

Damiano Zeffiro (*)

Abstract. The well known gradient descent method for smooth unconstrained optimization can be extended in a straightforward way to problems with convex constraints by using projections. However, in many cases there are more effective ways to generate feasible descent directions. One of the most popular alternatives to the projected gradient method is the Frank-Wolfe method, characterized by a linear minimization subproblem replacing the projection subproblem.
In this seminar, after a brief review of the above mentioned methods, some examples of sets commonly used in optimization where linear minimization is cheaper than projection will be discussed. Then, variants to improve the convergence rate of the Frank-Wolfe method will be presented, together with a general framework to study such variants. Finally, an algorithm for fast cluster detection in networks based on a Frank-Wolfe variant will be described.

1 Introduction

We discuss in these notes some iterative methods for the solution of the problem

$$\begin{aligned} &\min_x f(x) \\ &s.t. \quad x \in \Omega \end{aligned} \tag{1.1}$$

with Ω a convex and closed subset of $\mathbb{R}^n$, and $f : \Omega \to \mathbb{R}$ a differentiable function with Lipschitz continuous gradient with constant L, that is

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

for every $x, y \in \Omega$.

The classic gradient descent method [10] for unconstrained optimization can be extended in a straightforward way to problems of the form (1.1) by performing a projection on Ω at every iteration to mantain feasibility. However, in many cases there are more effective ways to improve the objective by finding feasible descent directions. Perhaps

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: zeffiro@math.unipd.it. Seminar held on 19 January 2022.

the most popular strategy is the Frank-Wolfe (FW) method [6], characterized by a linear minimization subproblem replacing the projection subproblem. In these notes, after some basic examples and definitions, we present a framework recently introduced in [12] to study variants of this method.
The notes are structured as follows: after introducing a general algorithmic framework for smooth constrained optimization in section 2, we review some examples of sets where linear minimization is cheaper than projecting in section 3. Some classic strategies to improve the convergence rate of the Frank-Wolfe method are reported in section 5, followed by the framework of [12] in section 6. The notes then conclude with the main results derived with the framework in section 7, in particular some improvements on the convergence rates of Frank Wolfe variants and an application to a cluster detection problem in networks.

2 General scheme

The methods we consider in these note generate a sequence of feasible points $\{x_k\}$ as described in Algorithm 1, converging to a first order stationary point x^*.

Algorithm 1 `First order method for problem (1.1).`

1 Choose a point $x_0 \in \Omega$
2 For $k = 0, \dots$
3 If x_k satisfies some specific condition, then STOP
4 Choose a feasible direction d_k
5 Set $x_{k+1} = x_k + \alpha_k d_k$, with $\alpha_k \in (0, \alpha_k^{max}]$
suitably chosen stepsize
6 End for

All the algorithms following the scheme above are clearly in the class of first order methods, that is the methods use only the gradient ∇f and the objective f to choose the next iterate. The stepsize α_k can be for instance chosen by linesearch in the interval $(0, \alpha_k^{max}]$.
We now discuss two basic instances of the above scheme.

2.1 Gradient descent

In the case where $\Omega = \mathbb{R}^n$, the well known gradient descent method [10] follows scheme 1 with $d_k = -\nabla f(x_k)$. The method can be extended to a generic convex Ω by projecting, i.e. setting

$$x_{k+1} = P_\Omega(x_k - \bar{\alpha}_k \nabla f(x_k)) \tag{2.1}$$

for $P_\Omega(y) = \text{argmin}_{z\in\Omega}\|y - z\|$ the projection operator on Ω. Equivalently, x_{k+1} minimizes a linearized and regularized objective:

$$x_{k+1} \in \text{argmin}_{z\in\Omega} f(x_k) + \langle \nabla f(x_k), z - x_k\rangle + \frac{\bar{\alpha}_k}{2}\|z - x_k\|^2 . \tag{2.2}$$

The projection P_Ω is uniquely defined since Ω is convex.

2.2 The Frank Wolfe method

In the Frank-Wolfe method the descent direction d_k is defined by

$$d_k = d_k^{\mathrm{FW}} := s_k - x_k \text{ with } s_k \in \operatorname{argmin}_{s\in\Omega}\langle \nabla f(x_k), s\rangle \,. \tag{2.3}$$

Thus the descent direction d_k points toward a minimizer s_k of the linearized objective:

$$s_k = \mathrm{LMO}_\Omega(\nabla f(x_k)) \in \operatorname{argmin}_{x\in\Omega} f(x_k) + \langle \nabla f(x_k), x - x_k\rangle \tag{2.4}$$

where the operator LMO_Ω used to compute s_k is typically referred to as linear minimization oracle [9].
One important property of the FW method is that it finds sparse approximate solutions, a feature that has lead to applications in many large scale optimization problems including maximum clique, SVM training, minimum enclosing ball, traffic assignment, submodular optimization, etc. (see, e.g., [7, 3] for surveys on the method and its applications).

3 Projection and linear minimization

To motivate the FW method, we report in this section some examples from [5] of sets commonly used in optimization where linear minimization is faster than projection.

Set $\mathcal{C}$	**Linear minimization**	**Projection**
ℓ_p-ball, $p \in \{1, 2, +\infty\}$	$\mathcal{O}(n)$	$\mathcal{O}(n)$
ℓ_p-ball, $p \in]1, 2[\cup]2, +\infty[$	$\mathcal{O}(n)$	$\mathcal{O}(n\rho^2 \|y - x^*\|_2^2/\varepsilon^2)$
Nuclear norm-ball	$\mathcal{O}(\nu \ln(m+n)\sqrt{\sigma_1}/\sqrt{\varepsilon})$	$\mathcal{O}(mn \min\{m, n\})$
Flow polytope	$\mathcal{O}(m+n)$	$\tilde{\mathcal{O}}(m^3 n + n^2)$
Birkhoff polytope	$\mathcal{O}(n^3)$	$\mathcal{O}(n^2 d_z^2/\varepsilon^2)$
Permutahedron	$\mathcal{O}(n \ln(n))$	$\mathcal{O}(n \ln(n) + n)$

Table 1. Complexities of linear minimization and projection (see [5, Table 1] for details).

3.1 l^p ball

Consider the case where Ω is the l^p ball:

$$\Omega = \{x \in \mathbb{R}^n \mid \sqrt[p]{\left(\sum |x_i|^p\right)} = \|x\|_p \leq 1\}$$

For $p = 1, 2, +\infty$, there are fast $O(n)$ algorithms for both projection and linear optimization, with linear optimization outperforming projection for $p = 1$ in large dimensions.
For $p \in (1, 2) \cup (2, \infty)$ linear minimization still costs $O(n)$, since we have the closed form expression

$$\mathrm{LMO}_\Omega(c) = (\operatorname{sign}(c_i)\frac{|c_i|^{\frac{p^*}{p}}}{\|c\|_{p^*}^{p^*}})_{i=1}^n \tag{3.1}$$

for $p^* = \frac{p}{p-1}$. Instead, projections must be approximated via an iterative method. In [5] for instance a method with $O(n/\varepsilon^2)$ complexity is proposed, for ε desired precision on the solution.

3.2 Nuclear norm ball

Consider now the case where $\Omega \subset \mathbb{R}^{n\times m}$ is the nuclear norm ball, that is the set of matrices with sum of singular values at most 1:

(3.2)
$$\Omega = \{X \in \mathbb{R}^{n\times m} \mid \sum \Sigma_{ii} \leq 1, \text{ where } U\Sigma V^t \text{ is the singular value decomposition of } X\}\,.$$

In this case, we have

$$\mathrm{LMO}_\Omega(X) = uv^t \tag{3.3}$$

with u, v top left and right singular vector of X, while

$$P_\Omega(X) = U\hat{\Sigma}V^t \tag{3.4}$$

for $\hat{\Sigma}$ diagonal matrix with $\mathrm{diag}(\hat{\Sigma})$ projection on the simplex of $\mathrm{diag}(\Sigma)$. In particular, computing P_Ω requires the full SVD decomposition, while computing LMO_Ω only requires the top left and right singular vector.

4 Convergence rates

We assume in the rest of these notes that the objective f is convex and that the feasible set Ω is compact, with x^* global minimizer of f in Ω. Recall that f is said to be strongly convex if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2 \tag{4.1}$$

for every $x, y \in \mathbb{R}^n$. We have the following convergence results for the projected gradient method and the FW method respectively (see, e.g., see [10] and [3]).

Proposition 4.1 *If f is convex then for the projected gradient method*

$$f(x_k) - f(x^*) = O(1/k)\,. \tag{4.2}$$

Furthermore, if f is strongly convex

$$f(x_k) - f(x^*) = O\left(\left(1 - \frac{\mu}{L}\right)^k\right)\,. \tag{4.3}$$

Proposition 4.2 *If f is convex and $\{x_k\}$ is generated by the Frank Wolfe method*

$$f(x_k) - f(x^*) = O(1/k)\,. \tag{4.4}$$

Furthermore, this rate is optimal even for the class of strongly convex objectives.

Thus, the FW method has a slower convergence rate for strongly convex objectives. The reason why this is the case is a well understood zig zagging behaviour ([9], [3]), taking place when the method approaches a solution on the boundary (see Figure 1).

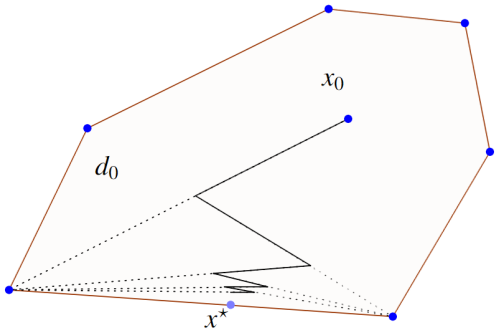


Figure 1. FW method approaching a solution on the boundary.

5 FW variants

The zig zagging behaviour of the classic FW method has motivated the introduction of variants considering different feasible directions instead of the FW direction. A popular choice [9, 3] is to consider away steps, that is steps pointing away from a bad vertex v_k in the minimal face $\mathcal{F}(x_k)$ containing the current iterate x_k:

$$d_k = d_k^{AS} = x_k - v_k\,. \tag{5.1}$$

There are several different strategies to select v_k [3]. For instance, in the special case of the Frank Wolfe method with in face directions (FDFW), we have

$$v_k \in \operatorname{argmax}_{v\in\mathcal{F}(x_k)}\langle \nabla f(x_k), v\rangle \tag{5.2}$$

and

$$d_k = \begin{cases} d_k^{FW} & \text{if } \langle -\nabla f(x_k), d_k^{FW}\rangle \geq \langle -\nabla f(x_k), d_k^{AS}\rangle \\ d_k^{AS} & \text{otherwise.} \end{cases} \tag{5.3}$$

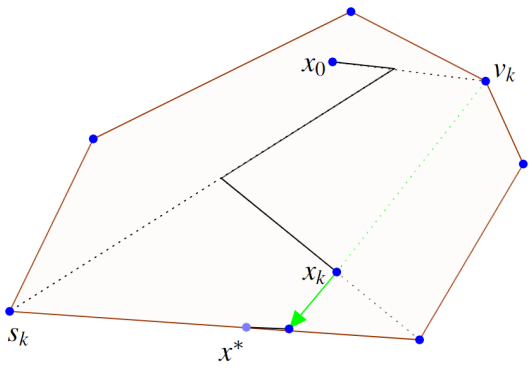


Figure 2. Behavior of the Frank Wolfe method with in face directions.

We have the following properties [2].

Proposition 5.1 *If Ω is a polytope, f is $\mu-$strongly convex and $\{x_k\}$ is generated by the FDFW, then $f(x_k) - f(x^*) \to 0$ with a linear convergence rate dependent only from μ, L and Ω. Furthermore, if strict complementarity conditions hold at x^*, $\mathcal{F}(x^*)$ is identified in a finite number of iterations, or in other words $x_k \in \mathcal{F}(x^*)$ for k large enough.*

While several Frank Wolfe variants solve the issue of zig zagging, they can still exhibit a slow initial convergence for non convex objectives (see Figure 3). This is due to bad steps, defined as maximal steps (i.e. for which $\alpha_k = \alpha_k^{\max}$) that are not FW steps. When a Frank Wolfe variants does a bad step, it can "waste" a gradient computation and a linear minimization without improving much the value of the objective.

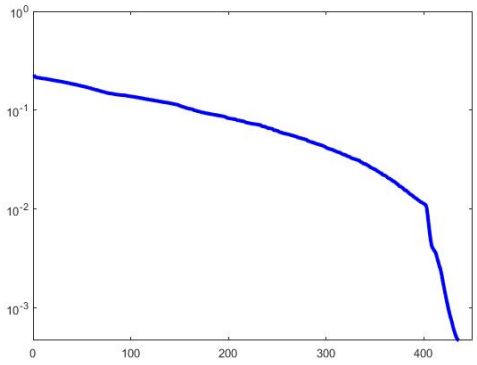


Figure 3. Iterations vs error for the FDFW applied to a continuous formulation of the max clique problem.

6 Framework for linearly convergent projection free optimization

We describe in this section the framework for linearly convergent FW variants first introduced in the paper [12]. The main motivation of the framework is to give a unifying analysis for linearly convergent FW variants, and to define a procedure that does not waste gradient and LMO computations in bad steps. This is achieved by specifying an angle condition ruling out the zig zagging behaviour, and by defining a procedure to recycle gradient and LMO computations in consecutive bad steps.

6.1 Angle condition

In the unconstrained case, a well known condition [1] to ensure convergence of a first order method following the general scheme 1 is that the angle between the gradient $-\nabla f(x_k)$ and the descent direction d_k must not exceed a certain threshold. In other words, for every k and some fixed $\tau > 0$:

$$\frac{\langle d_k, -\nabla f(x_k)\rangle}{\|d_k\|\|\nabla f(x_k)\|} \geq \tau\,. \tag{6.1}$$

In order to extend the condition to the constrained case, we first need to introduce the tangent cone.

Definition 6.1 For $x \in \Omega$, let

$$T_\Omega(x) = \mathrm{cl}\{v \in \mathbb{R}^n \mid x + \lambda v \in \Omega \text{ for some } \lambda > 0\}$$

be the tangent cone to Ω in x.

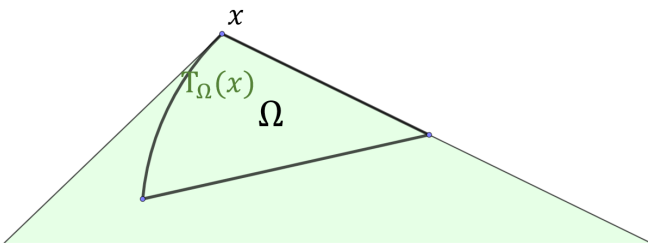


Figure 4. Tangent cone.

We can now define the angle condition in the constrained case.

Definition 6.2 We say that a method following the scheme 1 satisfies the angle condition at iteration k and for $\tau > 0$ if

$$\frac{\langle d_k, -\nabla f(x_k)\rangle}{\|d_k\|\|\pi(x_k, -\nabla f(x_k))\|} \geq \tau\,, \tag{6.2}$$

with $\pi(x, \cdot)$ projection on $T_\Omega(x)$. A method satisfies the angle condition for $\tau > 0$ if (6.2) holds for every k.

As an example, the following proposition states that the FW direction satisfies the angle condition for any fixed k, with a bound going to 0 if $\{x_k\}$ converges to the boundary from the interior, as expected given the zig zagging behaviour.

Proposition 6.3 *If Ω is a polytope and $x_k \in \Omega$, and $d_k = d_k^{FW}$, then the angle condition is satisfied with $\tau = \beta_k/D$, for D the diameter of the polytope and*

$$\beta_k := \min_{\substack{F \in \mathrm{faces}(\Omega),\\ x_k \notin F}} \mathrm{dist}(x_k, F)\,. \tag{6.3}$$

We have the following convergence theorem.

Theorem 6.4 *Under the angle condition, if f is strongly convex and x^* is the solution of problem (1.1)*

$$f(x_k) - f(x^*) = O\left(\left(1 + \tau^2 \frac{\mu}{L}\right)^{-\bar\gamma(k)}\right)\,, \tag{6.4}$$

for $\bar\gamma(k)$ number of good steps.

We remark that the above theorem holds also without assuming convexity, under the gradient inequality

$$\frac{\|\pi(x, -\nabla f(x))\|^2}{2\mu} \geq f(x) - f(x^*)\,.$$

6.2 Short step chain

The short step chain (SSC) procedure performs several steps keeping the gradient fixed, until a sufficient progress is achieved (we refer the reader to the paper [11] for details).

The following is the main convergence result proved in the paper for methods satisfying the conditions of the framework and using the SSC.

Theorem 6.5 *If f is strongly convex, a first order method with SSC always finite and satisying the angle condition converges with*

$$f(x_k) - f(x^*) = O\left(\left(1 + \frac{\mu}{L}\frac{\tau^2}{(1+\tau)^2}\right)^{-k}\right)\,. \tag{6.5}$$

7 Main results

7.1 Improving the rates of convergence of some FW variants

A consequence of Theorem 6.5 is that by applying the SSC we can improve the rates of existing methods as shown in Table 2 (see also [3] for a description of the related FW variants).

Algorithm	Objective	$\gamma(k)$	I_b	h_k/h_0 upper bound
AFW	SC	$k/2$	$\lvert S_0\rvert - 1$	$\left(1 - \frac{\mu}{L}\frac{\tau_p^2}{4}\right)^{\frac{k}{2}}$
PFW	SC	$k/(3\lvert A\rvert! + 1)$	-	$\left(1 - \frac{\mu}{L}\tau_p^2\right)^{\frac{k}{3\lvert A\rvert!+1}}$
FDFW	SC	$k/(\Delta(\Omega) + 1)$	$\dim(\mathcal{F}(x_0))$	$\left(1 - \frac{\mu}{L}\frac{\tau_v^2}{4}\right)^{\frac{k}{\Delta(\Omega)+1}}$
AFW + SSC	NC, KL	k	-	$\left(1 + \frac{\mu}{L}\frac{\tau_p^2}{(2+\tau_p)^2}\right)^{-k}$
PFW + SSC	NC, KL	k	-	$\left(1 + \frac{\mu}{L}\frac{\tau_p^2}{(1+\tau_p)^2}\right)^{-k}$
FDFW + SSC	NC, KL	k	-	$\left(1 + \frac{\mu}{L}\frac{\tau_v^2}{(1+\tau_v)^2}\right)^{-k}$

Table 2. $\Omega = \mathrm{conv}(A)$ with $|A| < \infty$. SC = strongly convex, NC = non convex, KL = KL property. $\gamma(k)$: lower bound on the number of good steps after k steps, counting from the first good step. I_b: bound on the number of bad steps before the first good step. h_k/h_0 upper bound: worst case rate assuming no initial bad steps. $\Delta(\Omega)$ = maximum increase in face dimension after a FW step. S_0 = active set for x_0. $\tau_p, \tau_p/2$ and τ_v are the angle condition constants for the AFW, PFW and FDFW respectively.

7.2 s-defective cliques

Let $\mathcal{G} = (V, E)$ be a graph with vertices V and edges E. A subset C of V is an $s-$ defective clique if at most s links are missing between vertices in C:

$$|\binom{C}{2} \cap E| \geq |\binom{C}{2}| - s$$

Then the maximum $s-$defective clique problem consists in finding an $s-$defective clique of maximum cardinality. We applied a first order method with SSC to (a regularized version of) the following continuous cubic formulation of this problem from [13]:

$$\max\{x^\intercal (A_{\mathcal{G}} + A(y))x \mid (x, y) \in \Delta_{|V|-1} \times \mathcal{D}'_s(\mathcal{G})\}, \tag{7.1}$$

with

- $\bar{E} = \binom{E}{2} \setminus E$.
- $D'_s(\mathcal{G}) = \{y \in [0,1]^{\bar{E}} \mid e^\intercal y \leq s\}$
- $\Delta_{|V|-1}$ the $|V| - 1$ dimensional simplex
- $A_{\mathcal{G}}$ adjacency matrix of $\mathcal{G}$
- $A(y)_{ij} = y_{\{i,j\}}$ for $\{i, j\} \in \bar{E}$, 0 otherwise (the so colled "fake edge matrix").

The method we applied combines the FDFW on the x components, with the FW method on the y component.

Algorithm 2 `FW for` $s-$`defective clique (FWdc)`

1 Choose $z_0 := (x_0, y_0) \in \mathcal{P}$, $k := 0$
2 If z_k is stationary then STOP
3 Compute a descent direction d_k^x on the x component with the FDFW
4 Compute a descent direction d_k^y on the y component with FW
5 If $\langle \nabla_y h_{\mathcal{G}}(x_k, y_k), d_k^y \rangle \geq \langle \nabla_x h_{\mathcal{G}}(x_k, y_k), d_k^x \rangle$ then:
6 set $x_{k+1} = x_k$, $y_{k+1} = y_k + d_k^y$.
7 Else set $x_{k+1} = x_k + \alpha_k d_k^x$, $y_{k+1} = y_k$.
8 Set $k := k + 1$. Go to step 2.

By results proved in [4], the above method always identifies an $s-$defective clique in a finite number of iterations, provided that it does not converge to a saddle point (which is never an issue in numerical tests; see also [8]). Moreover, Algorithm 2 outperforms the approach proposed in [13], based on the CONOPT solver and a combinatorial procedure to process solutions (see [4] for numerical results on a simplified variant of Algorithm 2).

References

[1] Pierre-Antoine Absil, Robert Mahony, and Benjamin Andrews, *Convergence of the iterates of descent methods for analytic cost functions.* SIAM Journal on Optimization, 16(2): 531–547, 2005.

[2] Mohammad Ali Bashiri and Xinhua Zhang, *Decomposition-invariant conditional gradient for general polytopes with line search.* In NIPS, pages 2690–2700, 2017.

[3] Immanuel Bomze, Francesco Rinaldi, and Damiano Zeffiro, *Frank-Wolfe and friends: a journey into projection-free first-order optimization methods.* 4OR, 19: 313–345, 2021.

[4] Immanuel Bomze, Francesco Rinaldi, and Damiano Zeffiro, *Fast cluster detection in networks by first-order optimization.* SIAM Journal on Mathematics of Data Science 4/1 (2022), 285–305.

[5] Cyrille W Combettes and Sebastian Pokutta, *Complexity of linear minimization and projection on some sets.* Operations Research Letters, 2021.

[6] Marguerite Frank and Philip Wolfe, *An algorithm for quadratic programming.* Naval research logistics quarterly, 3(1-2): 95–110, 1956.

[7] Martin Jaggi, *Revisiting Frank-Wolfe: Projection-free sparse convex optimization.* In ICML (1), pages 427–435, 2013.

[8] Chi Jin, Praneeth Netrapalli, and Michael I Jordan, *Accelerated gradient descent escapes saddle points faster than gradient descent.* In Conference On Learning Theory, pages 1042–1085. PMLR, 2018.

[9] Simon Lacoste-Julien and Martin Jaggi, *On the global linear convergence of Frank-Wolfe optimization variants.* arXiv preprint arXiv:1511.05932, 2015.

[10] Yurii Nesterov et al, "Lectures on convex optimization". Volume 137. Springer, 2018.

[11] Francesco Rinaldi and Damiano Zeffiro, *Avoiding bad steps in Frank Wolfe variants.* arXiv preprint arXiv:2012.12737, 2020.

[12] Francesco Rinaldi and Damiano Zeffiro, *A unifying framework for the analysis of projection-free first-order methods under a sufficient slope condition.* arXiv preprint arXiv:2008.09781, 2020.

[13] Vladimir Stozhkov, Austin Buchanan, Sergiy Butenko, and Vladimir Boginski, *Continuous cubic formulations for cluster detection problems in networks.* Mathematical Programming, pages 1–29, 2020.

Modular curves and Heegner points

Daniele Troletti (*)

Abstract. One the the main open conjecture is the one due to Birch and Swinnerton-Dyer about elliptic curves. There are many attempts to prove it but they were able to prove only some special cases, like the rank 1 case proven by Kolyvagin using the Heegner points method. This seminar will give an introduction on the basis required to define the Heegner points, such as elliptic and modular curves. After that we are going to define Heegner points and show some results achieved using them.

1 Elliptic curves

One of the main subject of study in modern number theory are the elliptic curves. They are special curves with many interesting properties. A good reference on this topic is [8]. Since we are interested mainly in elliptic curves over $\mathbb{C}$ all the fields will have characteristic zero.

Definition 1.1 An Elliptic curve E is a plane complex curve whose points are the solution of the cubic equation

$$y^2 = 4x^3 - g_2 x - g_3$$

where $g_2, g_3 \in \mathbb{C}$ and $\Delta = g_2^3 - 27g_3^2 \neq 0$.

Such an equation is called the *Weierstrass equation* of the elliptic curve. The quantity Δ is called the discriminant. We have an important invariant associated to every elliptic curve which will play an important role later on:

Definition 1.2 The *j-invatiant* of an elliptic curve is

$$j(E) = \frac{1728 g_2^3}{g_2^3 - 27g_3^2}$$

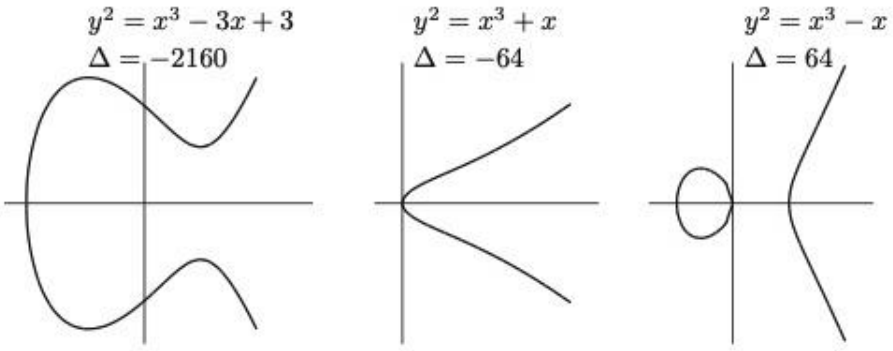
In a more geometric language an elliptic curve is a smooth complex algebraic cubic curve. This class of curves has a special feature that make them interesting in number theory: the set of their points has the structure of an abelian group.

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: daniele.troletti@math.unipd.it . Seminar held on 2 February 2022.

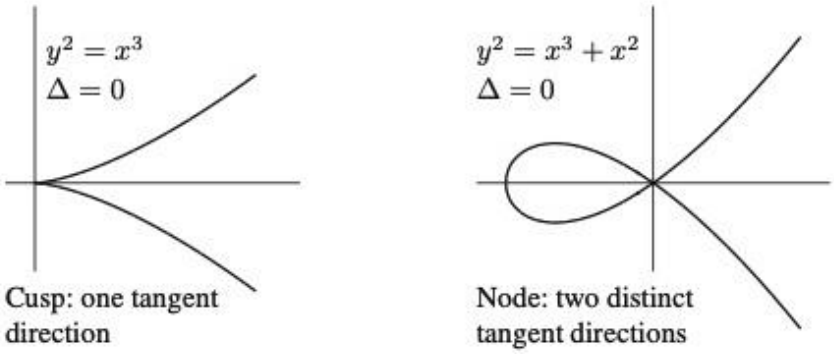
Remark 1.3 Why an elliptic curve has this cubic equation? This derives from the geometric description: wanting a simple object we take a curve (so dimension 1), smooth and with algebraic genus 1. The Riemann-Roch theorem implies that any curve with this characteristics must have an equation like the one of definition 1.1.

In order to define this structure we need to compactify the curve adding a *point at infinity* (using projective geometry, it is the point at infinity of the y-axis). We denote this extra point as ∞. Hence the set of the complex (projective) points of the elliptic curve is

$$E(\mathbb{C}) = \left\{(x, y) \in \mathbb{C}^2 \mid y^2 = 4x^3 - g_2 x - g_3\right\} \cup \{\infty\}$$



(a) Examples of elliptic curves



(b) Examples of singular cubic curves

Figure 1. An example of the real points of some elliptic curves (a) and some cubic curves that are not elliptic (b). Note that the elliptic curves are smooth, while the others have a singular point. Indeed, there is a classification theorem of cubic complex plane curves up to linear change of coordinates: either they are elliptic curves, or they have exactly one singular point that can be a node or a cusp. The two figures are taken from [8, Figures 3.1 and 3.2].

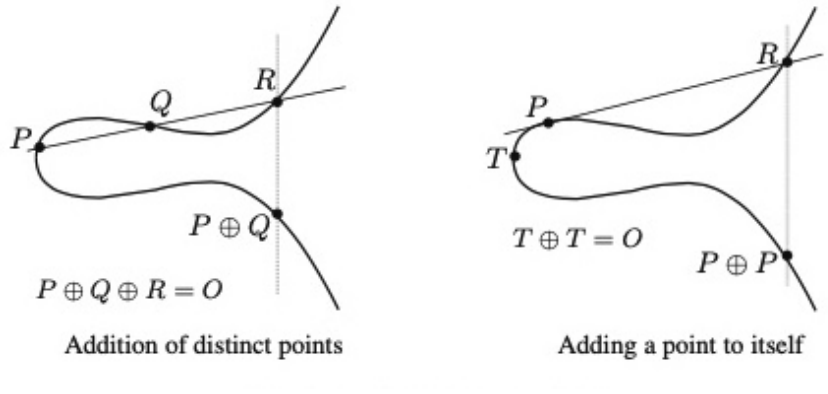


Figure 2. Geometric description of the composition law on the elliptic curve E. On the left we see the sum of two points $P \neq Q$: in this case $P \oplus Q$ is obtained reflecting w.r.t the x-axis the point R, namely the third point in which the line through P and Q meets E. On the right side we see the sum $P \oplus P$: it is the reflection of the other point in which the tangent line at P meets E. This figure is taken from [8, Figure 3.3].

Proposition 1.4 *The composition law described in geometric terms in Figure 2 makes $E(\mathbb{C})$ into an abelian group, whose zero is ∞ and that, in coordinates, is defined algebraically (i.e. by rational functions, in other words by quotients of polynomials).*

A complex projective algebraic variety (i.e. the zero locus into the complex n-th projective space of a set of homogeneous polynomials) whose points form a group (which is necessarily abelian) whose composition law is algebraically defined is called an *abelian variety* over $\mathbb{C}$. Therefore the previous proposition says that elliptic curves are abelian varieties. It turns out that all abelian varieties (over $\mathbb{C}$) of dimension 1 are elliptic curves. We begin to do Number Theory when the equation of an elliptic curve has only rational coefficients (ore more generally when the coefficient are in a number fields, i.e. finite degree extensions of $\mathbb{Q}$). The next definition give us a little bit of language on that.

Definition 1.5 If K is any subfield of $\mathbb{C}$ and $g_2, g_3 \in K$, we say that E is defined over K. In this case for any field extension $K \subseteq L \subseteq \mathbb{C}$ we define

$$E(L) = \left\{(x, y) \in \mathbb{C}^2 \mid y^2 = 4x^3 - g_2 x - g_3, \text{ such that } x, y \in L\right\} \cup \{\infty\} = E(\mathbb{C}) \cap L^2$$

the set of L-rational point of E.

It is easy to see using the expression of the sum in coordinates that $E(L)$ is a subgroup of $E(\mathbb{C})$. One of the most important theorem about this subject is the one regarding the structure of the group $E(L)$, which gives a precise description on how this group is made.

Theorem 1,6 (Mordell-Weil) *Let E be an elliptic curve defined over a number field K, then $E(K)$ is a finitely generated abelian group, hence isomorphic to $E(K)_{tors} \times \mathbb{Z}^r$, where $E(K)_{tors}$ denotes the K-rational torsion points (i.e. points of finite order, that is a point P such that $nP = \infty$ for some $n \in \mathbb{N}, n > 0$), and $r \in \mathbb{N}_{\geq 0}$ is called the* (algebraic) rank *of E.*

The (algebraic) rank of an elliptic curve is an important arithmetic invariant and is the subject of important conjectures, e.g. the Birch and Swinnerton-Dyer conjecture who relates it with the analytic rank which is the order of vanishing of the L-function attached to E at $s = 1$. Indeed this happens because this theorem is not effective, i.e. it does not provide a method for computing the algebraic rank (there are no known finite time algorithm in general).

As always in math, after introducing a class of object we define the maps among them. Since these curves have a group structure it is a natural choice to require some compatibility condition on the functions.

Definition 1.7 Let E_1 and E_2 be elliptic curves. An *isogeny* from E_1 to E_2 is a morphism of algebraic curves $\varphi\colon E_1 \to E_2$ such that $\varphi(\infty) = \infty$.

We can prove that an isogeny either is the zero map or it is surjective. Since the elliptic curves are groups then also the set of the maps between them is a group, we denote the set of isogenies from E_1 to E_2 as

$$\mathrm{Hom}(E_1, E_2) = \{\text{isoginies } E_1 \to E_2\}$$

The sum is of isogenies is defined pointwise. We define the *endomorphism ring* of E as the group $\mathrm{End}(E) = \mathrm{Hom}(E, E)$ endowed with the composition of functions as the multiplication map.

Example 1.8 Let $n \in \mathbb{Z}$, the *multiplication-by-n isogeny*

$$[n]\colon E \to E$$

is defined in the natural way:

$$[m]P = \begin{cases} P + \cdots + P \ m \text{ times} & \text{if } m > 0, \\ \infty & \text{if } m = 0, \\ (-P) + \cdots + (-P) \ m \text{ times} & \text{if } m < 0. \end{cases}$$

In the case of elliptic curves we can find the structure of the endomorphism ring.

Definition 1.9 Let $\mathcal{K}$ be a $\mathbb{Q}$-algebra that is finitely generated over $\mathbb{Q}$. An *order* $\mathcal{R}$ of $\mathcal{K}$ is a subring of $\mathcal{K}$ that is finitely generated as a $\mathbb{Z}$-module and satisfies $\mathcal{R} \otimes \mathbb{Q} = \mathcal{K}$.

Example 1.10 Let F be a number field, i.e. an extension of $\mathbb{Q}$, then its ring of integers is an order. Suppose $F = \mathbb{Q}[i]$ then $\mathcal{R} = \mathbb{Z} + \mathbb{Z}i$ is an order of $\mathbb{Q}[i]$.

Remark 1.11 If the algebra $\mathcal{K}$ is an imaginary quadratic field, i.e. a field of the form $\mathbb{Q}(\sqrt{d})$ for some $d \in \mathbb{Z}$, $d < 0$, and $\mathcal{O}_\mathcal{K}$ is its ring of integers then all the orders has the form $\mathcal{R} = \mathbb{Z} + c\mathcal{O}_\mathcal{K}$, where c is a positive integer. We call c the conductor of the order $\mathcal{R}$.

Theorem 1.12 *Let E be an elliptic curve, then* $\mathrm{End}(E)$ *is either* $\mathbb{Z}$ *or an order in an imaginary quadratic field.*

Many properties of E depends on the endomorphism ring, in particular:

Definition 1.13 Let E be an elliptic curve, $\mathcal{R}$ an order in an imaginary quadratic field K. We say that E has *complex multiplication* by $\mathcal{R}$ if $\mathrm{End}(E) = \mathcal{R}$

This property is not extremely rare and elliptic curves with CM are really nice to study and used in many construction, for example the Heegner points we are going to define later on. We now give an explicit example of an elliptic curve with CM.

Example 1.14 The curve E defined by the equation

$$y^2 = x^3 + x$$

is an elliptic curve with complex multiplication. We can make a change of coordinates and transform the equation in the form used in the definition but one of the two coefficient is not over $\mathbb{Q}$ and has a complex expression: $g_3 = 0$ and

$$g_2 = 64 \left(\int_0^1 \frac{dt}{\sqrt{1-t^4}} \right)^4$$

We can compute the j-invariant and we get $j(E) = 1728$. We can consider the morphism

$$\psi(x, y) = (-x, iy)$$

winch is an endomorphism since $(iy)^2 = (-x)^3 + (-x)$. Furthermore ψ^4 is the identity map and it is not a multiplication-by-n map for any n. So we have that $\mathrm{End}(E) \cong \mathbb{Z}[i]$ and E has complex multiplication.

2 Modular curves

In the previous section we have seen elliptic curves under an algebraic viewpoint, but there is also another approach to them: we can describe them as Riemann surfaces, i.e. complex differentiable manifold of dimension 1. A good reference about this description of elliptic curves is [1].

Definition 2.1 A Riemann surface is a topological space X such that locally at any point $x \in X$ there is an homeomorphism $\varphi \colon U \to V$ of an open neighbourhood U of x with an open subset V of $\mathbb{C}$ and such that the "local charts" are compatible, in the sense that the "change of coordinate maps" $\psi \circ \varphi^{-1} \colon \varphi^{-1}(U \cap U') \to \psi(U \cap U')$ are biholomorphic maps (holomorphic and bijective). Here $\varphi \colon U \to V$ and $\psi \colon U' \to V'$ are two local charts such that $U \cap U' \neq \emptyset$. The set of local charts of a Riemann surface is called a ?complex atlas?.

Example 2.2 The most trivial example of Riemann surface is of course $\mathbb{C}$ itself, with the identity as the unique chart; another easy example is the so called Riemann Sphere,

that is a sphere covered by two local charts: the two stereographic projections, one defined on the whole sphere without the North pole, the other on the sphere without the South one. Note that the latter example is a compact Riemann surface, that is topologically the compactification with one point of the first one.

Definition 2.3 A lattice Λ in $\mathbb{C}$ is a free subgroup of rank 2, i.e. a subset of $\mathbb{C}$ of the form $\Lambda = \mathbb{Z}\omega_1 \oplus \mathbb{Z}\omega_2$, where $\omega_1, \omega_2 \in \mathbb{C}$, such that $\mathrm{Im}(\omega_2/\omega_1) > 0$. The quotient group

$$\mathbb{C}/\Lambda = \{x + \Lambda \mid x \in \mathbb{C}\}$$

endowed with the quotient topology can be given a complex atlas. The resulting Riemann surface is called a complex torus (it is indeed topologically a torus).

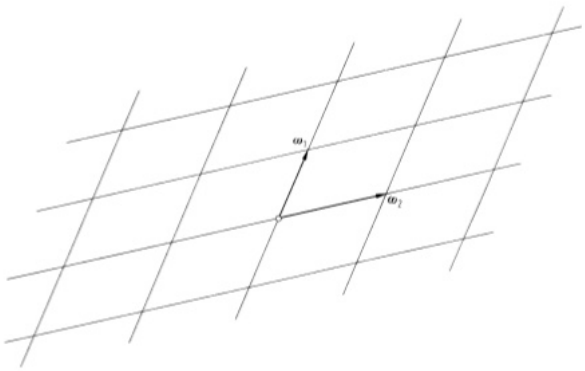


Figure 3. A lattice in the complex plane.

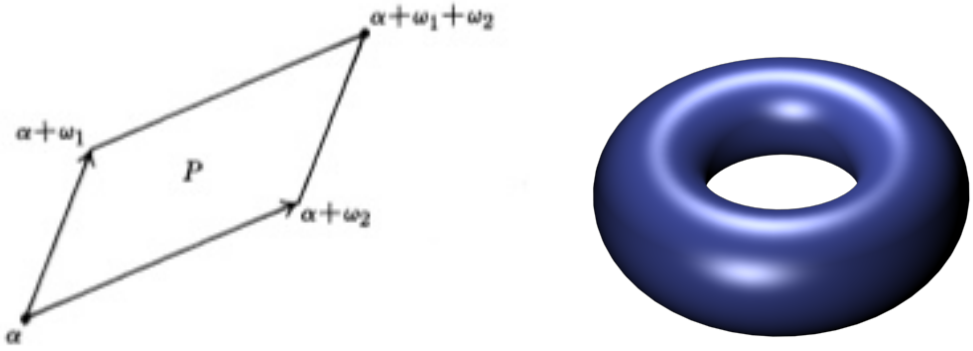


Figure 4. A fundamental parallelogram (a): up to the identification of the opposite sides it represent the quotient space $\mathbb{C}/\Lambda$. Once identified the opposite sides it becomes a topological torus (b).

The important result is that, using the Weierstrass $\wp$ function, any complex torus can be embedded into the complex projective plane in a biholomorphic way as an elliptic curve and all elliptic curves are obtained in this way, hence as Riemann surfaces elliptic curves and complex tori are the same thing.

Definition 2.4 Let Λ be a lattice, then the *Weierstrass $\wp$-function* associated to Λ is defined as

$$\wp(z,\Lambda) = \frac{1}{z^2} + \sum_{\substack{\omega\in\Lambda\\ \omega\neq 0}} \left(\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2} \right)$$

The *Eisenstein series of weight* $2k$ associated to Λ is the series

$$G_{2k}(\Lambda) = \sum_{\substack{\omega\in\Lambda\\ \omega\neq 0}} \omega^{-2k}$$

Theorem 2.5 *Let Λ be a lattice in $\mathbb{C}$, then:*

- *The Eisenstein series $G_{2k}(\Lambda)$ is absolutely convergent for all $k > 1$.*
- *The series defining the Weierstrass $\wp$-function converges absolutely and uniformly on every compact subset of $\mathbb{C} \setminus \Lambda$. The series defines a meromorphic function on $\mathbb{C}$ having a double pole with residue 0 at each lattice point and no other poles.*
- *For all $z \in \mathbb{C} \setminus \Lambda$, the Weierstrass $\wp$-function and its derivative satisfy the relation*

$$\wp'(z)^2 = 4\wp(z)^3 - 60G_4(\Lambda)\wp(z) - 140G_6(\Lambda)$$

Thus there exists a map ψ which sends biholomorphically the torus $\mathbb{C}/\Lambda$ to the elliptic curve E with $g_2 = 60G_4(\Lambda)$ and $g_3 = 140G_6(\Lambda)$

$$\begin{aligned} \psi\colon \mathbb{C}/\Lambda &\to E(\mathbb{C}) \\ z &\mapsto \left[\wp(z), \wp'(z), 1\right] \end{aligned}$$

We can construct an inverse to this map and we can prove that it maps biholomorphisms to isogenies, hence classify elliptic curves up to isogeny is the same thing as classify complex tori up to biholomorphism. In particular we are interested in classifying elliptic curves endowed with more structure.

Proposition 2.6 *Let Λ and Λ' be two lattices in $\mathbb{C}$. The two complex tori $\mathbb{C}/\Lambda$ and $\mathbb{C}/\Lambda'$ are biholomorphic if and only if Λ and Λ' are homothetic, i.e. $\Lambda' = \epsilon\Lambda$ for some $\epsilon \in \mathbb{C}^\times$.*

In particular given a lattice $\Lambda = \mathbb{Z}\omega_1 \oplus \mathbb{Z}\omega_2$, we can consider the complex number $\tau = \omega_2/\omega_1$, which has positive imaginary part, and we have that Λ is biholomorphic to the lattice $\Lambda_\tau = \mathbb{Z} \oplus \mathbb{Z}\tau$. Thus any complex tours is biholomorphic to one of the form $E_\tau = \mathbb{C}/\Lambda_\tau$ for some $\tau \in \mathbb{C}$ with $\operatorname{Im}\tau > 0$. Moreover:

Pro2.7 Two complex tori E_τ and $E_{\tau'}$ are biholomorphic if and only if there are four integers $a, b, c, d, \in \mathbb{Z}$ such that $ad - bc = 1$ and

$$\tau' = \frac{a\tau + b}{c\tau + d}$$

With a more sophisticate language this means that classify all elliptic curves up to biholomorphism is equivalent to classify the numbers τ that belong to the complex upper halfplane

$$\mathbb{H} = \{z \in \mathbb{C} \mid \mathrm{Im}(z) > 0\}$$

up to a left action of the special linear group

$$\mathrm{SL}_2(Z) = \left\{ \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M_2(\mathbb{Z}) \;\middle|\; \det(\gamma) = 1 \right\}$$

given by the formula

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \tau = \frac{a\tau + b}{c\tau + d}$$

We will denote the set of orbits by $Y(1) = \mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$. One can prove that $Y(1)$ has the structure of a Riemann surface, that we call an *open Modular curve*. We can see how it works topologically: Figure 5 represent a so called ?fundamental domain?, i.e. a connected subset of the complex plane that (outside the borders) is in bijection with the orbits of the action. Giving to $Y(1)$ the quotient topology is therefore equivalent to glue the left and right borders of the fundamental domain.

Example 2.8 We can consider the elliptic curve of the Example 1.14, its isogeny class corresponds to the point i in the fundamental domain of $Y(1)$.

Other (open) modular curves can be defined as the quotient $\Gamma\backslash\mathbb{H}$ of $\mathbb{H}$ by the induced action of some particular subgroups Γ of $\mathrm{SL}_2(\mathbb{Z})$, called congruence subgroups. We are interested in congruence subgroup of the form

$$\Gamma_0(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) \;\middle|\; c \equiv 0 \mod N \right\}$$

for some positive $N \in \mathbb{Z}$. The modular curve $Y_0(N) = \Gamma_0(N)\backslash\mathbb{H}$ can be interpreted as a classifying space, too:

Proposition 2.9 *The modular curve $Y_0(N) = \Gamma_0(N)\backslash\mathbb{H}$ parametrizes the couples (E, C), where E is an elliptic curve and C is a cyclic subgroup of E of order N.*

Another important features of modular curves is that they can be compactified with a finite number of points, called cusps, in a canonical way, giving rise to a compact Riemann surface, that we call a *closed Modular curve*. The compactification of a modular curve is denoted by the letter X, so e.g. $X_0(N)$ is the compactification of $Y_0(N)$. The importance of $X_0(N)$ is that it can be embedded (biholomorphically) into the complex projective plane as an algebraic curve defined over $\mathbb{Q}$ (i.e. it has an equation with all rational coefficients). Hence in the rest we will treat $X_0(N)$ as such a curve.

3 Modularity

In this brief section we introduce the concept of modularity for elliptic curves, which is the last ingredient we need in order to define the Heegner points.

Definition 3.1 Let E be an elliptic curve defined over $\mathbb{Q}$, then E is *modular* if there is a surjective morphism $\pi\colon X_0(N_E) \to E$ of algebraic curves defined over $\mathbb{Q}$.

Heuristically this means that an elliptic curve is a quotient of a modular curve.

Remark 3.2 The number N_E is the *conductor* of E and it does only depends on E.

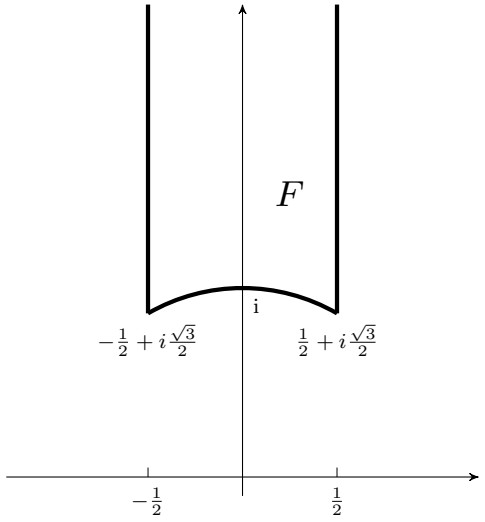


Figure 5. Fundamental domain for $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$.

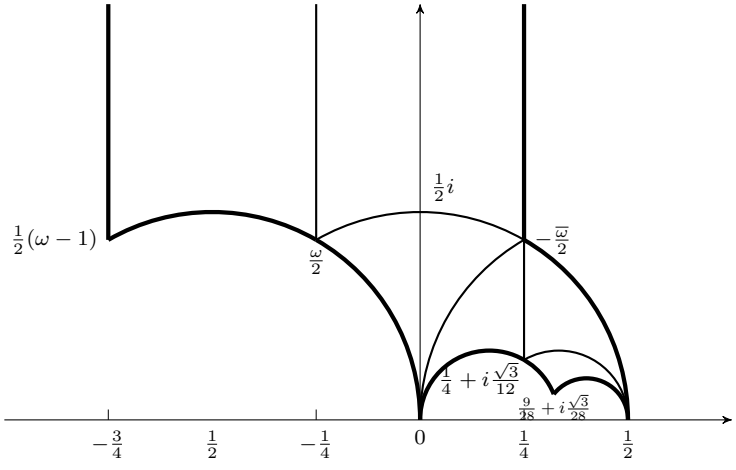


Figure 6. Fundamental domain for $X_0(4)$. Since we are taking a subgroup the fundamental domain is larger and more complex. In the picture $\omega = -\frac{1}{2}+i\frac{\sqrt{3}}{2}$.

Theorem 3.3 (Modularity) *Any elliptic curve E over $\mathbb{Q}$ is modular.*

This important theorem was proved for semistable elliptic curves by Wiles and Taylor in 1993-1995, and it was established in general only in 2001 in a joint paper by Breuil, Conrad, Diamond, Taylor and Richard.

The modularity theorem gives the existence of the modular parametrization π, but getting it explicitly following the idea in the proof of the theorem is extremely hard since it require passing through multiple Galois representation.

4 Heegner points

Let E be an elliptic curve of conductor N defined over $\mathbb{Q}$ and let $d \neq 3, 4$ be a positive integer. Let $K = \mathbb{Q}[\sqrt{-d}]$ be an imaginary quadratic field. We require that d satisfies the *Heegner Hypothesis*: every prime dividing N should split completely in K. By the Modularity Theorem we can fix a modular parametrization $\varphi\colon X_0(N) \to E$.

For every positive $n \in \mathbb{Z}$ we have an order $\mathcal{R}_n$ of conductor n in K. Using the theory of orders and complex multiplication we can construct an elliptic curve E_n with complex multiplication by $\mathrm{End}(E_n) = \mathcal{R}_n$ and a cyclic subgroup C_n of order N. The couple (E_n, C_n) defines a point x_n on the modular curve $X_0(N)$ via the identification of Proposition 2.9.

Definition 4.1 *Heegner points* on the elliptic curve E are the images $y_n = \varphi(x_n)$.

The points y_n are not defined in general over $\mathbb{Q}$ but over an extension (one for each n) of the quadratic field K. By the theory of complex multiplication and class field theory we know that this extension is the so-called ring class field of conductor n which is denoted by K_n.

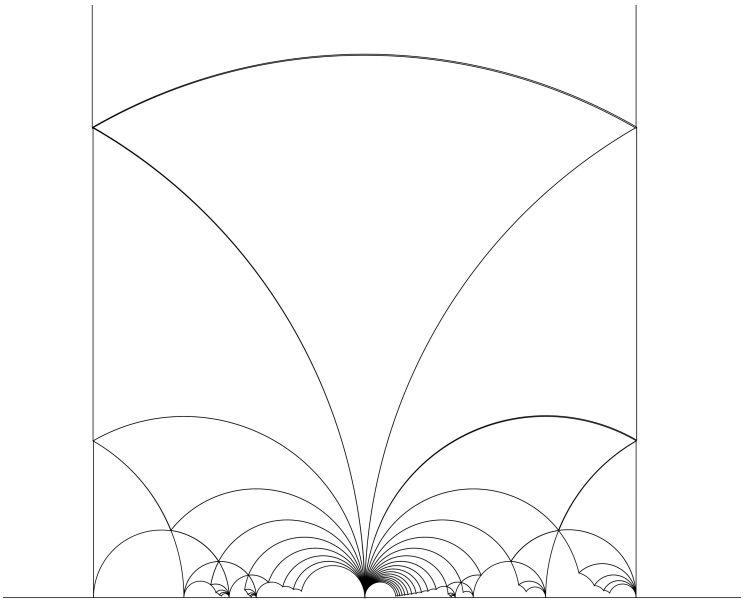


Figure 7. A fundamental domain for $X_0(53)$.

Example 4.2 Consider the elliptic curve $E\colon y^2+xy+y = x^3-x^2$. We can find a modular parametrization $X_0(53) \to E$. Using the technique of [3] we can find the point y_5. Let

$$P(x) = x^6 - 12x^5 + 1980x^4 - 5855x^3 + 6930x^2 - 3852x + 864$$

and let α be one of its roots. We have that the ring class field of conductor 5 is $K_5 = K[\alpha]$. We can compute the affine coordinates of the point y_5 which are

$$y_5 = \left(\alpha, -\frac{4}{315}\alpha^5 + \frac{43}{315}\alpha^4 - \frac{7897}{315}\alpha^3 + \frac{2167}{35}\alpha^2 - \frac{372}{7}\alpha + \frac{544}{35}\right)$$

Of particular interest is the point y_1. We can define the point

$$y_K = \mathrm{Tr}_{K_1/K}(y_1) = \sum_{\sigma \in \mathrm{Gal}(K_1/K)} \sigma(y_1)$$

where $\mathrm{Tr}_{K_1/K}$ the trace map and $\mathrm{Gal}(K_1/K)$ is the group of the automorphisms of K_1 fixing K. If we change the cyclic subgroup C_1 we get another point y_1 but the point y_K changes only up to a minus sign and a torsion point of E.

Theorem 4.3 (Kolyvagin) *If the point y_K has infinite order in $E(K)$, then*

- *The group $E(K)$ has rank* 1,
- *The group* $Ш(E/K)$ *is finite.*

The group $Ш(E/K)$ is an important group associated with an elliptic curve which measure how bad it behaves when trying to compute the algebraic rank of it. It is still open the question whether it is finite or not.

Further works of Kolyvagin and others using Heegner points were able to prove a special case of the Birch and Swinnerton-Dyer conjecture, the one where the algebraic rank of the elliptic curve is 1.

5 Further readings

The classical reference book about elliptic curves and their arithmetic properties is [8].

For more advanced topic, the complex viewpoint on elliptic curve and the theory of modular curve you can look at the first chapter of [7] and [1].

For the theory of complex multiplication and a brief review on class field theory you can look at the second chapter of [7].

If you are interested in the use of Heegner points to prove the rank 1 case of the Birch and Swinnerton-Dyer conjecture you can look at the original articles of Kolyvagin. A simpler article, which still requires some advanced knowledge in number theory, explaining the works of Kolyvagin is [2].

References

[1] Fred Diamond and Jerry Shurman, "A first course in modular forms". Vol. 228, Graduate Texts in Mathematics, Springer-Verlag, New York, 2005, pp. xvi+436, ISBN: 0-387-23229-X.

[2] Benedict H. Gross, *Kolyvagin's work on modular elliptic curves*. In: L-functions and arithmetic (Durham, 1989), vol. 153, London Math. Soc. Lecture Note Ser. Cambridge Univ. Press, Cambridge, 1991, pp. 235–256, DOI: 10.1017/CBO9780511526053.009, URL: https://doi.org/10.1017/CBO9780511526053.009.

[3] Dimitar Jetchev, Kristin Lauter and William Stein, *Explicit Heegner points: Kolyvagin's conjecture and non-trivial elements in the Shafarevich-Tate group*. In: J. Number Theory 129.2 (2009), pp. 284–302, ISSN: 0022-314X, DOI: 10.1016/j.jnt.2008.05.007, URL: https://doi.org/10.1016/j.jnt.2008.05.007.

[4] Neal Koblitz, "Introduction to elliptic curves and modular forms". Second, vol. 97, Graduate Texts in Mathematics, Springer-Verlag, New York, 1993, pp. x+248, ISBN: 0-387-97966-2, DOI: 10.1007/978-1-4612-0909-6, URL: https://doi.org/10.1007/978-1-4612-0909-6.

[5] V.A. Kolyvagin, *Euler systems*. In: The Grothendieck Festschrift, Vol. II, vol. 87, Progr. Math. Birkhäuser Boston, Boston, MA, 1990, pp. 435–483.

[6] Daniel A. Marcus, "Number fields". Universitext, Second edition of [MR0457396], With a foreword by Barry Mazur, Springer, Cham, 2018, pp. xviii+203, ISBN: 978-3-319-90233-3, DOI: 10.1007/978-3-319-90233-3, URL: https://doi.org/10.1007/978-3-319-90233-3.

[7] Joseph H. Silverman, "Advanced topics in the arithmetic of elliptic curves". Vol. 151, Graduate Texts in Mathematics, Springer-Verlag, New York, 1994, pp. xiv+525, ISBN: 0-387-94328-5, DOI: 10.1007/978-1-4612-0851-8, URL: https://doi.org/10.1007/978-1-4612-0851-8.

[8] Joseph H. Silverman, "The arithmetic of elliptic curves". Second, vol. 106, Graduate Texts in Mathematics, Springer, Dordrecht, 2009, pp. xx+513, ISBN: 978-0-387-09493-9, DOI: 10.1007/978-0-387-09494-6, URL: https://doi.org/10.1007/978-0-387-09494-6.

Beyond Nash Equilibria in Mean Field Games

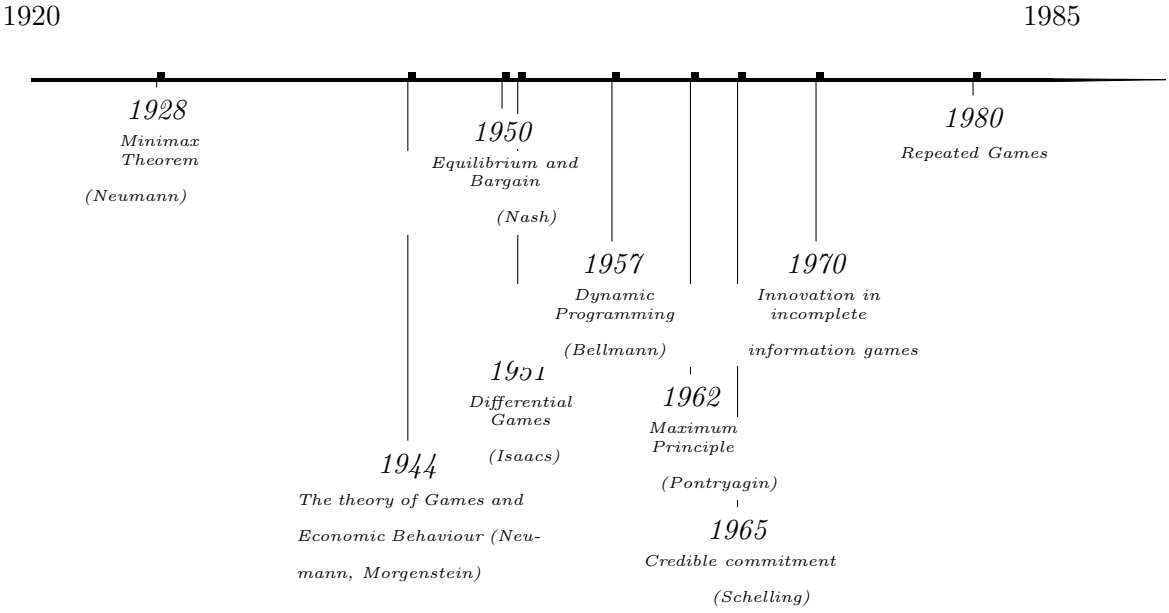
Ofelia Bonesini (*)

Abstract. The concept of *Nash Equilibrium* is the most important (and famous) notion in Game Theory. Assuming that the audience is not familiar with the topic, we will first warm up with an introduction to recall all the basic definitions and results. Then, we will focus on two extensions: *Correlated Equilibria* and *Mean Field Games*. Finally, we will gather things together to see how the definition of a *Correlated solution* can be formulated and its validity checked in the mean field context. Time permitting, I will mention some results of my research.

1 Introduction to Game Theory

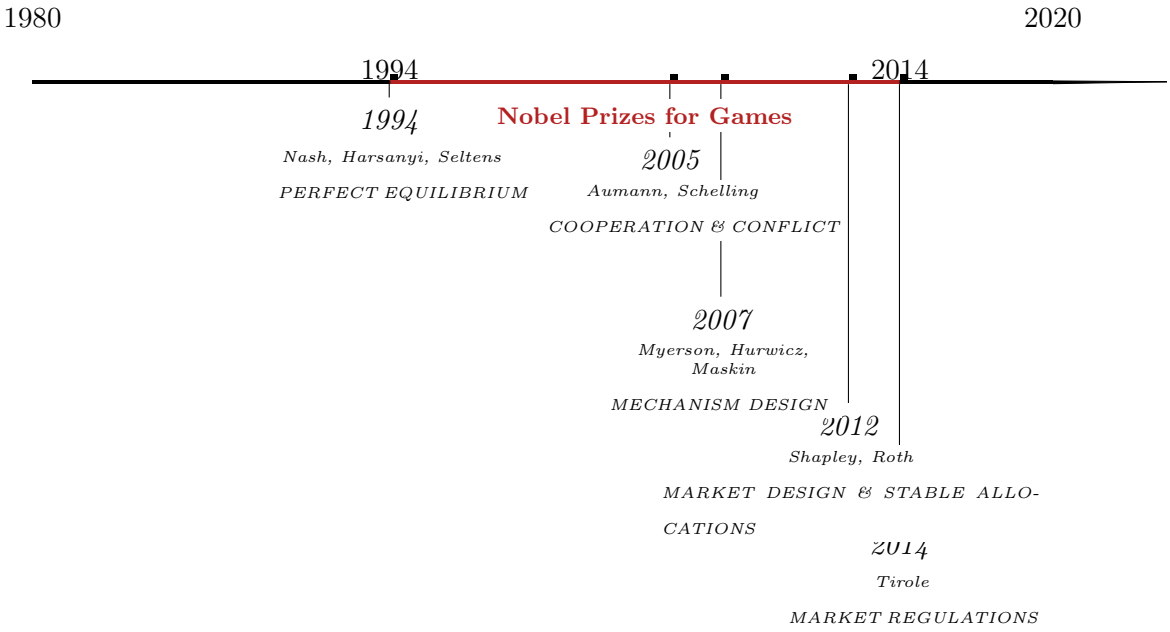
1.1 Historical backgroung and motivations

The purpose of this section is to provide the basic definitions and results of Game Theory. It does not aim at being exhaustive and the interested reader is referred to the first chapters in [2].



(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: bonesini@math.unipd.it . Seminar held on 23 February 2022.

The origins of Game Theory can be dated back to 1928 when Von Neumann proved the *Minimax Theorem*. Since then, the theory has experienced a steady growth over the following 50 years which have been characterized by milestones dates such as 1944 (the publishing of *The theory of Games and Economic Behaviour* by Von Neumann and Morgenstein) or 1950 (the theorization of the concept of Equilibrium within John Nash PhD Thesis). These are also the years in which Dynamic Programming, on one side, and Maximum Principle, on the other, were proved. In a nutshell, we conclude that this dense and full of innovations half-a-century has provided the fertile ground for what is well-portrayed by the following timeline. Indeed, although there is no Nobel Prize for Mathematics, five Nobel prizes in Economics have been assigned to works dealing with innovations in Game Theory, in the twenty years between 1994 and 2014.



This fact points out the key feature of game theory that we are going to discuss in the next section and which has been one of the motivations behind is success: its applicability.

1.2 Applications

For the sake of brevity, here we just list a few fields and some representative examples of applications, but the utility of game theory has no limit. This is a direct consequence of the fact that it provides a context-free set of mathematical tools that can be exploited in any situation in which a decision is made in an interactive context.

- **Theoretical economics:** buyers vs sellers, auctions...
- **Networks**: users vs providers, providers vs providers...
- **Political science**: political parties forming a governing coalition, voting methods and their properties...

- **Military applications**: a missile pursuing a fighter plane...
- **Inspection**: rules breaker vs inspector...
- **Biology**: *Evolutionarily Stable Strategy* (which is a variant of the notion of Nash equilibrium)...
- **Sports, Medicine, Psychology, Environment...**

1.3 Basic Notions

"Game Theory is the name given to the methodology of using mathematical tools to model and analyze situations of interactive decision making." [2].

This is the perfect answer to the question *"What is Game Theory?"*
It is informal but it is able to convey what the key ingredients are: Mathematical modelization and Interaction. Indeed, it is interactivity that distinguishes the discipline from standard control theory and optimization by taking into account the case in which more than one decision maker is pursuing his goals at the same time.

1.3.1 Games in Strategic Form

In order to keep things as simple as possible, we limit ourselves to a static context and to the following formulation[(1)].

Definition 1.1 A *Game in Strategic Form* (or in normal/matrix form) is an ordered triple $\mathbf{G} = (N, (S_i)_{i\in N}, (u_i)_{i\in N})$, where

- $N = \{1, 2, ..., n\}$ the finite set of players;
- For every player $i \in N$:
  - S_i the set of strategies of player i;
    $S = S_1 \times S_2 \times \cdots \times S_n$ the set of all vectors of strategies;
  - $u_i : S \to \mathbb{R}$ a function associating to each $s = (s_i)_{i\in N}$ the utility, or payoff, $u_i(s)$ of player i.

Now, let's introduce a notation that is used several times in the following. Given $s \in S$, we set $s_{-i} := (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n)$, that is the vector of all strategies but s_i.

We end this section presenting a well-known example of game to motivate the name *Matrix Form Games*. This should be considered as a prototype for the examples that are presented in the following. Table 1 represents *Rock, Paper and Scissors game*. We see that the indexes of the rows in the matrix are the strategies of the first player while the indexes of the columns represent the strategies of the second. Each entry of the matrix is a two-dimensional vector where the i-th component, with $i \in \{1, 2\}$, corresponds to the utility

[(1)] Static games are often presented in the so-called *Extended Form*.

function of the i-th player evaluated at the strategies corresponding to that index. Thus, when Player 1 wins over Player 2 their utilities are, respectively, $(1, -1)$, whereas, when Player 2 is the winner, the utility vector reads $(-1, 1)$. Finally, if a tie occurs $u = (0, 0)$.

		Player 2		
		R	P	S
Player 1	R	$(0,0)$	$(-1,1)$	$(1,-1)$
	P	$(1,-1)$	$(0,0)$	$(-1,1)$
	S	$(-1,1)$	$(1,-1)$	$(0,0)$

Table 1. Rock, Paper and Scissors game in matrix form.

1.3.2 Nash Equilibria

Definition 1.2 Consider a vector of strategies $s \in S$. If $u_i(\widehat{s}_i, s_{-i}) > u_i(s)$, we call the strategy $\widehat{s}_i$ a *profitable deviation* of player i at the strategy vector $s \in S$.

Definition 1.3 A strategy vector $s^* = (s_1^*, \dots, s_n^*)$ is defined a *Nash equilibrium* if

$$u_i(s^*) \geq u_i(s_i, s_{-i}^*),$$

for each strategy $s_i \in S_i$ and for each player $i \in N$. Thus, a *Nash equilibrium* is a vector of strategies such that none of the players has a profitable deviation.

An alternative, but completely equivalent, definition can be given in terms of the *best reply*.

Definition 1.4 Denote with s_{-i} a strategy vector of all the players but player i. We say that player i's strategy $\widehat{s}_i$ is a *best reply* to s_{-i} if

$$u_i(\widehat{s}_i, s_{-i}) = \max_{t_i \in S_i} u_i(t_i, s_{-i}).$$

Definition 1.5 A strategy vector $s^* = (s_1^*, \dots, s_n^*)$ is called a *Nash equilibrium* if, for every player $i \in N$, s_i^* is a best reply to s_{-i}^*.

Remark 1.6 A Nash Equilibrium can be equivalently characterized as a fixed point of the best reply map $B : S \to S$, $B(s) := (B_1(s_{-1}), \dots, B_n(s_{-n}))$, with $B_i(t_{-i})$ best reply to t_i for player i.

We end this part presenting two of the most famous examples of strategic form games and the corresponding Nash Equilibria.

The Prisoner's Dilemma which is presented in Table 2 is endowed with the following background story. Two criminals are arrested and accused of a robbery but there is no evidence against them. Hence, the two are asked to confess with the promise of a discount of their penalty. They know that if they both confess they get a small reduction of 1 year (i.e. strategy (D, D)). If they both stay silent refusing to confess (that is strategy (C, C)), since there is no evidence against them they are just sentenced 1 year (w.r.t. 5 total). If one stays mute and the other confesses the one which confesses is set free and the other is condemned the whole period of 5 years (corresponding to strategy (C, D)). In Table 2 the utility functions represent the penalty discount in years. The only Nash equilibrium in this case is given when both the criminals decide to confess. This result might look counterintuitive at a first glance but it is due to the fact that Nash Equilibria are non-cooperative equilibria.

		Player II	
		D	C
Player I	D	$(1,1)$	$(5,0)$
	C	$(0,5)$	$(4,4)$

Table 2. Prisoner's Dilemma.

The last example that we have decided to present is the *Battle of the Sexes*. A couple wants to go out on a Saturday night. They must decide where to go. The girl prefers to go to the ballet (B), while the guy prefers going to a bar to see a football match (S). The payoffs presented in Table 3 can be interpreted as follows. If one goes to one place while the other goes to the other one the two end up being unhappy because they are not together. On the other hand, if they go to the ballet the girl is going to have more fun because she is doing what she wanted but the guy will be enjoying himself too because they are spending time together. In this game the Nash equilibria are two, corresponding to the cases in which the guys are going to the same place, namely strategies (B, B) and (S, S).

		Male	
		B	S
Female	B	$(2,1)$	$(0,0)$
	S	$(0,0)$	$(1,2)$

Table 3. The battle of sexes.

`https://www.youtube.com/watch?v=LJS7Igvk6ZM`

This is the url of a 2-minutes clip of the famous scene of the bar in the film "A beautiful mind". This film is based on the life of John Nash and this scene is supposed to present a Nash Equilibrium. This is not true and can easily checked watching the video and comparing it with what we have discussed above.

1.3.3 Nash Equilibria in Mixed Strategies

What we have presented until now are the so-called *Pure Strategies*. Now, we would like to generalize this idea letting a player randomize his strategy. A randomization is understood in the following sense. We leave the players the possibility (one independently of the others) to toss a coin and decide which strategy to play according to the outcome of the tossing.

Definition 1.7 Consider a strategic-form game $\mathbf{G} = (N, (S_i)_{i\in N}, (u_i)_{i\in N})$ in which the set of pure strategies S_i is nonempty and finite, for every player $i \in N$ and denote the set of pure strategy vectors $S := \times_{i\in N} S_i$. We call *mixed extension* of $\mathbf{G}$ the game

$$\Gamma = (N, (\Sigma_i)_{i\in N}, (U_i)_{i\in N}),$$

in which, for each player i in N:

- $\Sigma_i = \mathcal{P}(S_i)$ is the set of probability measures on S_i;
  $\Sigma := \times_{i\in N}\Sigma_i$;
- $U_i : \Sigma \to \mathbb{R}$, is his utility function function associating to each $\sigma = (\sigma_1, \ldots, \sigma_n)$ the payoff
$$U_i(\sigma) = \mathbb{E}_\sigma[u_i] = \sum_{(s_1,\ldots,s_n)\in S} u_i(s)\sigma_1(s_1)\cdot \ldots \cdot \sigma_n(s_n).$$

Definition 1.8 Consider a strategic-form game $\mathbf{G}$ and its mixed extension Γ. We call *equilibrium in mixed strategies* of $\mathbf{G}$ every equilibrium of Γ.

The following result is a useful instrument for finding equilibria.

Theorem 1.9 (Indifference Principle) *Let σ^* be an equilibrium in mixed strategies of a strategic-form game $\mathbf{G}$, and let s_i and s_i' be a couple of pure strategies of player i. Furthermore, assume that $\sigma_i^*(s_i) > 0$ and $\sigma_i^*(s_i') > 0$. Then,*

$$U_i(s_i, \sigma_{-i}^*) = U_i(s_i', \sigma_{-i}^*).$$

Remark 1.10 The set of Nash Equilibria in pure strategies is included in the set of Nash Equilibria in Mixed Strategies.

In Tables 4 and 5 we present the Nash Equilibria in mixed strategies for the couple of examples introduced before. In particular, for the Prisoner's Dilemma the only equilibrium in mixed strategies is provided by the Nash equilibrium itself, whereas, for the battle of sexes, there is a new Nash equilibrium in mixed strategies which is $\left(\frac{1}{3}\delta_B + \frac{2}{3}\delta_S\right)\left(\frac{2}{3}\delta_B + \frac{1}{3}\delta_S\right)$.

		Player II	
		D	C
Player I	D	$(1,1)$	$(5,0)$
	C	$(0,5)$	$(4,4)$

Table 4. Prisoner's Dilemma. No mixed Nash equilibrium apart from $\delta_{D,D}$.

		Male B	Male S
Female	B	$(2,1)$	$(0,0)$
	S	$(0,0)$	$(1,2)$

Table 5. The battle of sexes. One new mixed Nash equilibrium at $\left(\frac{1}{3}\delta_B + \frac{2}{3}\delta_S\right)\left(\frac{2}{3}\delta_B + \frac{1}{3}\delta_S\right)$.

1.3.4 Results

We end the introductory section on Game Theory presenting the two most famous results on the existence of Nash Equilibria.

Theorem 1.11 (Kuhn) *In every finite game with perfect information there exists at least one Nash equilibrium.*

Theorem 1.12 (Nash: 1950, 1951) *Let* $\mathbf{G}$ *be a game in strategic form with a finite number of players and in which every player has a finite number of pure strategies. Then, there is an equilibrium in mixed strategies.*

2 Correlated Equilibria

2.0.1 Aumann's Idea

Correlated Equilibria were introduced by Robert J. Aumann [3] and his idea can be roughly sketched as follows. First, we introduce a mediator, who suggests strategies to the players to play in the game. The suggestions given by the observer are chosen probabilistically, according to a probability distribution that is common knownledge among the players. Furthermore, the recommendations are private, that is each player only knows the recommendation provided to him. Finally, the mechanism is common knowledge among the players: each player knows that this is the mechanism used in the game, each player knows that the other players know that this is the mechanism used in the game, each player knows that the other players know that the other players know that this is the mechanism used in the game, and so forth.

2.0.2 Formal Construction

Let's try now to give a mathematical formulation to the idea we have sketched above.

Let $\mathbf{G} = (N, (S_i)_{i\in N}, (u_i)_{i\in N})$ be a strategic-form game. For every probability distribution $p \in \mathcal{P}(S)$, we define the following game $\Gamma^*(p)$:

- According to p, an outside observer (mediator) probabilistically selects an action vector $s \in S$.
- The mediator tells each player $i \in N$ s_i , but not s_{-i}. So, to each player i is revealed (or better recommended) his coordinate in the action vector that was selected.

- Each player i chooses an action $s_i' \in S_i$. This is not necessarily the one suggested by the mediator.
- Each player i has a payoff $u_i(s_i', \dots, s_n')$.

Definition 2.1 A *(pure) strategy* of player i in the game $\Gamma^*(p)$ is a map $\tau_i : S_i \to S_i$ that associates every recommendation s_i given by the mediator to an action $\tau_i(s_i) \in S_i$.

Any player i can follow the mediator's recommendation. For each player $i \in N$, define a strategy τ_i^* by

$$\tau_i^*(s_i) = s_i, \quad \text{for any } s_i \in S_i.$$

Theorem 2.2 *A strategy vector τ^* represents an equilibrium of the game $\Gamma^*(p)$ if and only if*

$$\sum_{s_{-i} \in S_{-i}} p(s_i, s_{-i}) u_i(s_i, s_{-i}) \geq \sum_{s_{-i} \in S_{-i}} p(s_i, s_{-i}) u_i(s_i', s_{-i}), \text{ for all } s_i, s_i' \in S_i.$$

Definition 2.3 We call a probability distribution $p \in \mathcal{P}(S)$ a *correlated equilibrium* in the game $\mathbf{G}$, if the strategy vector τ^* is a Nash equilibrium of the game $\Gamma^*(p)$. Equivalently,

$$\sum_{s_{-i} \in S_{-i}} p(s_i, s_{-i}) u_i(s_i, s_{-i}) \geq \sum_{s_{-i} \in S_{-i}} p(s_i, s_{-i}) u_i(s_i', s_{-i}), \text{ for all } s_i, s_i' \in S_i,$$

for every player $i \in N$.

It is straightforward to see that the set of correlated equilibria includes the set of Nash equilibria in mixed strategies, and consequently, the set of Nash equilibria in pure strategies.

Theorem 2.4 *The probability distribution $p_{\sigma^*} := \sigma_1^*(\cdot) \dots \sigma_n^*(\cdot)$ is a correlated equilibrium, for every Nash equilibrium σ^*.*

The following corollary follows from Theorem 2.4 together with Theorem 1.12.

Corollary 2.5 *There exists a correlated equilibrium in every finite strategic-form game.*

Furthermore, the set of correlated equilibria, as opposed to Nash equilibria in pure and mixed strategies, enjoys the following nice property.

Theorem 2.6 *In any finite game the set of correlated equilibria is convex and compact.*

Finally, we end the session discussing correlated equilibria for our ongoing examples. On one side, for the Prisoner's Dilemma, the only correlated equilibrium is the Nash equilibrium in pure strategies. On the other side, for the Battle of the Sexes, the set of correlated equilibria, $\{\alpha\delta_{(B,B)} + \beta\delta_{(S,B)} + \gamma\delta_{(B,S)} + \eta\delta_{(S,S)}\colon\ 2\alpha \geq \beta, \eta \geq 2\gamma, 2\eta \geq \beta$ and $\alpha \geq 2\gamma\}$, has infinite cardinality.

		Player II	
		D	C
Player I	D	$(1,1)$	$(5,0)$
	C	$(0,5)$	$(4,4)$

Table 6. Prisoner's Dilemma. The only correlated equilibrium is the Nash equilibrium in pure strategies.

		Male	
		B	S
Female	B	$(2,1)$	$(0,0)$
	S	$(0,0)$	$(1,2)$

Table 7. The battle of sexes. Correlated Equilibria $\alpha\delta_{(B,B)} + \beta\delta_{(S,B)} + \gamma\delta_{(B,S)} + \eta\delta_{(S,S)}$, with $2\alpha \geq \beta, \eta \geq 2\gamma, 2\eta \geq \beta$ and $\alpha \geq 2\gamma$.

3 Mean Field Games

3.0.1 Origins

Mean field games were introduced by simultaneously but independently by Huang, Malhamé and Caines [6] and Lasry and Lions [8]. They arise as limit systems, as the number of players, N, goes to infinity, for certain N-player games. In particular, the starting N-player games should be symmetric (in the sense that the components have to be statistically indistinguishable, or equivalently, the joint laws have to be exchangeable), with mean field interaction (the influence of each single player on the whole system diminishes as $N \to \infty$). The passage to the limit is completely analogous to the one for McKean-Vlasov limit of weakly interacting particle systems. The difference is in the fact that the systems here are controlled and, in particular, the notion of optimality at prelimit level is the one of (approximate) Nash equilibria.

3.0.2 A One-Period Deterministic Game

A simple but prototypical example of MFG is presented now. It is meant to provide the idea of the kind of problems treated and of the technique exploited without the technical details that are usually needed for this kind of problems. In the form presented there it is taken from the book by Carmona and Delarue [1] but this example first appeared in Lyons' lectures.

We aim at getting a well-motivated answer to the following question: *"Where do I put my towel on the beach?"* The mathematical setting is the following. Consider $i \in [\![1, N]\!]$, a population of N individuals, with N large ($N \to \infty$). Let $\alpha_i \in A_i$ ($A_i = (A, d)$ compact metric space) be a point chosen by player $i \in [\![1, N]\!]$ and corresponding to the place on the beach where the i-th player has decided to place his towel. Fix $\alpha, \beta \in \mathbb{R}_+$ and α_0, a special point of interest (for example α_0 can be the location of the unique bar on the beach). Each

player wants to minimize the quantity

$$J_i(\alpha_1, \dots, \alpha_N) := \gamma d(\alpha_i, \alpha_0) - \frac{\beta}{N-1} \sum_{j \neq i, j=1}^{N} d(\alpha_i, \alpha_j).$$

The objective functional is, hence, given by the sum of two components to which we can give the following interpretation. The first part, which has a positive contribution, is a penalization term to force the system to be as close as possible to the point of interest. The second term is introduced since individuals want to avoid crowds and so to be as far as possible from the other individuals in mean. Notice that each of the functionals J_i can be rewritten as

$$J_i(\alpha_1, \dots, \alpha_N) = \widetilde{J}(\alpha_i, \frac{1}{N-1} \sum_{j \neq i, j=1}^{N} \delta_{\alpha_j}),$$

for $\widetilde{J}(\alpha, m) := \gamma d(\alpha, \alpha_0) - \beta \int_A d(\alpha, \lambda) m(d\lambda)$. In particular, this game possesses all the nice properties that we were asking for in the previous section and, consequently, we can think of approximating its solution by solving a limit problem linked to the functional $\widetilde{J}$. Roughly speaking, in the limit $N \to \infty$, we proceed mimicking the construction of Nash equilibria for N- player games. First, we fix a probability distribution μ. Then, we solve the minimization problem

$$\inf_{\alpha \in A} \widetilde{J}(\alpha, \mu).$$

Finally, as a fixed point argument, we look for a measure $\widehat{\mu}$ concentrated on the arguments of the minimization.

A *MFG problem* consists in the three-step problem above.

4 Correlated Solutions for MFGs

In the final section we wrap things up together to formulate a suitable definition of *correlated solution* in the Mean Field context.

4.0.1 Set Up

We consider the following setting for the game. For a fixed T, finite time horizon, let $[\![0, T]\!] := \{0, 1, \dots, T-1, T\}$ be the discrete time steps. The space of individual states, $\mathcal{X}$, and of individual control action, Γ, are finite. The space of idiosyncratic noise, $\mathcal{Z} = [0, 1]$, is equipped with $\nu = \text{Uniform}([0, 1])$. The measurable system function that determines the dynamics of the states is given by $\Psi : [\![0, T-1]\!] \times \mathcal{X} \times \mathcal{P}(\mathcal{X}) \times \Gamma \times \mathcal{Z} \to \mathcal{X}$. Finally, $\mathcal{R} := \{\varphi : [\![0, T-1]\!] \times \mathcal{X} \to \Gamma, \text{measurable}\}$ is set of admissible individual strategies.

Remark 4.1 Notice that the strategies only depend on time and players' own positions, restricted strategies (also referred to as decentralized Markov strategies).

4.0.2 The dynamics in the N-player Game

Let $\mathfrak{m}^N \in \mathcal{P}(\mathcal{X}^N)$ be an initial distribution. We call a $\gamma^N \in \mathcal{P}(\mathcal{R}^N)$ *correlated profile* and a mapping $u : \mathcal{R} \to \mathcal{R}$ *strategy modification.*
The tuple $((\Omega_N, \mathcal{F}_N, \mathbb{P}_N), (\Phi_j^N)_{j=1}^N, (X_0^{j,N}, \dots, X_T^{j,N})_{j=1}^N, (\xi_1^{j,N}, \dots, \xi_T^{j,N})_{j=1}^N)$ is a *realization* of the triple $(\mathfrak{m}^N, \gamma^N, u)$ for player i if:

i) $\mathbb{P}_N \circ (X_0^{1,N}, \dots, X_0^{N,N})^{-1} = \mathfrak{m}^N$;

ii) $\mathbb{P}_N \circ (\Phi_1^N, \dots, \Phi_N^N)^{-1} = \gamma^N$;

iii) $\xi_t^{j,N}$, $j \in [\![1, N]\!]$, $t \in [\![1, T]\!]$ are i.i.d. according to $\nu(\cdot)$;

iv) $(\xi_t^{j,N})_{j\in[\![1,N]\!],t\in[\![1,T]\!]}$, $(X_0^{j,N})_{j\in[\![1,N]\!]}$ e $(\Phi_j^N)_{j\in[\![1,N]\!]}$ are independent;

v) For any $t \in [\![0, T-1]\!]$, $\mathbb{P}_N$-a.s.:

$$
\begin{aligned}
X_{t+1}^{i,N} &= \Psi\left(t, X_t^{i,N}, \mu_t^{i,N}, u \circ \Phi_i^N(t, X_t^{i,N}), \xi_{t+1}^{i,N}\right), \\
X_{t+1}^{j,N} &= \Psi\left(t, X_t^{j,N}, \mu_t^{j,N}, \Phi_j^N(t, X_t^{j,N}), \xi_{t+1}^{j,N}\right), \quad j \neq i.
\end{aligned} \tag{1}
$$

where $\mu_t^{l,N} := \frac{1}{N-1}\sum_{j=1, j\neq l}^N \delta_{X_t^{j,N}}$, for all $l \in [\![1, N]\!]$, $t \in [\![0, T]\!]$.

4.0.3 The costs in the N-player Game

Player i faces costs associated with $(\mathfrak{m}^N, \gamma^N, u)$ that are given by

$$
J_i^N(\mathfrak{m}^N, \gamma^N, u) := \mathbb{E}\left[\sum_{t=0}^{T-1} f\left(t, X_t^{i,N}, \mu_t^{i,N}, u \circ \Phi_i^N\left(t, X_t^{i,N}\right)\right) + F\left(X_T^{i,N}, \mu_T^{i,N}\right)\right].
$$

where f represents the running costs and F the terminal costs.

Remark 4.2 We can give the following interpretation to the dynamics and costs in the N-player game. In the definition of costs for player i, he applies modified strategy $u \circ \Phi_i^N$ (instead of Φ_i^N), while all the other players apply the strategies recommended by the mediator, that is Φ_j^N. When $u = \mathrm{Id}$ player i accepts the mediator's suggestion .

4.0.4 Correlated Equilibria in N-player game

Definition 4.3 Let $\varepsilon \geq 0$. We name a distribution $\gamma^N \in \mathcal{P}(\mathcal{R}^N)$ an *ε-correlated equilibrium* with initial distribution $\mathfrak{m}^N \in \mathcal{P}(\mathcal{X}^N)$ if, for any $i \in [\![1, N]\!]$ and any strategy modification $u : \mathcal{R} \to \mathcal{R}$, we have

$$
J_i^N(\mathfrak{m}^N, \gamma^N, \mathrm{Id}) \leq J_i^N(\mathfrak{m}^N, \gamma^N, u) + \varepsilon.
$$

In particular, we call γ^N an *correlated equilibrium*, simply denoted by CE, if $\varepsilon = 0$.

Remark 4.4 In this setting, Correlated equilibria are defined with respect to restricted strategies but analogous definitions can be formulated for full feedback strategies. When γ^N is a Dirac distribution ($\Phi_1^N, \dots, \Phi_N^N$ constant), then the definition above reduces to that of a Nash equilibrium in pure strategies. When γ^N has product form ($\Phi_1^N, \dots, \Phi_N^N$ independent), then the definition above corresponds to Nash equilibria in randomized strategies.

Proposition 4.5 (Proposition 3.1, in [4]) *Let the distribution $\mathfrak{m}^N$ be exchangeable. Then, there exists a symmetric correlated equilibrium with initial distribution $\mathfrak{m}^N$.*

4.0.5 The dynamics in the Mean Field Game

Let $\mathfrak{m}_0 \in \mathcal{P}(\mathcal{X})$ be an initial distribution. We call $\rho \in \mathcal{P}(\mathcal{R} \times \mathcal{P}(\mathcal{X})^{T+1})$ *correlated solution* and a mapping $u : \mathcal{R} \to \mathcal{R}$ *strategy modification.*
The tuple $((\Omega, \mathcal{F}, \mathbb{P}), \Phi, (X_t)_{t\in[\![0,T]\!]}, (\xi_t)_{t\in[\![1,T]\!]})$ is a realization of the triple $(\mathfrak{m}_0, \rho, u)$ if:

i) $\mathbb{P} \circ X_0^{-1} = \mathfrak{m}_0$;

ii) $\mathbb{P} \circ (\Phi, \mu)^{-1} = \rho$;

iii) ξ_t, $t \in [\![1, T]\!]$ are i.i.d. according to ν;

iv) $(\xi_t)_{t\in[\![1,T]\!]}$, X_0 e (Φ, μ) are independent;

v) For any $t \in [\![0, T-1]\!]$, $\mathbb{P}$-a.s.:

$$X_{t+1} = \Psi\left(t, X_t, \mu_t, u \circ \Phi(t, X_t), \xi_{t+1}\right). \tag{2}$$

4.0.6 The costs in the Mean Field Game

The representative player's costs associated with $(\mathfrak{m}_0, \rho, u)$ that are given by

$$J(\mathfrak{m}_0, \rho, u) := \mathbb{E}\left[\sum_{t=0}^{T-1} f\left(t, X_t, \mu_t, u \circ \Phi\left(t, X_t\right)\right) + F\left(X_T, \mu_T\right)\right].$$

where, as in the N-plyer game, f represents the running costs and F the terminal costs.

Remark 4.6 The independence properties and the iterative structure guarantee that any two realizations of $(\mathfrak{m}_0, \rho, u)$ share the same expected value in the definition of $J(\mathfrak{m}_0, \rho, u)$ and, consequently, the cost functional is well posed.
When $u = \mathrm{Id}$ the representative player accepts the mediator's suggestion .

4.0.7 Correlated Solutions in the Mean Field Game

Definition 4.7 We name a distribution $\rho \in \mathcal{P}(\mathcal{R} \times \mathcal{P}(\mathcal{X})^{T+1})$ a *correlated solution* with initial distribution $\mathfrak{m}_0 \in \mathcal{P}(\mathcal{X})$ if the following two conditions hold:

- For any strategy modification $u : \mathcal{R} \to \mathcal{R}$, we have:

$$J(\mathfrak{m}_0, \rho, \mathrm{Id}) \leq J(\mathfrak{m}_0, \rho, u).$$

- For any realization $((\Omega, \mathcal{F}, \mathbb{P}), \Phi, (X_t)_{t\in[\![0,T]\!]}, (\xi_t)_{t\in[\![1,T]\!]})$ of the triple $(\mathfrak{m}_0, \rho, \mathrm{Id})$, it holds

$$\mu_t(\cdot) = \mathbb{P}(X_t \in \cdot | \mathcal{F}_T^{\mu}).$$

Remark 4.8

- In the definition above the first condition is called ***Optimality***, while the second is called ***Consistency***.
- The definition above is again with respect to restricted strategies.
- Consistency condition implies $\mu_t(\cdot) = \mathbb{P}(X_t \in \cdot | \mathcal{F}_t^{\mu})$.

4.0.8 Assumption

We introduce the following set of assumptions which are needed to guarantee the validity of the following results.

Assumption 4.9

(A1) Continuity of $\Psi : [\![0, T-1]\!] \times \mathcal{X} \times \Gamma \times \mathcal{Z} \to \mathcal{X}$:

1) For every $(t, x, \gamma) \in [\![0, T-1]\!] \times \mathcal{X} \times \Gamma$ and for all $m, \widetilde{m} \in \mathcal{P}(\mathcal{X})$,

$$\int_{\mathcal{Z}} \mathbb{I}_{\Psi(t,x,m,\gamma,z)\neq\Psi(t,x,\widetilde{m},\gamma,z)} \nu(dz) \leq w(\mathrm{dist}(m, \widetilde{m})),$$

where $w : [0, +\infty) \to [0, 1]$ is a measurable function with $\lim_{s\to 0^+} w(s) = 0$.

2) For any $t \in [\![0, T-1]\!]$, $\Psi(t, \cdot)$ is $\tau \otimes \nu$-almost everywhere continuous, for every $\tau \in \mathcal{P}(\mathcal{X} \times \mathcal{P}(\mathcal{X}) \times \Gamma)$.

(A2) The functions f and F, running cost and terminal cost, are continuous.

4.1 Justification of the Definition

The definition introduced above is justified in two ways. On one side, it can be justified by showing convergence of N-player correlated equilibria to the mean field game limit.

Theorem 4.10 (Theorem 6.1 in [4]) *Let* $((\Omega_N, \mathcal{F}_N, \mathbb{P}_N), (\Phi_j^N)_{j=1}^N, (X_0^{j,N}, \dots, X_T^{j,N})_{j=1}^N,$ $(\xi_1^{j,N}, \dots, \xi_T^{j,N})_{j=1}^N)$ *be a realization of the triple* $(\mathfrak{m}^N, \gamma^N, Id)$*, for any* $N \in \mathbb{N} \setminus \{1\}$*, where*

γ_N is a symmetric ε_N-correlated equilibrium in restricted strategies with initial distribution $\mathfrak{m}^N = \mathfrak{m}_{0,N}^{\otimes N}$, such that $\mathfrak{m}_{0,N} \overset{N\to\infty}{\longrightarrow} \mathfrak{m}_0$. Assume that $\varepsilon_N \to 0$, as N goes to infinity. Set

$$\rho^N := \mathbb{P}_N \circ (\Phi_1^N, \mu^{1,N})^{-1},$$

where $\mu_t^{1,N}$, $t \in [\![0,T]\!]$ is the empirical measure flow.
Then, under the set of assumptions in 4.9, the sequence $(\rho^N)_{N\in\mathbb{N}}$ is relatively compact as a subset of $\mathcal{P}(\mathcal{R} \times \mathcal{P}(\mathcal{X})^{T+1})$ and any limit point of the sequence is a correlated solution of the mean field game having initial distribution $\mathfrak{m}_0$.

On the other side, we can construct approximate correlated equilibria starting from a solution of the mean field game.

Theorem 4.11 (Theorem 7.1 in [4]) *Let $\mathfrak{m}_0 \in \mathcal{P}(\mathcal{X})$ and let $\mathfrak{m}^N = \mathfrak{m}_{0,N}^{\otimes N} \in \mathcal{P}(\mathcal{X}^{T+1})$, with $\mathfrak{m}_{0,N} \overset{N\to\infty}{\longrightarrow} \mathfrak{m}_0$, in $\mathcal{P}(\mathcal{X})$. Assume that $\rho \in \mathcal{P}(\mathcal{R} \times \mathcal{P}(\mathcal{X})^{T+1})$ is a correlated solution of the mean field game with initial distribution $\mathfrak{m}_0$. For any natural number $N \in \mathbb{N} \setminus \{1\}$, define $\gamma^N \in \mathcal{P}(\mathcal{R}^N)$ through*

$$\gamma^N(C_1 \times \cdots \times C_N) = \int_{\mathcal{P}(\mathcal{X})^{T+1}} \prod_{j=1}^{N} \rho_1(C_j|m)\rho_2(dm).$$

Then, under technical assumptions, there exists $(\varepsilon_N)_{N\in\mathbb{N}}$ such that γ^N is an ε_N-correlated equilibrium with initial distribution $\mathfrak{m}^N$, and $\lim_{N\to\infty} \varepsilon_N = 0$.

References

[1] Carmona, R. and Delarue, F., "Probabilistic Theory of Mean Field Games with Applications". Probab. Theory Stoch. Model., 83, Springer, 2018.

[2] Maschler, M.; Solan, E. and Zamir, S., "Game Theory". Cambridge University Press, Cambridge, 2013.

[3] Aumann, R.J., *Subjectivity and correlation in randomized strategies.* J. Math. Econom., 1, pp. 67–96, 1974.

[4] Campi, L. and Fischer, M., *Correlated equilibria and mean field games: a simple model.* Math. Oper. Res., Published online at `https://doi.org/10.1287/moor.2021.1206`.

[5] Hart, S. and Schmeidler, D., *Existence of correlated equilibria.* Math. Oper. Res., 14(1), pp. 18–25, 1989.

[6] Huang, M.; Malhamé, R.P. and Caines, P.E., *Large population stochastic dynamic games: Closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle.* Commun. Inf. Syst., 6(3), pp. 221–252, 2006.

[7] Lacker, D., *On the convergence of closed-loop Nash equilibria to the mean field game limit.* Ann. Appl. Probab., 2018.

[8] Lasry, J.-M. and Lions, P.-L., *Mean field games.* Japan. J. Math., 2(1), pp. 229–260, 2007.

Optimal control problems: existence of minimizers, necessary conditions, and gap phenomena

Giovanni Fusco (*)

Abstract. By *optimal control problem* we mean the minimization of a functional over arcs that satisfy certain constraints (dynamics, control, endpoint and state constraints). After a brief introduction on the subject, we will discuss the notion of closure of trajectories associated with a controlled differential equation, so that to present an existence theorem for optimal control problems. Then, we will announce the Pontryagin's Maximum Principle, that is, the most known set of necessary conditions that has to be fulfilled by a minimizer. Afterwards, we will introduce the most common extensions for the optimal control problems which do not admit minimizers and we will analyze the properness of such extensions. In particular, we will deal with the issue of *gap phenomena* between an optimal control problem and an its extension and we will prove a link between this occurrence and a topological property of the trajectories which is usually called *isolation*. Finally, we will establish that isolated trajectories satisfy the Maximum Principle in abnormal form, i.e. there exists at least a set of multipliers with cost multiplier equal to zero. We will conclude with some examples that illustrate the outcomes.

Mathematics Subject Classification (2020): 34H, 49N, 49K

Keywords: Optimal control problems, Control systems, Maximum Principle, Gap phenomena, Normality

1 Introduction

Since their introduction in the mathematical analysis theory, differential equations have been largely employed in order to model physical phenomena whose rate of change is known and depends on the current state itself, as for instance

$$\dot{y}(t) = f(t, y(t)), \qquad y(0) = x_0.$$

Notice that if we know the initial position of the system we are able to solve the above Cauchy problem, so that to forecast the future evolution. However, even if we can under-

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: fusco@math.unipd.it . Seminar held on 2 March 2022.

stand a process and its progression in time, we can not affect its behavior in any way: we are here taking a spectator point of view without the possibility of taking action.

In control theory the perspective is reversed: we now assume the presence of an external agent (i.e. a controller) who can play an action on the system, altering it. This new situation is designed by a control system, that is

$$\dot{y}(t) = f(t, y(t), \alpha(t)), \qquad y(0) = x_0, \qquad \alpha(t) \in A.$$

In this case, the dynamics of the system relies not only on the current state itself, but also on the control $\alpha(\cdot)$, which represents the active external influence selected by the controller in order to adjust the behavior of the system and reach certain preassigned goals: steer the system from an initial point to a fixed target, maximize a profit, minimize a cost. We will discuss topics about control systems in Section 2, which is mainly bases on [2]. In particular, we present an existence theorem for control systems and then we establish sufficient conditions to ensure the closure of the set of trajectories.

A very important area of control theory is concerned with optimal control. In many applications, among all strategies which accomplish a certain task, one seeks an optimal one, based on a given performance criterion. Optimal control emerged as a distinct field of research in the 1950's, to address optimization problems arising in aerospace engineering. Now, fifty years on, there are ample applications in new areas ranging from process control, resource economics, ecology, to robotics and epidemiology. Equally significant is the stimulus optimal control has given to research in related branches of mathematics such as convex analysis, nonlinear analysis, functional analysis and dynamical systems.

A classic example of (fixed end-time) optimal control problem is given by

$$\begin{cases} \text{Min} \quad \int_0^T L(t, y(t), \alpha(t))dt \\ \qquad \text{over processes } (y, \alpha) \text{ satisfying} \\ \dot{y}(t) = f(t, y(t), \alpha(t)); \quad \alpha(t) \in A; \\ y(0) = x_0; \quad y(T) \in \mathcal{T}. \end{cases}$$

Here, among all controls which steer the system from the initial point to some point on the target set, we may seek the one that minimizes the lagrangian functional. We point out that if $f(t, x, a) \equiv a$, $A = \mathbb{R}^m$ and $\mathcal{T} = \{x_{_T}\}$, we have a classic calculus of variations problem: optimal control can be seen as a generalization of calculus of variations. Roughly speaking, the main difference between a calculus of variations problem and an optimal control problem is that in the former the derivative is unrestricted, while in latter it is constrained.

The basic theory of optimal control has been concerned with three main issues: existence of optimal controls, necessary conditions for the optimality of a control and sufficient conditions for optimality. The first two points are deeply investigated in Section 3, while on the contrary the last one, which is mostly related to dynamic programming and Hamilton Jacobi Belmann equations, is not analyzed in this article.

The basic aim of any set of necessary conditions is to select possible candidates for the minimum. The major result in this direction is the celebrated Pontryagin's Maximum

Principle, which extends to control systems the Euler-Lagrange and the Weierstrass necessary conditions for a strong local minimum in the calculus of variations. Nevertheless, as we will see, many optimal control problems do not admit minimizers, and the Maximum Principle becomes useless. In this situations it is a common practice to embed our original optimization problem into an auxiliary optimal control problem which has minimizers. Then, once found a minimizer for the auxiliary problem, thanks to numerical approximation techniques it is possible to find ε-minimizers for the original problem, i.e. feasible processes whose cost exceeds the infimum cost of the original problem at most of a quantity ε. Of course, a fundamental requirement for a good extension is that there is no gap between the infimum of the original problem and that of the auxiliary problem, and the goal of Section 4 is to provide sufficient conditions in order to avoid gap phenomena. The most studied auxiliary problems are the extended (see [9, 5, 10]) and the relaxed problem (see [11, 14, 10]) and are usually studied separately. Basing our exposition on [6, 7, 8], we will present them in an original unified framework and then we will illustrate our results with some examples.

1.1 Notations and preliminaries

Given an interval $I \subseteq \mathbb{R}$ and a set $X \subseteq \mathbb{R}^k$, we write $W^{1,1}(I,X)$, $C(I,X)$, $\mathcal{M}(I,X)$, $L^1(I,X)$ for the space of absolutely continuous functions, continuous functions, measurable functions, and Lebesgue integrable functions defined on I and with values in X, respectively. For all the classes of functions introduced so far, we will not specify domain and codomain when the meaning is clear. Given $y \in W^{1,1}(I,X)$, its (weak) derivative belongs to $L^1(I,X)$ and it is denoted by $\dot{y}$. We will use $\|\cdot\|_{L^\infty(I)}$ to denote the ess-sup norm on I. When the domain is clear, we will sometimes simply write $\|\cdot\|_{L^\infty}$. Furthermore, we denote by $\operatorname{co} X$, $\overline{X}$ the convex hull and the closure of X, respectively. In particular the convex hull of X is defined to be the smallest convex set that contains X and, by the Caratheodory's theorem [1,Prop. 0.5.5], it can be represented by

$$\operatorname{co} X = \left\{ \sum_{j=0}^{k} \lambda_j x_j \,:\, x_j \in X \ \forall j,\ (\lambda_0, \dots, \lambda_k) \in \Delta_k \right\}, \tag{1.1}$$

where $\Delta_k := \{(\lambda_0, \dots, \lambda_k) \in \mathbb{R}^{k+1} : \lambda_i \geq 0 \ \forall i,\ \sum_{i=0}^{k} \lambda_i = 1\}$ is the symplex in $\mathbb{R}^{k+1}$. As customary, $I \cdot X$ denotes the set $\{r\,x \,|\, r \in I,\, x \in X\}$. Given two nonempty subsets X_1, X_2 of $\mathbb{R}^k$, we denote by $X_1 + X_2$ the set $\{x_1 + x_2 \,|\, x_1 \in X_1,\, x_2 \in X_2\}$. We set $\mathbb{R}_{\geq 0} := [0, +\infty[$. We denote by $NBV^+([0,S];\mathbb{R})$ the space of increasing, real valued functions μ on $[0,S]$ of bounded variation, vanishing at the point 0 and right continuous on $]0,S[$. Each $\mu \in NBV^+([0,S];\mathbb{R})$ defines a Borel measure on $[0,S]$, still denoted by μ, its total variation function is indicated by $\|\mu\|_{TV}$ or equivalently by $\mu([0,S])$, and its support by $\operatorname{spt}\{\mu\}$. Let $G : \mathbb{R}^k \to \mathbb{R}^l$ be a continuously differentiable function, we denote by ∇G the Jacobian matrix of G and when $l = 1$, with a small abuse of notation, we still denote by ∇G the usual gradient operator. If $G : \mathbb{R}^{k_1} \times \mathbb{R}^{k_2} \to \mathbb{R}^l$ and $x = (x_1, x_2) \in \mathbb{R}^{k_1} \times \mathbb{R}^{k_2}$, we use $\nabla_{x_i} G$ to denote partial Jacobian matrix (partial gradient operator when $l = 1$) of G with respect to x_i, for $i = 1, 2$.

2 Control systems

A control system is a dynamical system governed by a controlled differential equation of the form

$$\text{(CS)} \qquad \begin{cases} \dot{y}(t) = f(t, y(t), \alpha(t)) & \text{a.e. } t \in [0,T] \\ y(0) = x_0 \\ \alpha(t) \in \mathcal{M}([0,T], A), \end{cases}$$

where $f : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is the *dynamics function*, $A \subset \mathbb{R}^m$ is the *control set*, $y(\cdot)$ is an absolutely continuous *trajectory* associated with the measurable *control* $\alpha(\cdot)$ and $x_0 \in \mathbb{R}^n$ is the initial position. Notice that $\alpha(\cdot)$ is the action that an agent can play on the system since it affects the dynamics and, subsequently the evolution of the system: it is for this reason that the function $\alpha(\cdot)$ is called control. A pair (y, α) that satisfies the constraints in (CS) is called *process*. In this section we assume the following hypotheses:

(H1) The control set $A \subset \mathbb{R}^m$ is compact, the dynamics function $f(t, x, a)$ is continuous in all variables and continuously differentiable with respect to x. Moreover, there exists $C > 0$ such that for all $(t, x, a) \in \mathbb{R} \times \mathbb{R}^n \times A$ it holds

$$|f(t, x, a)| \le C, \qquad |\nabla_x f(t, x, a)| \le C.$$

Noticing that, given a measurable control $\alpha(\cdot)$, we can define the function $(t, x) \mapsto g(t, x) := f(t, x, \alpha(t))$, we can evoke the existence theorems for ordinary differential equations (see [2, Sec. 2.1]) and deduce the following existence result.

Theorem 2.1 ([2], Thm. 3.2) *Assume* **(H1)**. *Then, for any $T > 0$ and any control $\bar{\alpha} \in \mathcal{M}([0,T], A)$, there exists a unique absolutely continuous solution $y(\cdot, \bar{\alpha})$ to*

$$\dot{y}(t) = f(t, y(t), \bar{\alpha}(t)) \quad \text{a.e. } t \in [0,T], \qquad y(0) = x_0.$$

Moreover, if $(\alpha_n) \subset \mathcal{M}([0,T], A)$ converges to $\bar{\alpha} \in \mathcal{M}([0,T], A)$ in $L^1(0,T)$, then the sequence of associated trajectories $(y(\cdot, \alpha_n))$ converges to $y(\cdot, \bar{\alpha})$ in $C^0(0,T)$.

Remark 2.2 The regularity assumptions on f in **(H1)** could be considerably weakened. In particular, it suffices that (1): the function $t \mapsto f(t, x, a)$ is measurable for any $(x, a) \in \mathbb{R}^n \times A$, (2): the function $(x, a) \mapsto f(t, x, a)$ is continuous for every $t \in \mathbb{R}$ and (3): there exists $\psi \in L^1(\mathbb{R}, \mathbb{R}_{\ge 0})$ such that

(i) $|f(t, x, a)| \le \psi(t)$ for any $(t, x, a) \in \mathbb{R} \times \mathbb{R}^n \times A$,
(ii) $|f(t, x, a) - f(t, x', a)| \le \psi(t)|x - x'|$ for any $(t, x, a), (t, x', a) \in \mathbb{R} \times \mathbb{R}^n \times A$.

For more details see [1, 3, 12], even if the results are expressed for the equivalent differential inclusion formulation. For the differential inclusion formulation see also [2, Sec. 3.1], which is simpler to read. In addition, in order to prove the only existence and uniqueness we can further remove (3)(i), and if we remove point (3) existence still holds without uniqueness. Furthermore we could slightly weaken condition (3) for local existence results.

Now we introduce the concept of closure of the set of trajectories with an example.

Example 2.3 Consider the following scalar control system

$$\dot{y}(t) = \alpha(t), \quad y(0) = 0, \quad \alpha(t) \in \{-1, 1\},$$

and consider the sequence (α_n) of highly oscillatory controls defined by

$$\alpha_n(t) := \begin{cases} 1 & \text{if } \sin(nt) \geq 0 \\ -1 & \text{otherwise.} \end{cases}$$

It is easy to see that the sequence $(y(\cdot, \alpha_n))$ of the trajectories associated with (α_n) converges uniformly to 0, which is not a trajectory of the considered control system. In fact, there are no controls whose associated trajectory is 0, because the control should be constantly equal to 0. The problem here, that might cause confusion with respect to Theorem 2.1, is that the sequence of control is not convergent in L^1 to an admissible control.

We want to find sufficient conditions for the set of trajectories to be closed, in order to avoid the situation in Example 2.3.

Definition 2.4 We say that the set of trajectories of control system (CS) is **close** if for any given sequence of processes (y_n, α_n) such that $y_n \to \bar{y}$ in $C^0(0, T)$ for some contiuous arc $\bar{y}$, then there exists an admissible control $\bar{\alpha} \in \mathcal{M}([0, T], A)$ such that $\bar{y}(\cdot) = y(\cdot, \bar{\alpha})$.

To this aim we define the *set of velocities* $F(t, x)$ as follows

$$F(t, x) := \{f(t, x, a) : a \in A\} \qquad \forall (t, x) \in \mathbb{R} \times \mathbb{R}^n.$$

Theorem 2.5 *rm ([2], Thm. 3.5) Assume* **(H1)** *and that the set of velocities $F(t, x)$ is convex for all $(t, x) \in \mathbb{R} \times \mathbb{R}^n$. Then the set of trajectories is closed.*

Remark 2.6 For instance, $F(t, x)$ is convex if $f(t, x, a) = M(t, x) + N(t, x)a$ and A is convex.

The importance of Theorem 2.5 relies on the fact that it is quite common in analysis to use subsequence extraction techniques and limit procedures, as for instance the Ascoli Arzelá's theorem, and as we will see later in Theorem 3.2. Since the main assumption of the Theorem 2.5 is the convexity of the set of velocities, it is natural to ask which is the relation a control system with non convex set of velocities and its relaxation, namely, the control system that we obtain by replacing the original set of velocities with its convex hull.

Given $T > 0$, we call *relaxed process* any $(y, (\alpha_0, \dots, \alpha_n), (\lambda_0, \dots, \lambda_n))$ that satisfies

$$\text{(RCS)} \qquad \begin{cases} \dot{y}(t) = \sum_{j=0}^{n} \lambda_j(t) f(t, y(t), \alpha_j(t)) \ \text{ a.e. } t \in [0, T] \\ (\alpha_0, \dots, \alpha_n), (\lambda_0, \dots, \lambda_n) \in \mathcal{M}([0, T], A^{n+1}) \times \mathcal{M}([0, T], \Delta_n) \\ y(0) = x_0. \end{cases}$$

Roughly speaking, control system (RCS) is obtained from (CS) by replacing the dynamics function with its convexification. In fact, in view of (1.1) one has

$$\operatorname{co} F(t,x) = \left\{ \sum_{j=0}^{n} \lambda_j f(t,x,a_j) \,:\, a_j \in A,\ (\lambda_0,\dots,\lambda_n) \in \Delta_n \right\}.$$

Remark 2.7 We notice that a process (y,α) for (CS) can be identified with the process $\big(y,(\alpha,\dots,\alpha),\big(\frac{1}{n+1},\dots,\frac{1}{n+1}\big)\big)$ for (RCS), hence there is an embedding between the two sets of processes.

We now state the well known relaxation theorem.

Theorem 2.8 ([2], Thm. 3.7) *Assume* **(H1)**. *Then the set of trajectories for (CS) is dense in the set of trajectories for (RCS) in the C^0-norm. That is, given $\varepsilon > 0$ and a process $(\bar y,(\bar\alpha_0,\dots,\bar\alpha_n),(\bar\lambda_0,\dots,\bar\lambda_n))$ for (RCS) defined on the interval $[0,T]$, then there exists a process $(y_\varepsilon,\alpha_\varepsilon)$ for (CS) defined on $[0,T]$ such that $\|\bar y - y_\varepsilon\|_{L^\infty(0,T)} \le \varepsilon$.*

Remark 2.9 We point out that, for Theorem 2.5 and Theorem 2.8 it suffices that $f(t,x,a)$ is measurable in t, measurably Lipschitz in x (i.e. it satisfies condition (3)(ii) in Remark 2.2 and continuous in a, see [12, Ch. 2]OptV, but also [1, 3].

3 Optimal control problems

A general (fixed end-time) optimal control problem is a problem of minimization of a certain functional over arcs that satisfy some constraints, as the following

$$\text{(P)}\quad \begin{cases} \text{Min}\ \ \Phi(y(T)) + \int_0^T L(t,y(t),\alpha(t))dt \\ \qquad \text{over } (y,\alpha) \in W^{1,1}([0,T],\mathbb{R}^n) \times \mathcal{M}([0,T],\mathbb{R}^m) \text{ satisfying} \\ \dot y(t) = f(t,y(t),\alpha(t)) \quad \text{a.e. } t \in [0,T] \\ \alpha(t) \in A \quad \text{a.e. } t \in [0,T] \\ (y(0),y(T)) \in \{x_0\} \times \mathcal{T} \\ h(t,y(t)) \le 0 \quad \text{for any } t \in [0,T]. \end{cases}$$

We refer to $\Phi(\cdot)$ as the *cost function*, to $\dot y = f(t,y,\alpha)$ as the *dynamics constraint*, to $\alpha \in A$ as the *control constraint*, to $(y(0),y(T)) \in \{x_0\} \times \mathcal{T}$ as the *endpoint constraint* and to $h(t,y) \le 0$ as the *state constraint*. A pair (y,α) is called *process* if it satisfies the dynamics and the control constraint. We refer to y as trajectory and to α as control. We call a process (y,α) *feasible* if y satisfies the endpoint and the state constraint. We say that a feasible process $(\bar y,\bar\alpha)$ is a *minimizer* for (P) if

$$\Phi(\bar y(T)) + \int_0^T L(t,\bar y(t),\bar\alpha(t))dt \le \Phi(y(T)) + \int_0^T L(t,y(t),\alpha(t))dt \qquad \text{for all } (y,\alpha) \text{ feasible.}$$

Remark 3.1 We notice that we can assume without loss of generality that $L \equiv 0$. Otherwise we could consider an auxiliary optimal control problem with cost function given by $\tilde{\Phi}(y(T), y_{n+1}(T)) = \Phi(y(T)) + y_{n+1}(T)$, where $y_{n+1} : [0,T] \to \mathbb{R}$ is a new variable such that $\dot{y}_{n+1}(t) = L(t, y(t), \alpha(t))$ and $y_{n+1}(0) = 0$.

In addition to **(H1)**, we make the following assumptions on the data.

(H2) We assume that $\mathcal{T} := \{x \in \mathbb{R}^n : \eta(x) \leq 0\}$ for some continuously differentiable function $\eta : \mathbb{R}^n \to \mathbb{R}$, $L \equiv 0$, $\Phi : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, $h : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ is continuous, continuously differentiable with respect to the x variable, and such that $|\nabla_x h(t,x)| \leq L$ for some $L > 0$.

We are now ready for an existence result, of which we give a proof inspired by that presented in [2, Thm. 5.1].

Theorem 3.2 *Assume* **(H1)-(H2)** *and that the set of the velocities* $F(t,x) = \{f(t,x,a) : a \in A\}$ *is convex for all* $(t,x) \in [0,T] \times \mathbb{R}^n$. *Then, if there exists at least one feasible process, the optimal control problem* (P) *admits a minimizer* $(\bar{y}, \bar{\alpha})$.

Proof. We take a minimizing sequence (y_n, α_n), that exists since there is at least a feasible process. We notice that

$$|\dot{y}_n(t)| \leq |f(t, y_n(t), \alpha_n(t))| \leq C \quad \text{and} \quad |y_n(t)| \leq |x_0| + \int_0^T |\dot{y}_n(t)| dt \leq |x_0| + CT.$$

By Ascoli-Arzelá's theorem we deduce that (y_n) has a convergent subsequence in $C^0(0,T)$ (we do not relabel) to some $\bar{y}$. By Theorem 2.5 there exists $\bar{\alpha}$ such that $\bar{y}(\cdot) = y(\cdot, \bar{\alpha})$. Clearly $\bar{y}(0) = x_0$ since $y_n(0) \equiv x_0$ and $\lim_n y_n(T) = \bar{y}(T) \in \mathcal{T}$ since $(y_n(T)) \subset \mathcal{T}$ and $\mathcal{T}$ is closed, because it is the preimage of the closed set $]-\infty, 0]$ through the continuous function $\eta(\cdot)$.

For $t \in [0,T]$ one has

$$h(t, \bar{y}(t)) \leq h(t, y_n(t)) + (h(t, \bar{y}(t)) - h(t, y_n(t))) \leq L|y_n(t) - \bar{y}(t)|.$$

Since $y_n(t) \to \bar{y}(t)$, letting $n \to +\infty$ in the previous relation we obtain $h(t, \bar{y}(t)) \leq 0$. Moreover, since (y_n, α_n) is minimizing sequence and $y_n(T) \to \bar{y}(T)$, then $\Phi(\bar{y}(T))$ coincides with the infimum of (P). □

We now announce the most common set of necessary conditions for a minimum of problem (P), usually known as the Pontryagin's Maximum Principle. Our formulation is based on the nonsmooth state constrained Maximum Principle [12, Thm. 9.3.1] (see also [12, Thm. 6.2.3] for the details about the right endpoint under consideration).

Theorem 3.3 *Assume* **(H1)-(H2)** *and let* $(\bar{y}, \bar{\alpha})$ *be a minimizer for* (P). *Then there exist* $\gamma \in \mathbb{R}_{\geq 0}$, $\beta \in \mathbb{R}_{\geq 0}$, $p \in W^{1,1}([0,T], \mathbb{R}^n)$ *and* $\mu \in NBV([0,T], \mathbb{R})$ *that fulfill the following*

conditions:

$$\|p\|_{L^\infty} + \|\mu\|_{TV} + \gamma + \beta \neq 0, \tag{3.1}$$

$$-\dot{p}(t) = q(t) \cdot \nabla_x f(t, \bar{y}(t), \bar{\alpha}(t)) \quad a.e.\ t \in [0,T], \tag{3.2}$$

$$-q(T) = \gamma \nabla \Phi(\bar{y}(T)) + \beta \nabla \eta(\bar{y}(T)), \quad \beta = 0 \text{ if } \eta(\bar{y}(T)) < 0 \tag{3.3}$$

$$q(t) \cdot f(t, \bar{y}(t), \bar{\alpha}(t)) = \max_{a \in A} \{q(t) \cdot f(t, \bar{y}(t), a)\} \quad a.e.\ t \in [0,T], \tag{3.4}$$

$$\mathrm{spt}(\mu) \subseteq \{t \in [0,T] \,:\, h(t, \bar{y}(t)) = 0\}, \tag{3.5}$$

where $q : [0,T] \to \mathbb{R}^n$ *is defined by*

$$\begin{cases} q(t) = p(t) + \int_{[0,t[} \nabla_x h(s, \bar{y}(s)) \mu(ds) & \text{if } t \in [0,T[, \\ q(T) = p(T) + \int_{[0,T]} \nabla_x h(t, \bar{y}(t)) \mu(dt) & \text{if } t = T. \end{cases} \tag{3.6}$$

We refer to (3.1) as *nontriviality condition*, to (3.2) as *adjoint equation*, to (3.3) as *transversality condition* and to (3.4) as *maximality condition*. Condition (3.5) locates the support of the measure μ inside the set of points in which the state constraint evaluated along the optimal trajectory is active.

Definition 3.4 A feasible process $(\bar{y}, \bar{\alpha})$ that fulfills conditions (3.1)–(3.5) is said to be an *extremal* for (P). We will call it *normal* if all possible choices of (γ, β, p, μ) as above have $\gamma > 0$ and *abnormal* when it is not normal.

Remark 3.5 Both Theorem 3.2 and Theorem 3.3 can be adapted to the case in which hypotheses **(H1)-(H2)** are replaced by the following assumptions: $f(t,x,a)$ satisfies conditions (1)-(2)-(3) of Remark 2.2. the control set $A(\cdot)$ is a measurable set valued function with compact values, Φ is locally Lipschitz continuous, h is an upper semicontinuous function which is locally Lipschitz continuous in x, $\mathcal{T}$ is some general closed subset of $\mathbb{R}^n$. For further details see [12, Ch. 6.9] and [3].

Remark 3.6 The MP remains true for *local minimizer*. In fact, in view of cut-off arguments, the conditions (3.1)–(3.5) are satisfied even if $(\bar{y}, \bar{\alpha})$ is an L^∞-local minimizer, that is, $\Phi(\bar{y}(T)) \leq \Phi(y(T))$ for any feasible process (y, α) for (P) such that $\|\bar{y} - y\|_{L^\infty(0,T)} \leq \delta$ for some $\delta > 0$.

Some final words about the different techniques that can be employed in order to prove the Maximum Principle, which are mainly two:

- *Variational approach:* take a minimizer for the original problem (P) and, for given $\varepsilon > 0$, with the help of the Ekeland's theorem, find a minimizer for a perturbed problem (P_ε) for which the necessary conditions are easy to derive since it has no right endpoint constraint. Finally, pass to the limit in these conditions for $\varepsilon \to 0$ and deduce necessary conditions for (P).

- *Set separation method:* notice that a minimizer for (P) defines a boundary point of the reachable set that comprises images of the terminal values of the state trajectories under a function having components the cost and equality constraint functions. Construct a (convex) approximating cone to the reachable set and use the Schauder fixed point theorem and a relaxation lemma to show that there exists a linear hyperplane separating this approximating cone by 0. This latter assertion easily implies the MP.

References for the first approach are [12, 3], while for the second one see [13]. These methods are complementary and it has often been the case that progresses based on one approach were followed by separate proofs based on the other approach, such as for the necessary conditions for nonsmooth problems proved with variational methods by Clarke and at the same time independently by Warga with set separation arguments.

4 Extension of optimal control problems and gap phenomena

The Maximum Principle provides a powerful tool for finding extremals that are canditate minimizers. However, there are many optimal control problems for which a minimizer does not exist. For instance, consider the following example that we will resume throughout the section.

Example 4.1 Consider the optimization problem given by

$$\text{(P)} \quad \begin{cases} \text{Min} \quad (y^2(1))^2 \\ \qquad \text{over } ((y^1,y^2),\alpha) \in W^{1,1}([0,1],\mathbb{R}^2)\times \mathcal{M}([0,1],\mathbb{R}) \text{ satisfying} \\ (\dot{y}^1(t),\dot{y}^2(t)) = (\alpha(t),(y^1(t))^2) \qquad \text{a.e. } t\in[0,1] \\ \alpha(t)\in A := \{-1,1\} \quad \text{a.e. } t\in[0,1] \\ (y^1(0),y^2(0)) = (0,0). \end{cases}$$

Notice that there are no feasible processes $((y^1,y^2),\alpha)$ with $(y^2(1))^2 = 0$. In fact,

$$0 = y^2(1) - y^2(0) = \int_0^1 (y^1(t))^2 dt \;\Rightarrow\; y^1 \equiv 0 \;\Rightarrow 0 \equiv \dot{y}^1 = \alpha.$$

A contradiction, since $\alpha \in \{-1,1\}$. However, we can exhibit a minimizing sequence. Consider (α_n) defined by

$$\alpha_n(t) := \begin{cases} 1 & \text{if } \sin(nt) \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

It is very easy to see that the sequence (y_n^1) such that $y_n^1(0) = 0$ and $\dot{y}_n^1 = \alpha_n$ converges uniformly to 0 and its absolute value is bounded by 1. Therefore the sequence (y_n^2) such that $y_n^2(0) = 0$ and $\dot{y}_n^2 = (y_n^1)^2$ satisfies

$$y_n^2(1) = \int_0^1 (y_n^1(t))^2 dt \to 0,$$

by the dominated convergence theorem. Hence this optimal control problem admits infimum but not minimum and the Maximum Principle becomes useless.

In control theory, when an optimal control problem (P) does not admit minimizers, it is a common practice to construct an *auxiliary* optimal control problem (P_a) by enlarging the set of admissible solutions. Then, once found a minimizer for (P_a), thanks to numerical approximation methods it is possible to find an ε-minimizer for (P), that is a feasible process $(y_\varepsilon, \alpha_\varepsilon)$ for (P) whose cost exceeds the infimum cost for (P) at most of a quantity ε, that is

$$\Phi(y_\varepsilon(T)) \leq \inf\{\Phi(y(T)) : (y,\alpha) \text{ feasible process for (P)}\} + \varepsilon.$$

In order that this procedure works it is necessary that (P_a) admits a minimizer and that there is *no infimum gap*, namely, the infimum of (P) coincides with the minimum of (P_a). Therefore, from now on our goal is to seek for sufficient conditions in order to ensure that there is no infimum gap between (P) and a suitable auxiliary problem (P_a). We will illustrate the results with some examples.

In Theorem 3.2 we have seen that two fundamental requirements for the existence of minimizers of an optimal control problem are the compactness of the control set A and the convexity of the set of the velocities $F(t,x) := \{f(t,x,a) : a \in A\}$. Accordingly, the most considered auxiliary problems are the following:

- The *extended problem* (P_e), in which we replace the original bounded (but not necessarily close) control set with its closure. It is a type of extension commonly used in optimal control problems with unbounded dynamics, usually known as *impulsive optimal control problems*.

- The *relaxed problem* (P_r), in which we replace the original dynamics with its convexification (from $\dot{y} = f(t,y,\alpha)$ to $\dot{y} = \sum_{j=0}^{n} \lambda_j f(t,y,\alpha_j)$, $\underline{\lambda} \in \Delta_n$). It is equivalent to a probabilistic extension of the original control problem (see [13]).

These different auxiliary problems have been always studied separately (see [14, 11, 10] for relaxed problem and [9, 5] for the extended problem). In [6] we managed to define for the first time an original unified framework that brings them together and that comprises as special case (relaxed) control polynomial impulsive optimization problems that have a wide application in Lagrangian mechanics. Further results based on this new unified framework can be found in [7, 8].

We consider the following optimization problem

$$\text{(P)} \quad \begin{cases} \text{Min} \quad \Phi(y(T)) \\ \qquad \text{over } (y,(\omega,\alpha)) \in W^{1,1}([0,T],\mathbb{R}^n) \times \mathcal{M}([0,T],\mathbb{R}^q \times \mathbb{R}^m) \text{ satisfying} \\ \dot{y}(t) = f(t,y(t),\omega(t),\alpha(t)) \qquad \text{a.e. } t \in [0,T], \\ (\omega(t),\alpha(t)) \in V \times A \qquad \text{a.e. } t \in [0,T], \\ (y(0),y(T)) = \{x_0\} \times \mathcal{T}, \\ h(t,y(t)) \leq 0 \qquad \forall t \in [0,T]. \end{cases}$$

In addition to **(H2)**, we make the following assumptions on the data

(H3) The control set $A \subset \mathbb{R}^m$ is compact, the control set $V \subset \mathbb{R}^q$ is bounded but not necessarily closed. The dynamics function $f(t,x,w,a)$ is continuous in all variables, continuously differentiable with respect to x. Moreover, there exists $C > 0$ such that for all $(t,x,w,a) \in \mathbb{R} \times \mathbb{R}^n \times V \times A$ it holds

$$|f(t,x,w,a)| \leq C, \qquad |\nabla_x f(t,x,w,a)| \leq C.$$

Furthermore, for any $(t,x,w,a),(t,x,w',a) \in \mathbb{R} \times \mathbb{R}^n \times V \times A$ it holds

$$|\nabla_x f(t,x,w,a) - \nabla_x f(t,x,w',a)| \leq C|w-w'| \tag{4.1}$$

Remark 4.2 Hypothesis **(H3)** emphasizes the different role played by the controls ω and α, as only the first one is extended. A situation where condition **(H3)** is verified, is when the dynamic function has a polynomial dependence on the control variable w, with bounded and globally Lipschitz continuous coefficients in the state variable. All the results of this section can be extended to a nonsmooth context as explained in Remark 3.5. In this case, condition (4.1) turns into an uniform continuity requirement in the variable w of the partial Clarke's generalized Jacobian with respect to x. See [6] for more details.

Let $W = \overline{V}$. Then we define the *extended optimal control problem* (P_e) of (P) as

$$(\mathrm{P}_e) \begin{cases} \text{Min} \quad \Phi(y(T)) \\ \qquad \text{over } (y,(\omega,\alpha)) \in W^{1,1}([0,T],\mathbb{R}^n) \times \mathcal{M}([0,T],\mathbb{R}^q \times \mathbb{R}^m) \text{ satisfying} \\ \dot{y}(t) = f(t,y(t),\omega(t),\alpha(t)) \qquad \text{a.e. } t \in [0,T], \\ (\omega(t),\alpha(t)) \in W \times A \qquad \text{a.e. } t \in [0,T], \\ (y(0),y(T)) \in \{x_0\} \times \mathcal{T}, \\ h(t,y(t)) \leq 0 \qquad \forall t \in [0,T]. \end{cases}$$

Moreover, the *relaxed (extended) optimal control problem* (P_r) of (P) is given by

$$(\mathrm{P}_r) \begin{cases} \text{Min} \quad \Phi(y(T)) \\ \text{over } y \in W^{1,1}([0,T],\mathbb{R}^n),\ \underline{\omega} = (\omega_0,\dots,\omega_n) \in \mathcal{M}([0,T],(\mathbb{R}^q)^{n+1}) \\ \underline{\alpha} = (\alpha_0,\dots,\alpha_n) \in \mathcal{M}([0,T],(\mathbb{R}^m)^{n+1}),\ \underline{\lambda} := (\lambda_0,\dots,\lambda_n) \in \mathcal{M}([0,T],\mathbb{R}^{n+1}) \text{ satisfying} \\ \dot{y}(t) = \sum_{j=0}^n \lambda_j(t) f(t,y(t),\omega_j(t),\alpha_j(t)) \qquad \text{a.e. } t \in [0,T], \\ (\underline{\omega},\underline{\alpha},\underline{\lambda}) \in W^{n+1} \times A^{n+1} \times \Delta_n \qquad \text{a.e. } t \in [0,T], \\ (y(0),y(T)) \in \{x_0\} \times \mathcal{T}, \\ h(t,y(t)) \leq 0 \qquad \forall t \in [0,T]. \end{cases}$$

Let Γ, Γ_e, Γ_r be the set of feasible processes for (P), feasible processes for (P_e) and feasible processes for (P_r), respectively. Of course, the fact that $V \subseteq W$ and Remark 2.7 imply that

$$\Gamma \subseteq \Gamma_e \subseteq \Gamma_r. \tag{4.2}$$

Definition 4.3 Let $\tilde{z} := (\tilde{y}, (\tilde{\omega}_0, \dots, \tilde{\omega}_n), (\tilde{\alpha}_0, \dots, \tilde{\alpha}_n), (\tilde{\lambda}_0, \dots, \tilde{\lambda}_n)) \in \Gamma_a$ for $a \in \{e, r\}$ and assume that $\tilde{z}$ is a minimizer for (P_a). We say that at $\tilde{z}$ there is a *local infimum gap* if, for some $\delta > 0$ one has

$$\Phi(\tilde{y}(T)) < \inf\{\Phi(y(T)) : y \in \Gamma,\ \|y - \tilde{y}\|_{L^\infty(0,T)} \leq \delta\}.$$

Definition 4.4 Let $\tilde{z} := (\tilde{y}, (\tilde{\omega}_0, \dots, \tilde{\omega}_n), (\tilde{\alpha}_0, \dots, \tilde{\alpha}_n), (\tilde{\lambda}_0, \dots, \tilde{\lambda}_n)) \in \Gamma_a$ for $a \in \{e, r\}$. We say that $\tilde{z}$ is *isolated* if for some $\delta > 0$ one has

$$\{y \in \Gamma : \|y - \tilde{y}\|_{L^\infty(0,T)} \leq \delta\} = \emptyset.$$

The following proposition relates the occurrence of gap phenomena with the topological property of isolation.

Proposition 4.5 *Let $\tilde{z} := (\tilde{y}, (\tilde{\omega}_0, \dots, \tilde{\omega}_n), (\tilde{\alpha}_0, \dots, \tilde{\alpha}_n), (\tilde{\lambda}_0, \dots, \tilde{\lambda}_n)) \in \Gamma_a$, $a \in \{e, r\}$, be a minimizer for (P_a). Then at $\tilde{z}$ there is a local infimum gap if and only if $\tilde{z}$ is isolated.*

Proof. Let $\tilde{z} \in \Gamma_a$, $a \in \{e, r\}$, be a minimizer for (P_a) and suppose that at $\tilde{y}$ there is a local infimum gap. Assume by contradiction that $\tilde{z}$ is not isolated, then there exists a sequence $(y_n) \subset \Gamma$ such that $\|y_n - \tilde{y}\|_{L^\infty} \to 0$. In particular $y_n(T) \to \tilde{y}(T)$. Hence, for some $\delta > 0$, the continuity of Φ implies

$$\Phi(\tilde{y}(T)) < \inf\{\Phi(y(T)) : y \in \Gamma,\ \|y - \tilde{y}\|_{L^\infty} \leq \delta\} \leq \lim_n \Phi(y_n(T)) = \Phi(\tilde{y}(T)).$$

Conversely, let $\tilde{z} \in \Gamma_a$, $a \in \{e, r\}$, be a minimizer for (P_a) and suppose that $\tilde{z}$ is isolated. Assume by contradiction that at $\tilde{z}$ there is no local infimum gap. Therefore, there exists a sequence $(y_n) \subset \Gamma$ such that

$$\|y_n - \tilde{y}\|_{L^\infty} \leq \frac{1}{n} \tag{4.3}$$

and $\Phi(y_n(T)) \to \Phi(\tilde{y}(T))$. However, (4.3) contradicts the fact that $\tilde{y}$ is isolated. □

Now we state the most important results of this paper, recalling that a feasible process (y, ω, α) for (P_e) can be interpreted as a feasible process $(\tilde{y}, (\tilde{\omega}_0, \dots, \tilde{\omega}_n), (\tilde{\alpha}_0, \dots, \tilde{\alpha}_n), (\tilde{\lambda}_0, \dots, \tilde{\lambda}_n))$ for (P_r) where $\tilde{y} = y$, $\tilde{\omega}_j = \omega$, $\tilde{\alpha}_j = \alpha$, $\tilde{\lambda}_j = \frac{1}{n+1}$ for any $j = 0, \dots, n$.

Theorem 4.6 *Assume* **(H2)**-**(H3)** *and let $\bar{z} := (\bar{y}, (\bar{\omega}_0, \dots, \bar{\omega}_n), (\bar{\alpha}_0, \dots, \bar{\alpha}_n), (\bar{\lambda}_0, \dots, \bar{\lambda}_n))$ be a minimizer for (P_a), for $a \in \{e, r\}$. If at $\bar{y}$ there is a local infimum gap, then $\bar{z}$ is an abnormal extremal for problem (P_a).*

As a corollary we deduce the following sufficient condition to avoid gap phenomena

Theorem 4.7 *Assume* **(H2)**-**(H3)** *and let $\bar{z} := (\bar{y}, (\bar{\omega}_0, \dots, \bar{\omega}_n), (\bar{\alpha}_0, \dots, \bar{\alpha}_n), (\bar{\lambda}_0, \dots, \bar{\lambda}_n))$ be a minimizer for (P_a), for $a \in \{e, r\}$. If $\bar{z}$ is a normal extremal for problem (P_a), then at $\bar{y}$ there is no local infimum gap.*

The proof is divided into several steps in which successive sequences of optimization problems are introduced that have as admissible controls only strict sense controls, and costs that measure how much a process violates the constraints. Using the Ekeland's variational Principle, minimizers are then built for these problems, which converge to the initial isolated process. Furthermore, applying a Maximum Principle to these approximate problems with reference to the above mentioned minimizers, we obtain in the limit a set of multipliers with $\gamma = 0$.

Remark 4.8 We point out what does "abnormal extremal for (P_a)" mean (see the statement of Theorem 4.6 in the case $a = r$. This should be enough also for the case $a = e$ in view of (4.2) and the embedding in Remark 2.7. In particular, Theorem 4.6 and the Maximum Principle applied to the particular optimization problem (P_r) imply that, if $(\bar{y}, (\bar{\omega}_0, \dots, \bar{\omega}_n), (\bar{\alpha}_0, \dots, \bar{\alpha}_n), (\bar{\lambda}_0, \dots, \bar{\lambda}_n))$ is a minimizer for (P_r) and at $\bar{y}$ there is a local infimum gap, then there exist $\beta \in \mathbb{R}_{\geq 0}$, $p \in W^{1,1}([0,T], \mathbb{R}^n)$ and $\mu \in NBV([0,T], \mathbb{R})$ that fulfill conditions (3.3), (3.5), and the following requirements:

$$\|p\|_{L^\infty} + \|\mu\|_{TV} + \beta \neq 0, \tag{4.4}$$

$$-\dot{p}(t) = \sum_{j=0}^{n} \bar{\lambda}_j(t) q(t) \cdot \nabla_x f(t, \bar{y}(t), \bar{\omega}_j(t), \bar{\alpha}_j(t)) \quad \text{a.e. } t \in [0,T], \tag{4.5}$$

(4.6)
$$q(t) \cdot f(t, \bar{y}(t), \bar{\omega}_j(t), \bar{\alpha}_j(t)) = \max_{(w,a) \in W \times A} \{q(t) \cdot f(t, \bar{y}(t), w, a)\} \quad \text{a.e. } t \in [0,T],\ \forall j = 0, \dots, n$$

where $q : [0,T] \to \mathbb{R}^n$ is defined as in (3.6). We could similarly rewrite the statement of Theorem 4.7.

Remark 4.9 Theorem 4.6 and Theorem 4.7 can be easily extended to nonsmooth free end time problems with Lipschitz continuous dependence in the time variable, through a time reparameterization technique (see [6, Sec. 4]). We managed to adapt these result even for nonsmooth free end time problems with measurable time dependence in t (see [8]): this kind of problems have financial applications, as we encounter such phenomena in a variety of threshold problems associated, for instance, with abrupt changes in a tariff or rate of return on investment at prespecified times.

Remark 4.10 When $h(0, x_0) = 0$, every minimizer turns out to be abnormal, so that Theorem 4.7 becomes inapplicable to deduce that there is no infimum gap. By adding suitable *nondegeneracy* hypotheses, we managed to convert the *normality test* of Theorem 4.7 into a *nondegenerate normality test* (see [6, Sec. 3]): in order to prove that there is no local infimum gap, it is sufficient to prove that normality holds not for all sets of multipliers, but only among the nondegenerate ones, i.e. multipliers (γ, β, p, μ) that satisfy the following strengthened nontriviality condition

$$\|q\|_{L^\infty} + \gamma + \beta + \mu(]0,T]) \neq 0,$$

where q is as in (3.6).

Remark 4.11 Instead of checking that all possible sets of multipliers have $\gamma > 0$, which is a really hard task in certain problems with several variables, in literature can be found some easily verifiable conditions on the data that ensure the normality of the extremals. See for instance [4, 5].

We conclude illustrating Theorem 4.6 and Theorem 4.7 with some examples.

Example 4.12 Consider the optimization problem in Example 4.1, which we have seen that does not admit minimizer. Since $A = \{-1, 1\}$ is compact, the extended problem (P_e) coincides with (P). On the contrary, in view of Remark 2.6, the relaxed problem (P_r) can be obtained by (P) by replacing A with $\operatorname{co} A = [-1, 1]$ in the control constraint. Since the cost is nonnegative, we deduce that $\bar{z} := ((\bar{y}^1, \bar{y}^2), \bar{\alpha}) = ((0, 0), 0)$ is feasible and also a minimizer for (P_r). We now show that $\bar{z}$ is a normal extremal for (P_r), hence the absence of gap follows from Theorem 4.7, and it is confirmed by the minimizing sequence constructed in Example 4.1. From Theorem 3.3 we know that $\bar{z}$ is an extremal for (P_r) for some set of multipliers (γ, β, p, μ). Since there are no state constraint (i.e. $h \equiv -1$) and right endpoint constraint (i.e. $\eta \equiv -1$), from (3.5) and (3.3) we immediately deduce that $\mu = 0$ and $\beta = 0$. By (3.2) and (3.3) we deduce

$$\begin{cases} (-\dot{p}_1, -\dot{p}_2)(t) = (2p_2(t)\bar{y}^1(t), 0) = (0, 0) & \text{a.e. } t \in [0, 1] \\ (-p_1, -p_2)(T) = \gamma(0, 2\bar{y}^2(1)) = (0, 0). \end{cases}$$

Hence, $(p_1, p_2) \equiv 0$ and (3.1) implies $\gamma > 0$, so that $\bar{z}$ is a normal extremal for (P_r).

Now we suggest an optimal control problem which completely exploit the strength of our unified framework considered along the section and takes into account also state and endpoint constraints. Thanks to the following example, it can also be understood that gap phenomena can occur also when the set of trajectories (not necessarily feasible) for the original optimization problem (P) is dense in the set of trajectories for the suitable auxiliary problem (P_a) under consideration, due to the presence of constraints. In the following example you can notice that Theorem 2.8 implies that the set of trajectories (not necessarily feasible) for (P_e) is dense in the set of trajectories for (P_r). Nevertheless, there is an infimum gap also between (P_e) and (P_r).

Example 4.13 Consider the following optimization problem

$$(\mathrm{P}) \quad \begin{cases} \text{Min } \ -y^1(1) \\ \qquad \text{over } y = (y^1, y^2, y^3) \in W^{1,1}([0, 1], \mathbb{R}^3),\ (\omega, \alpha) \in \mathcal{M}([0, 1], \mathbb{R} \times \mathbb{R}) \text{ satisfying} \\ (\dot{y}^1(t), \dot{y}^2(t), \dot{y}^3(t)) = (0, \ \ y^1(t)\alpha(t), \ \ (y^2(t))^2 + \omega(t)) \qquad \text{a.e. } t \in [0, 1] \\ \omega(t) \in V :=]0, 1], \ \ \alpha(t) \in A := \{-1, 1\} \quad \text{a.e. } t \in [0, 1], \\ y^2(0) = y^3(0) = 0, \ \ y^3(1) \le 0, \\ y^1(t) - 1 \le 0 \ \ \forall t \in [0, 1]. \end{cases}$$

Problem (P) do not admit feasible processes. In fact, if (y, ω, α) were a feasible process, then

$$0 \ge y^3(1) - y^3(0) = \int_0^1 [(y^2(t))^2 + \omega(t)]dt > 0. \tag{4.7}$$

The extended problem (P_e) is obtained from (P) by replacing the control constraint $\omega(t) \in V$ with $\omega(t) \in \overline{V} = [0,1]$, and $(\check{y}, \check{\omega}, \check{\alpha}) \equiv ((0,0,0),0,1)$ is a minimizer for (P_e) with $-\check{y}^1(1) = 0$. In fact, if (y, ω, α) is feasible for (P), reasoning as in (4.7) one deduces $\omega \equiv 0$ and $y^2 \equiv 0$, so that $0 = \dot{y}^2 = \alpha y^1$. Since $\dot{y}^1 = 0$, then $y^1 \equiv y^1(0)$. Hence, $0 = \alpha y^1(0)$ and $\alpha \in \{-1, 1\}$ imply $y^1 \equiv 0$.

The relaxed problem (P_r) is obtained from (P_e) by replacing the control constraint $\alpha(t) \in A$ with $\alpha(t) \in \operatorname{co} A = [0,1]$. Reasoning as before and taking account of the state constraint $y^1(t) - 1 \leq 0$ and the fact that now the control function α may assume the value 0, one deduces that $(\bar{y}, \bar{\omega}, \bar{\alpha}) \equiv ((1,0,0),0,0)$ is a minimizer for (P_r) with $-\bar{y}^1(1) = -1$. In particular we have shown

$$\inf_{y \in \Gamma_r} -y^1(1) = -1 \quad < \quad \inf_{y \in \Gamma_e} -y^1(1) = 0 \quad < \quad \inf_{y \in \Gamma} -y^1(1) = +\infty.$$

Because of the presence of infimum gap, we can illustrate Theorem 4.6. In fact, $(\check{y}, \check{\omega}, \check{\alpha}) \equiv ((0,0,0),0,1)$ and $(\bar{y}, \bar{\omega}, \bar{\alpha}) \equiv ((1,0,0),0,0)$ are abnormal extremals for (P_e) and (P_r), respectively: a set of (abnormal) multipliers for both of them is given by

$$(\gamma, \beta, p, \mu) \equiv (0, 1, (0,0,-1), 0).$$

References

[1] Aubin, J.-P., Cellina, A., "Differential inclusions. Set-valued maps and viability theory". Springer-Verlag, Berlin, 1984.

[2] Bressan A., Piccoli B., "Introduction to the Mathematical Theory of Control". AIMS Series on Applied Mathematics, vol. 2, Springfield, MO, 2007.

[3] Clarke F.H., "Optimization and Nonsmooth Analysis". Wiley-Interscience, New York, 1983, reprinted as vol. 5 of Classics in Applied Mathematics, SIAM, Philadelphia, 1990.

[4] Fontes F.A.C.C., Frankowska H., *Normality and nondegeneracy for optimal control problems with state contraints*. J. Opt. Theory Appl., vol. 166, no. 1, pp. 115–136, 2015.

[5] Fusco G., Motta M., *No Infimum Gap and Normality in Optimal Impulsive Control Under State Constraints*. Set-Valued Var. Anal., vol. 29, no. 2, pp. 519–550, 2021.

[6] Fusco G., Motta M., *Nondegenerate abnormality, controllability, and gap phenomena in optimal control with state constraints*. SIAM J. Control Optim., vol. 60, no. 1, pp. 280–309, 2022.

[7] Fusco G., Motta M., *Strict sense minimizers which are relaxed extended minimizers in general optimal control problems*. For the Proc. of the 60th IEEE Conference on Decision and Control, CDC 2021, Austin (Texas), to appear.

[8] Fusco G., Motta M., *Gap phenomena and controllability in free end-time problems with active state constraints*. To appear in Journal of Mathematical Analysis and Applications.

[9] Motta M., Rampazzo F., Vinter R.B., *Normality and gap phenomena in optimal unbounded control*. ESAIM: Contr., Opt. and Calc. of Var., vol. 24, no. 4, pp. 1645–1673, 2018.

[10] Palladino, M.; Rampazzo, F., *A geometrically based criterion to avoid infimum gaps in optimal control*. J. Differential Equations, vol. 269, no. 11, pp. 10107–10142, 2020.

[11] Palladino M., Vinter R.B., *When are minimizing controls also minimizing extended controls?*. Discr. Cont. Dyn. Syst., vol. 35, no. 9, pp. 4573–4592, 2015.

[12] Vinter R.B., "Optimal control". Birkhäuser, Boston, 2000.

[13] Warga J., "Optimal control of differential and functional equations". Academic Press, New York, 1972.

[14] Warga J.,, *Controllability, extremality, and abnormality in nonsmooth optimal control*. J. Optim. Theory Appl. vol. 41, no. 1, pp. 239–260, 1983.

The Graph p-Laplacian Eigenvalue Problem

Piero Deidda (*)

Graphs are used to model a large variety of physical phenomena typically involving interactions between particles/components, such as, for example, the graph rapresentation of a molecule, the connections between users on social media, the transportation networks idealized as road maps or air routes. The study of the topology of a graph as well as the problem of computing topological invariants become important tools to acquire information about the distributions and the relations between the data represented in the graph. Within this framework we could think of two different approaches to the problem. The first tries to approximate topological invariants of the graph by developping specific algorithms and studying their accuracy and robustness, i.e. their convergence toward the real solution. The second approach wonders about what we can learn about the topology of the graph starting from some invariants that we already know how to calculate. Think for example to the spectrum of the Laplacian operator and to the famous problem *Can one hear the shape of a drum?* [20], meaning does the spectrum of the Laplace-Beltrami operator fully characterize the domain? This conjecture was then proved to be false [16] but in the meantime many different geometrical properties had been related to properties of the Laplacian spectrum [25]. In this manuscript we deal with the second approach and a after a short introduction about the graph setting and the p-Laplacian operator and its spectrum we show some old and new results about the topological information that can be deduced from the study of the p-Laplacian eigenvalues and eigenfunctions.

1 The Graph Setting

An undirected graph, $\mathcal{G}$, can be denoted by a triple $\mathcal{G} := (V, E, \omega)$, where V is the discrete set of the nodes (or vertices) of the graph, $E \subset V \times V$ denotes the set of the edges and is such that if $(u, v) \in E$ then also $(v, u) \in E$, and finally $\omega : E \to \mathbb{R}$ is a function on the edges such that $\omega(u, v) = \omega(v, u)$ that can be thought to represent the reciprocal of the edge length. To simplify the notation in the following we will often use the notation $\omega_{uv} := \omega(u, v)$. Using these definitions we can introduce a distance between two nodes u

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: deidda@math.unipd.it . Seminar held on 13 April 2022.

and v of the graph defined as the length of the shortest path joining them:

$$d(u,v) = \min_{u=u_1,\dots,u_n=v} \sum_{i=1}^{n-1} \frac{1}{\omega(u_i, u_{i+1})} \tag{1}$$

Denote by $\mathcal{H}(V)$ and $\mathcal{H}(E)$ the Hilbert spaces of the functions on the nodes and on the edges of the graph, respectively, endowed with the scalar products:

$$\langle f, g\rangle_{\mathcal{H}(V)} = \sum_{u\in V} f(u)g(u) \qquad \langle F, G\rangle_{\mathcal{H}(E)} = \frac{1}{2} \sum_{(u,v)\in E} F(u,v)G(u,v)\,. \tag{2}$$

We can also introduce the basic equivalents to differential operators of the continuous setting. Start with the gradient of a nodal function defined as the function that reproduces the slope of f on the edges:

$$\begin{aligned} \nabla :\ \mathcal{H}(V) &\longrightarrow \mathcal{H}(E) \\ f\ &\longrightarrow \nabla f(u,v) = \omega_{uv}(f(v) - f(u))\,, \end{aligned} \tag{3}$$

with u and v being vertices of the edge (u,v) and with the obvious property that $\nabla f(u,v) = -\nabla f(v,u)$. Next we introduce the divergence operator. Not considering a boundary on graphs is usually understood to be analogous to having homogeneous Neumann boundary conditions, thus to preserve the classical divergencce theorem in the continuous setting, i.e.

$$-\langle f, \operatorname{div} G\,\rangle_{\mathcal{H}(V)} = \langle \nabla f, G\rangle_{\mathcal{H}(E)}\,, \tag{4}$$

we may define the divergence as the half of minus the adjoint of the gradient, that in matrix form reads $-\mathrm{div} = \frac{1}{2}\nabla^T$, i.e.

$$\begin{aligned} \operatorname{div} :\ \mathcal{H}(E) &\longrightarrow \mathcal{H}(V) \\ G\ &\longrightarrow \operatorname{div} G(u) = \frac{1}{2} \sum_{v\sim u} \omega_{uv}\big(G(u,v) - G(v,u)\big)\,, \end{aligned} \tag{5}$$

where $\{v\,|v \sim u\}$ are the nodes connected to the node u by an edge, i.e. such that $(u,v) \in E$. Given the definitions of gradient and divergence we can introduce the graph Laplacian operator ($p = 2$) and the more general p-Laplacian operator $\big(p \in (1,\infty)\big)$, whose definitions are similar to the one used in the continuous setting:

$$\Delta_p f(u) = -\operatorname{div}\big(|\nabla f|^{p-2} \odot \nabla f\big)(u) = \sum_{v\sim u} \omega_{uv} |\nabla f(v,u)|^{p-2} \nabla f(v,u)\,, \tag{6}$$

where $|\nabla f|^{p-2}$ has to be understood entrywhise and $\odot$ denotes the entrywhise product (we will omit this symbol in the following).

2 p-Laplacian eigenpairs

Now we can introduce the p-Laplacian eigenvalue problem. Mimicking the continuous setting, we look for the critical points equation of the Rayleigh quotient:

$$\mathcal{R}_p(f) = \frac{\|\nabla f\|_p}{\|f\|_p} = \frac{\left(\frac{1}{2}\sum_{(u,v)\in E} |\nabla f(u,v)|^p\right)^{\frac{1}{p}}}{\left(\sum_{u\in V} |f(u)|^p\right)^{\frac{1}{p}}}\,, \tag{7}$$

whose critical point equation $\big(p \in (1,\infty)\big)$, up to rescalings, reads:

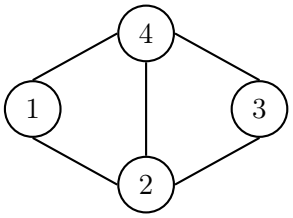
$$\Delta_p f(u) = \mathcal{R}_p^p(f)|f(u)|^{p-2}f(u) \quad \forall u \in V\,. \tag{8}$$

We thus define (f,λ) to be a p-Laplacian eigenapair iff

$$\Delta_p f(u) = \lambda|f(u)|^{p-2}f(u) \quad \forall u \in V\,. \tag{9}$$

Multiplying the above equation by $f(u)$ and then summing over the vertices shows that if λ is an eigenvalue corresponding to the eigenfunction f, necessarily $\lambda = \mathcal{R}_p^p(f)$.

We highlight in Figure 1, from [11], that differently from the linear case $p = 2$, where Δ_2 is a symmetric positve semidefinite matrix, for a generic p the number of p-Laplacian eigenpairs can be greater than the dimension of the space, i.e $|V|$, the eigenpairs are not in general orthogonal and there is no clear notion of eigenspaces or multiplicity. In [29, 1] other examples and discussions on the problem.



(a) $f_1 = (1,1,1,1), \quad \lambda_1 = 0$

(b) $f_2 = (1,0,-1,0), \quad \lambda_2 = 2$

(c) $f_3 = (0,1,0,-1), \quad \lambda_3 = 2 + 2^{p-1}$

(d) $f_4 = (1,0,1,-2^{\frac{1}{p-1}}),\ (1,-2^{\frac{1}{p-1}},1,0)$
$\lambda_4 = 1 + \big(1 + 2^{\frac{1}{p-1}}\big)^{p-1}$

(e) $f_5 = (1,-1,1,-1), \quad \lambda_5 = 2^p$

Figure 1. Left: Example graph in which the corresponding p-Laplacian Δ_p with $\omega_{uv} = 1\ \forall (u,v) \in E$, has more eigenvalues then the dimesion of the space. Right: Set of five eigenvalues and corresponding eigenfunctions.

Nevertheless, using some classical results from calculus of variations it is possible to characterize a set of "variational" eigenvalues whose multiplicity is equal to the dimension of the space, $|V|$, and that somehow are representants of the whole spectrum. In detail, we observe that, because of the homogeneity of the Rayleigh quotient $\mathcal{R}_p$, we can restrict the study of its critical points on the p-sphere, $S_p := \{f \in \mathcal{H}(V) \,|\, \|f\|_p = 1\}$. Defined $\mathcal{R}_p^c := \{f \mid \mathcal{R}_p(f) < c\}$ consider the follwing deformation lemma and its direct consequence

given in Theorem 2.2 (a particular case of more general and classic results, see e.g. [24, 27, 15, 14].

Lemma 2.1 (Deformation Lemma) *Assume c to be a regular value of $\mathcal{R}_p$, then there exist $\epsilon > 0$ and a continuous family of deformations, $\phi \in C([0,1] \times S_p, S_p)$ such that $\phi(t,f) = -\phi(t,-f)\ \forall (t,f)$, $\phi(1, \mathcal{R}_p^{c+\epsilon}) \subset \mathcal{R}_p^{c-\epsilon}$, $\phi(0,f) = f$.*

Proof. We give a skecth of proof that is quite intuitive. Consider a neighborhood B, of $\{f \mid \mathcal{R}_p(f) = c\}$ without critical points, a cutoff function $\xi(f)$ that is zero outside B and the projection of the gradient of $\mathcal{R}_p(f)$ on the tangent space of S_p, denoted, with a small abuse of notation, by $\frac{\partial}{\partial f}\mathcal{R}_p(\cdot)$. Finally define $\phi(t,f)$ as the solution to the gradient flow

$$\begin{cases} \frac{\partial}{\partial t}\phi(t,f) = -\xi(\phi(t,f))\frac{\partial}{\partial f}\mathcal{R}_p(\phi(t,f)) \\ \phi(0,f) = f \end{cases}$$

which is a continuous map from $[0,1] \times S_p$ to S_p. □

Theorem 2.2 *Assume $\mathcal{F}$ to be a family of subsets of S_p such that for any regular value $c \in \mathbb{R}$ of $\mathcal{R}_p$, there exist $\epsilon > 0$ and a continuous deformation of the domain $\phi[0,1] \times S_p \to S_p$ s.t.*

$$\begin{cases} \phi : (0, \cdot) = id_{S_p}(\cdot) \\ \phi(1, \mathcal{R}_p^{c+\epsilon}) \subset \mathcal{R}_p^{c-\epsilon} \\ \phi(t, A) \in \mathcal{F}\,,\ \forall A \in \mathcal{F}\,,\ \forall t \in [0,1] \end{cases}$$

Then

$$\Lambda := \inf_{A \in \mathcal{F}} \sup_{f \in A} \mathcal{R}_p(f)\,,$$

is a critical value of $\mathcal{R}_p$, i.e. the p-th root of an eigenvalue of Δ_p.

Proof. The proof is a direct consequence of the deformation lemma 2.1. □

Based on the above Theorem we can introduce the variational eigenapairs of the p-Laplacian. The theorem says that we have to find families, $\mathcal{F}_k$, of subsets stable for deformations, i.e, if $A \in \mathcal{F}_k$ and ϕ is a deformation also $\phi(A) \in \mathcal{F}_k$. To understand how this works recall the Fisher-Courant min max characterization of the eigenvalues of a symmetric matrix (arising from the graph Laplacian) i.e.

$$\lambda_k(\Delta_2) = \min_{dim(A) \geq k} \max_{f \in A \setminus \{0\}} \frac{\langle \Delta_2 f, f \rangle}{\langle f, f \rangle} = \min_{dim(A) \geq k} \max_{f \in A \setminus \{0\}} \mathcal{R}_2^2(f)\,.$$

A possible strategy (not the unique one) to generalize this min max theorem to the non-linear case, using Theorem 2.2, is based on the idea of considering a generalized notion of dimension, the Krasnoselskii genus, that is based/related to the Lyusternik-Schnirelmann category of a space [27, 15, 14]. First of all observe that, as we are interested in studying

critical points of $\mathcal{R}_p$, which is an even functional, it would be enough to generalize the notion of dimension to the symmetric subsets. Thus we introduce

$$\mathcal{A} = \{A \subseteq \mathbb{R}^n |\ A \text{ closed}\,,\ A = -A\}$$

Then we observe that in the case of A a linear subspace of dimension k, $A \setminus \{0\}$ can be retracted with continuity on a sphere of dimension $k-1$, S^{k-1}. This notion can be generalized defining, for any $A \in \mathcal{A}$, the Krasnoselskii genus of A:

$$\gamma(A) = \begin{cases} \inf\{k \in \mathbb{N} \,|\exists\, \psi \in C(A, S^{k-1})\ \ s.t.\ \ \psi(x) = -\psi(-x)\} \\ +\infty \qquad\qquad\qquad\qquad \text{if } \nexists\, k \text{ as above} \end{cases}$$

We first note that, if $\gamma(A) \geq k$ and $\phi \in C(\mathbb{R}^n, \mathbb{R}^n)$, $\gamma(\phi(A)) \geq \gamma(A)$. Hence, the families $\mathcal{F}_k(S_p) = \{A \subseteq \mathcal{A} \cap S_p \,|\gamma(A) \geq k\}$ satisfy the hypotheses of Theorem 2.2 Thus we define the Krasnoselskii variational eigenvalues of Δ_p as

$$\lambda_k^{\frac{1}{p}} = \inf_{A \in \mathcal{F}_k} \sup_{f \in A} \mathcal{R}_p(f)\,. \tag{10}$$

Morover, since we are working in a finite dimensional space, it is possible to prove that the above inf sup is actually a min max.

2.1 Cases $p = 1, \infty$

A particular discussion is necessary for the two extreme cases $p = 1$ and $p = \infty$, observe that in these cases the Rayleigh quotients, $\mathcal{R}_1(f)$ and $\mathcal{R}_\infty(f)$, are still well defined but not differentiable anymore. This opens the problem of how to define the 1 and the infinity eigenpairs. The answer to this problem is not unique, and different approaches have been proposed in the literature. Here, we discuss an approach that has been initially used for the case $p = 1$ [6, 17], but that has been recently used also in the continuous setting for the infinity case [4, 3]. The idea is to define a generalized notion of critical points for the Rayleigh quotients $\mathcal{R}_1(f)$ and $\mathcal{R}_\infty(f)$. To this aim, we first observe that given f, a p-Laplacian eigenfunction, we can assume w.lo.g. $\|f\|_p = 1$, which is equivalent to saying that f is a point of the unit sphere S_p (the sphere of unit p-norm). Then observe that $|f|^{p-2}f$ is the outward normal to S_p in f. As a consequence, from (9), f being a p-Laplacian eigenfunction, is equivalent to have $\partial\|\nabla f\|_p/\partial f$ $(= C\Delta_p(f))$ equal, up to rescalings, to the outward normal to the manifold S_p in the point f. Trying to generalize the notion of critical points there are two difficulties. The first is the non differentiability of the 1 or ∞ norm of the gradient and the second is the fact that the outward normal to the spheres S_1 and S_∞ is not everywhere well defined. A solution to both of these problems comes from the notion of subgradients of a convex function [26]. Let $\Psi : \mathbb{R}^n \to \mathbb{R}$ be a convex function, e.g a norm. Its subgradient at a point f_0 is defined as:

$$\partial\Psi(f_0) = \left\{\xi \mid \Psi(g) - \Psi(f_0) \geq \langle \xi, g - f_0\rangle\ \forall g \in \mathbb{R}^n\right\}. \tag{11}$$

This is a generalization of the notion of gradient: if the function Ψ is differentiable at the point f_0, then $\partial\Psi(f_0) = \big(\partial\Psi/\partial f\big)(f_0)$. Morover it is possible to characterize the

composition of the subdifferential of a convex function with a linear transformation (see Theorem 23.9 [26] for weaker hypotheses):

Theorem 2.3 *Let $\Phi(f) = \Psi(Af)$, where Ψ is a convex function on $\mathbb{R}^m$, $|\Psi(f)| < +\infty \; \forall f \in \mathbb{R}^m$ and A is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$, then*

$$\partial\Phi(f) = A^T \partial(\Psi(Af))$$

These results allow us to define a generalized notion of $\partial\|\nabla f\|/\partial f$, meaningful also in case of 1 and infinity norms, that matches the classical definition for $1 < p < \infty$. Now we need to generalize the notion of outward normal to the spheres of p-norm equal to one. As the outward normal doesn't change, instead of considering the sphere let us consider the corresponding closed ball. Then

$$D_p = \{f \mid \|f\|_p \leq 1\}$$

is easily proved to be a convex set for any p and we can define the convex external cone in the generic point, f_0 s.t. $\|f_0\|_p = 1$ as

$$C_{Ext}(f_0) = \{\xi \,|\, \langle \xi, f_0 - g\rangle \leq 0 \; \forall \, g \in D_p\}\,.$$

Again it is possible to relate this set to the subgradient of $f \to \|f\|_p$ since it is possbile to prove that [26, 7]:

$$C_{Ext}(f_0) = \bigcup_{\lambda \geq 0} \lambda \partial \|f_0\|\,. \tag{12}$$

It follows that saying that f is an eigenfunction of the p-Laplacian is equivalent to asking that there exist $\Lambda > 0$ such that

$$\emptyset \neq \partial\|\nabla f\|_p \cap \Lambda \partial \|f\|_p\,, \tag{13}$$

and now this definition makes sense also when $p = 1$ and $p = \infty$.

To complete the discussion about the two nonsmooth cases we need to characterize the sets $\partial\|f\|_1$ and $\partial\|f\|_\infty$. To this end we recall the following result [5], of which we include the proof because of its simplicity.

Lemma 2.4 *Given a function f_0 and a norm $\|\cdot\|$,*

$$\partial\|f_0\| = \{\xi \mid \|g\| \geq \langle \xi, g\rangle \; \forall \, g, \; \|f_0\| = \langle \xi, f_0\rangle\}$$

Proof. Observe that the right-hand side is trivially included in the left one by definition of subgradient. About the opposite inclusion, by definition and the triangular inequality, we know that if $\xi \in \partial\|f_0\|$

$$\langle \xi, h - f_0\rangle = \langle \xi, h\rangle - \langle \xi, f_0\rangle \leq \|h\| - \|f_0\| \leq \|h - f_0\| \quad \forall g$$

Then, as $g := (h - f_0)$ spans the whole $\mathbb{R}^n$, the sup in the above inequality yields

$$0 \geq \sup_g \Big(\langle \xi, g \rangle - \|g\| \Big) \geq \langle \xi, f_0 \rangle - \|f_0\|,$$

on the other hand, still from the subgradient definition, we have:

$$\langle \xi, f_0 \rangle - \|f_0\| \geq \Big(\sup_g \langle \xi, g \rangle - \|g\| \Big) \geq 0\,.$$

Thus if $\xi \in \partial \|f_0\|$ it satisfies:

$$\langle \xi, f_0 \rangle - \|f_0\| = \sup_g \Big(\langle \xi, g \rangle - \|g\| \Big) = 0\,,$$

concluding the proof. □

Observe that from this lemma it follows that necessarily, if f is an eigenfunction as in (13), $\Lambda = \mathcal{R}_p(f)$ i.e. it is the p-th root of the eigenvalue defined in (9). However observe that when $p = \infty$, $\mathcal{R}_\infty^\infty$ is not defined while for $p = 1$ there is no difference between $\mathcal{R}_1(f)$ and its 1-th root, thus for the two extreme cases, we will call, with a small abuse of notation, Λ in (13) the eigenvalue corresponding to f.

The subgradients of the 1 and ∞ norm can be calculated from Lemma 2.4 and Theorem 2.3 yielding the following formulas:

$$\begin{aligned} \partial \|f\|_1 &= \Big\{ \xi_V \;\Big|\; \xi_V(u) = \operatorname{sign}\big(f(u)\big) \Big\} \\ \partial \|\nabla f\|_1 &= \Big\{ -\operatorname{div} \xi_E \;\Big|\; \xi_E(u,v) = \operatorname{sign}\big(\nabla f(u,v)\big) \Big\} \end{aligned} \tag{14}$$

where $\operatorname{sign}(x) = \begin{cases} 1 & if\ x > 0 \\ [-1,1] & if\ x = 0 \\ -1 & if\ x < 0 \end{cases}$.

$$\begin{aligned} \partial \|f\|_\infty &= \left\{ \xi_V \;\middle|\; \begin{array}{l} \|\xi_V\|_1 = 1,\ \xi_V(u) = 0\ if\ |f(u)| < \|f\|_\infty \\ \xi_V(u) f(u) = \big|\xi_V(u) f(u)\big| \end{array} \right\} \\ \partial \|\nabla f\|_\infty &= \left\{ -\operatorname{div} \xi_E \;\middle|\; \begin{array}{l} \|\xi_E\|_1 = 1,\ \xi_E(u,v) = 0\ if\ |\nabla f(u,v)| < \|\nabla f\|_\infty \\ \xi_E(u,v) \nabla f(u,v) = \big|\xi_E(u,v) \nabla f(u,v)\big| \end{array} \right\} \end{aligned} \tag{15}$$

We conclude this part recalling that, also in the degenerate cases, the min max in (10) characterizes eigenvalues as generalized critiacal values (13), allowing to define the variational eigenvalues also for $p = 1$ and $p = \infty$ provided subgradients are used (see [6, 7] for the details).

3 Nodal Domains and properties of the p-Laplacian eigenpairs

The nodal domains induced by a function f are generally the maximal subdomains induced by the sign of f. Their study in relation to the eigenpairs of the Laplacian dates back to the works of Sturm and Courant where they observed respectively for strings and membranes that the number of nodal domains induced by the eigenfunctions of the Laplacian-Beltrami operator, somehow, reflects the corresponding frequency (i.e. eigenvalue index).

We here briefly discuss and recall few results about the nodal domains induced by the p-Laplacian eigenfunctions. We start from the definition:

Definition 3.1 (Nodal domains) Given a graph $\mathcal{G}$ and a function $f : V \to \mathbb{R}$, a subset of the vertices, $A \subseteq V$, is a nodal domain induced by f if the subgraph $\mathcal{G}_A \subset \mathcal{G}$ with vertices in A is a maximal connected subgraph of $\mathcal{G}$ where f is nonzero and has constant sign. We denote by $\mathcal{N}(f)$ the number of nodal domains induced by a function f.

Exactly as in the case of the Laplacian, it turns out that also for the graph p-Laplacian it is possible to relate the number of nodal domains induced by an eigenfunction to the corresponding frequency, where now the frequency is intended in terms of position of the eigenvalue with respect to the variational spectrum. In particular, without entering in the details of the sharpest bounds, in [28, 11], it can be shown that

Theorem 3.2 *Suppose that $\mathcal{G}$ is a connected graph, $1 < p < \infty$ and $\lambda_1 < \lambda_2 \leq \cdots \leq \lambda_N$ are the variational eigenvalues of Δ_p.*

- *If f is an eigenfunction of Δ_p with eigenvalue λ such that $\lambda < \lambda_k$, then*
$$\mathcal{N}(f) \leq k - 1\,;$$
- *if f is an eigenfunction of Δ_p with eigenvalue λ such that $\lambda > \lambda_k$, then*
$$\mathcal{N}(f) \geq k - \beta - z(f) + 1\,,$$
where β is the number of independent loops of the graph, i.e. $\beta = |E| - |V| + 1$, and $z(f)$ is the number of nodes where f is zero.

These results, jointly with the fact that the nodal domains are expected to provide a sort of equilibrated partition of the graph, led to the idea of using the eigenvalues of the p-Laplacian to approximate the "costs" of some "optimal" partitions of the graph, providing geometrical information about the graph itself.

3.1 $p = 1$ and Cheeger constants

We start this subsection by introducing the family of the Cheeger constants of the graph that are able to provide information about the number and the goodness of the clusters of the graph. Let $A \subset V$ be a subset of the nodes and consider $E(A, \bar{A})$ be the set of the edges having one endpoint in A and the other in $\bar{A} := V \setminus A$. Consider, then, the quantity

$$c(A) = \frac{\|\omega\big(E(A,\bar{A})\big)\|_1}{|A|} = \frac{\frac{1}{2}\sum_{(u,v)\in E(A,\bar{A})}\omega(u,v)}{|A|}\,, \tag{16}$$

and, given an integer k, all the possible families of k nonempty and disjoint subsets of V

$$\mathcal{D}_k(\mathcal{G}) = \{A_1, \dots, A_k \subset V \mid A_i \neq \emptyset,\ A_i \cap A_j = \emptyset\ \forall\, i, j\}\,. \tag{17}$$

Define the k-th Cheeger constant (or isoperimetric constant) see [22, 21, 8], as

$$h_k(\mathcal{G}) := \min_{\{A_1,\dots A_k\}\in\mathcal{D}_k(\mathcal{G})} \max_{i=1,\dots,k} c(A_i)\,, \tag{18}$$

Observe that having a "small" value of h_k means that there exist k-good subset of nodes that are quite massive and "poorly connected" to each other, that is the idea of a cluster of nodes. The constant $h_k(\mathcal{G})$ can thus be considered as an indicator of how well the graph can be clusterized in k subgraphs with the corresponding family of subsets being the approximate clusters.
Now observe that given a subset, $A \subset V$, and considered its characteristic function, χ_A, we have that

$$\mathcal{R}_1(\chi_A) = \frac{\frac{1}{2}\sum_{(u,v)\in E} \omega_{uv}|\chi_A(u) - \chi_A(v)|}{\sum_{v\in V}|\chi_A(u)|} = c(A)\,. \tag{19}$$

It is then natural to study the Cheeger constants in relation to the eigenvalues of the 1-Laplacian. What is actually possible to prove is the following theorem [2, 6, 9, 17, 28]

Theorem 3.3 *Let Λ_k^1 be the k-th variational eigenvalue of the 1-Laplacian, then*

$$\Lambda_2^1 = h_2(\mathcal{G})\,, \qquad \Lambda_k^1 \leq h_k(\mathcal{G}) \quad \forall k. \tag{20}$$

Moreover using the p-Laplacian eigenpairs, when p goes to 1, it is possibile to prove the following theorem that relates the number of nodal domains induced by an eigenfunction and the Cheeger constants, [9, 28]

Theorem 3.4 *Let $(f_k, \lambda_k^{(p)})$ be the k-th variational eigenpair of the p-Laplacian, $p > 1$, then*

$$\frac{2^{p-1}}{\tau(\mathcal{G})^{p-1}} \frac{h^p_{\mathcal{N}(f)}}{p^p} \leq \lambda_k^{(p)}\,,$$

where $\tau(\mathcal{G}) = \max_{u\in V}\sum_{v\sim u}\omega(u,v)$

Combining the last two theorems we observe that whenever we have a variational eigenfunction whose nodal domain count reflects the corresponding frequency, letting p goes to one, the eigenvalue reproducess exactly an higher order Cheeger constant.

3.2 $p = \infty$

Analogous results relate the ∞-eigenpairs to the maximal distance among k nodes. It is worth mentioning that some of these results are similar to those obtained in the continuous

setting by [19, 18, 13] using an approach different from the one that employs the subgradients,.
To describe this case we start by introducing the k-th radius of the graph:

$$r_k = \max_{v_1,\dots,v_k \in V} \min_{i,j=1,\dots,k} \frac{d(v_i, v_j)}{2} \tag{21}$$

that can also be written as

$$r_k = \max\{r \mid \exists\, v_1, \dots, v_k \in V \ s.t.\ d(v_i, v_j) \geq 2r \ \forall\, i, j = 1, \dots, k\}\,. \tag{22}$$

Observe that $2r_2$ is obviously the diameter of the graph, while, for $k > 2$ we are computing the maximal reciprocal distance among k nodes. In terms of informations about a set of data represented by the graph, r_k measures a sort of higher order distribution width of the data.
Similarly to what shown before, (19), we can observe that given a node v and a radius r (not too small) we can build the cone function $f : V \to \mathbb{R}$, $f(u) := \max\{r - d(u, v), 0\}$, then it is trivial to observe that

$$\mathcal{R}_\infty(f) = \frac{1}{r}\,.$$

In addition, recalling the ∞-Laplacian eigenvalue problem from equation (15), if (f, Λ) is an ∞-eigenpair, there exist $\xi_V \in \partial\|f\|_\infty$ and $\xi_E \in \partial\|\nabla f\|_\infty$ such that

$$-\mathrm{div}(\xi_E) = \Lambda \xi_V\,.$$

We know that if $\xi_V(u) \neq 0$ then necessarily $f(u) = \|f\|_\infty$ which is equivalent to say that in any non extremal node v of f $\mathrm{div}(\xi_E)(v) = 0$. On the other hand, $\xi_E(u, v) \neq 0$ necessarily means $\nabla f(u, v) = \|\nabla f\|_\infty$. From these two observations, starting from an extremal node u of f ($\xi_V(u) \neq 0$) there has to exist a path, whose edges are here indicated by $\{(v_i, v_{i+1})\}_{i=1}^{n-1}$ with $v_1 = u$ $v_n = w$, such that $|\nabla f(v_i, v_{i+1})| = \|\nabla f\|_\infty \ \forall\, i$. Morover, the end point of the path can only be another point w such that $|f(w)| = \|f\|_\infty$, $f(w) = -f(u)$. It is then easy to arrive at the conclusion that

$$\Lambda = \frac{\|\nabla f\|_\infty}{\|f\|_\infty} = \frac{2}{d(u, w)} \tag{23}$$

Since any ∞-eigenpair is related to a distence and to some sort of cone function, it makes again sense to relate the infinite variational eigenvalues to the radii of the graph. The following results, very similar to the one presented for the $p = 1$ case, hold [10]:

Theorem 3.5 *Let $\Lambda_k^{(\infty)}$ be the k-th ∞-variational eigenvalue, then*

$$\Lambda_2^{(\infty)} = \frac{1}{R_2}, \qquad \Lambda_k^{(\infty)} \leq \frac{1}{R_k} \quad \forall k$$

Morover in the case of eigenpairs obtained as limit for $p \to \infty$ of p-Laplacian eigenapairs we have the following

Theorem 3.6 *Let (f, Λ^∞) be an ∞-eigenpair that is a limit of p-Laplacian eigenpairs as $p \to \infty$, then*

$$\frac{1}{R_{\mathcal{N}(f)}} \leq \Lambda^{(\infty)}$$

Thus we recover the same remarks made at the end of the previous subsection 3.1.

4 Computing the p-Laplacian eigenpairs

We conclude this summary presenting just a hint of a possible numerical approach for the computation of the p-Laplacian eigenpairs. From (9) we observe that , for $p > 2$, the p-Laplacian eigenvalue equation can be written as a constrained linear eigenvalue problem weighted with respect to two positive measures, one on the edges, $\mu_0 : E \to \mathbb{R}^+$, and one on the nodes, $\nu_0 : V \to \mathbb{R}^+$:

$$\begin{cases} \Delta_{\mu_0} f(u) = \Big(- \operatorname{div}\big(\operatorname{diag}(\mu_0)\big)\nabla f\Big)(u) = \lambda \nu_{0u} f(u) & \forall u \in V \\ \mu_{0uv} = |\nabla f(u,v)|^{p-2} & \forall (u,v) \in E \\ \nu_{0u} = |f(u)|^{p-2} & \forall u \in V \end{cases}$$

Then, for any couple of positive measures $\mu : E \to \mathbb{R}^+$ $\nu : V \to \mathbb{R}^+$ it is possible to consider the eigenpairs $(f, \lambda)(\mu, \nu)$ of the generalized eigenvalue problem

$$\Delta_\mu f(u) = \Big(- \operatorname{div}\big(\operatorname{diag}(\mu)\big)\nabla f\Big)(u) = \lambda \nu_u f(u)\,.$$

These are eigenpairs of a linear eigenvalue problem and thus can be enumerated from 1 to $|V|$, the cardinality of the node space.
For any $k \leq |V|$ we introduce the function

$$\mathcal{L}_k(\mu, \nu) = \frac{1}{\lambda_k(\mu,\nu)} + \frac{p-2}{p} \sum_{(u,v)\in E} (\mu_{uv}^{\frac{p}{p-2}}) - \frac{p-2}{p} \sum_{v \in V} (\nu_v^{\frac{p}{p-2}}) \tag{24}$$

Note that these functions are still well defined when $p = \infty$ by defining $\frac{p}{p-2} = 1$. Morover, as we see in the next theorem, [12], it is possible to relate the saddle points of these functions with the p-Laplacian eigenpairs. Denoting by $\mathcal{M}^+(V), \mathcal{M}^+(E)$ the space of the positive measures on V and E, we have the following result:

Theorem 4.1 *Let, $p > 2$ and $(\nu^*, \mu^*) := \operatorname{argmax}_{\nu \in \mathcal{M}^+(V)} \operatorname{argmin}_{\mu \in \mathcal{M}^+(E)} \mathcal{L}_k(\mu, \nu)$ be a smooth saddle point of the function $\mathcal{L}_k(\mu, \nu)$, then $\big(\lambda_k^{\frac{p}{2}}(\mu^*, \nu^*), f_k(\mu^*, \nu^*)\big)$ is a p-Laplacian eigenpair.*

This result leds to the construction of numerical algorithms based on gradient flows for these functionals that at each step only require the computation of an eigenpair of a weighted Laplacian, taking the possibility to use all the advantages of the linearity. We refer to [12] for more details.

References

[1] S. Amghibech, *Eigenvalues of the discrete p-Laplacian for graphs.* Ars Combinatoria 67: 283–302, 04 2003.

[2] T. Bühler and M. Hein, *Spectral clustering based on the graph.* Proceedings of the 26th International Conference on Machine Learning, 01 2009.

[3] L. Bungert and Y. Korolev, *Eigenvalue problems in L^∞: Optimality conditions, duality, and relations with optimal transport.* ArXiv preprint (`arxiv:2107.12117, 2021`).

[4] L. Bungert, Y. Korolev, and M. Burger, *Structural analysis of an l-infinity variational problem and relations to distance functions.* Pure and Applied Analysis 2, 2020.

[5] M. Burger, G. Gilboa, M. Moeller, L. Eckardt, and D. Cremers, *Spectral decompositions using one-homogeneous functionals.* SIAM Journal on Imaging Sciences, 9(3): 1374–1408, 2016.

[6] K.C. Chang, *Spectrum of the 1-laplacian and cheeger?s constant on graphs.* Journal of Graph Theory, 81(2): 167–207, 2016.

[7] K.-C. Chang, S. Shao, D. Zhang, and W. Zhang, *Nonsmooth critical point theory and applications to the spectral graph theory.* Science China Mathematics, 64(1): 1–32, 2021.

[8] A. Daneshgar, H. Hajiabolhassan, and R. Javadi, *On the isoperimetric spectrum of graphs and its approximations.* Journal of Combinatorial Theory, Series B, 100(4): 390–412, 2010.

[9] A. Daneshgar, R. Javadi, and L. Miclo, *On nodal domains and higher-order Cheeger inequalities of finite reversible Markov processes.* Stochastic Processes and their Applications 122(4): 1748–1776, 2012.

[10] P. Deidda, M. Putti, and M. Burger, *The graph ∞-laplacian eigenvalue problem* (tentative title). In preparation.

[11] P. Deidda, M. Putti, and F. Tudisco, *Lower bounds for the nodal domains of the generalized graph p-laplacian.* Submitted.

[12] P. Deidda, N. Segala, and M. Putti, *Computing p-laplacian eigenpairs by a dynamical method* (tentative title). In preparation.

[13] L. Esposito, B. Kawohl, C. Nitsch, and C. Trombetti, *The Neumann eigenvalue problem for the ∞-laplacian.* Rendiconti Lincei - Matematica e Applicazioni, 26, 05 2014.

[14] S. Fucik, J. Necas, J. Soucek, and V. Soucek, "Spectral analysis of nonlinear operators". Volume 346. Springer LNM, 2006.

[15] N. Ghoussoub, "Duality and Perturbation Methods in Critical Point Theory". Cambridge Tracts in Mathematics. Cambridge University Press, 1993.

[16] C. Gordon, D.L. Webb, and S. Wolpert, *One cannot hear the shape of a drum.* Bulletin of the American Mathematical Society, 27(1): 134–138, 1992.

[17] M. Hein and T. Bühler, *An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse pca.* Advances in Neural Information Processing Systems, 23, 2010.

[18] P. Juutinen and P. Lindqvist, *On the higher eigenvalues for the ∞-eigenvalue problem.* Calculus of Variations and Partial Differential Equations 23: 169–192, 06 2005.

[19] P. Juutinen, P. Lindqvist, and J. Manfredi, *The ∞-eigenvalue problem.* Archive for Rational Mechanics and Analysis 148: 89–105, 09 1999.

[20] M. Kac, *Can one hear the shape of a drum?.* The American Mathematical Monthly, 73(4P2): 1–23, 1966.

[21] G.F. Lawler and A.D. Sokal , *Bounds on the ℓ^2 spectrum for Markov chains and Markov processes: a generalization of Cheeger?s inequality.* Transactions of the American Mathematical Society 309(2): 557–580, 1988.

[22] J.R. Lee, S.O. Gharan, and L. Trevisan, *Multiway spectral partitioning and higher-order Cheeger inequalities.* Journal of the ACM (JACM), 61(6): 1–30, 2014.

[23] P. Lindqvist, "Notes on the infinity Laplace equation". Springer Briefs in Mathematics, 2016.

[24] R.S. Palais, *Critical point theory and the minimax principle.* In Proc. Symp. Pure Math, volume 15, pages 185–212. American Mathematical Society Providence, 1970.

[25] M. Protter, *Can one hear the shape of a drum? revisited.* SIAM Review, 29(2): 185–197, 1987.

[26] R.T. Rockafellar, "Convex Analysis, volume 36". Princeton University Press, 1970.

[27] M. Struwe, "Variational methods, volume 991". Springer, 2000.

[28] F. Tudisco and M. Hein, *A nodal domain theorem and a higher-order Cheeger inequality for the graph p-Laplacian.* Journal of Spectral Theory 8: 883–908, 03 2016.

[29] D. Zhang, *Homological eigenvalues of graph p-laplacians.* ArXiv, 2021.

Kolmogorov-Arnold-Moser (KAM) stability and its application in the planetary n-body problem

Rita Mastroianni (*)

Abstract. The study of exoplanetary systems with two or more planets in orbits with non-zero mutual inclination is an interesting topic of Hamiltonian dynamics, in view of the many applications related to the astronomical discovery, in the last 20 years, of several such systems. The present report discusses the mathematical context of the theory of the long term stability for nearly Keplerian perturbed n-body systems, following the so-called Kolmogorov-Arnold-Moser (KAM) Theorem. The KAM Theorem is a cornerstone of canonical perturbation theory: it allows to conjugate, through a convergent sequence of canonical trasformations, particular solutions of the "pertubed" dynamical system to the invariant dynamics on a torus. We provide a short summary of classical results of perturbation theory. We also briefly present some recent progress on the construction of the Kolmogorov normal form for 'isochronous systems'. Finally, we explain in an introductory manner, how the above concepts can be implemented in exoplanetary systems with a 3D-orbital architecture.

1 Introduction

The general *problem of dynamics*, as defined by Poincaré in *Les méthodes nouvelles de la mécanique céleste* ([11]) is described by the following Hamiltonian

$$\mathcal{H}(\boldsymbol{\varphi}, \boldsymbol{I}) = h(\boldsymbol{I}) + \varepsilon\, f(\boldsymbol{\varphi}, \boldsymbol{I}) \tag{0.1}$$

whith action variables $\boldsymbol{I} \in \mathcal{G} \subset \mathbb{R}^n$ ($\mathcal{G}$ an open set), angle variables $\boldsymbol{\varphi} \in \mathbb{T}^n$ and ε a "small" parameter. The Hamiltonian (1) is called "nearly-integrable": it is composed by two terms, $h(\boldsymbol{I})$ and $\varepsilon\, f(\boldsymbol{\varphi}, \boldsymbol{I})$. By the *Liouville-Arnold-Jost Theorem* 3.2 the first term $h(\boldsymbol{I})$ is integrable; thus we can establish that the orbits (i.e. the solutions of Hamilton's equations only for the part $h(\boldsymbol{I})$) lie on invariant n-dimensional tori, parametrized by the values of the actions $\boldsymbol{I}(0)$. However, the complete Hamiltonian $\mathcal{H}(\boldsymbol{\varphi}, \boldsymbol{I})$ is "perturbed" by the presence of the term $\varepsilon\, f(\boldsymbol{\varphi}, \boldsymbol{I})$. At first, one can (wrongly) conjecture that since the complete Hamiltonian $\mathcal{H}(\boldsymbol{\varphi}, \boldsymbol{I})$ is close to the integrable one $h(\boldsymbol{I})$, the same happens for its dynamics. However a wide range of new phenomena appear as a result of the perturbation.

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: rita.mastroianni@math.unipd.it. Seminar held on 11 May 2022.

In the present chapter we will discuss how Hamiltonian perturbation theory allows to find normalization procedures (i.e. a sequence of canonical change of coordinates) able to lead to a control of the dynamics of the complete system, as well as a quantification of the difference between the perturbed and the integrable part.

This work is organized as follows. In section 2 we give some basic definitions regarding Hamiltonian systems. Sections 3 and 4 we describe, respectively, the integrable and perturbative cases, explaining the theorems and the techniques required for the treatment of such systems. Finally, in section 5 we focus on applications. In particular we will see how the theory explained in the previous sections can be applied in order to understand and explain the main phenomena for a three body problem with a high mutual inclination.

2 Definitions and basic notions of Hamiltonian systems

Given the Hamiltonian function $\mathcal{H} : \mathcal{F} \to \mathbb{R}$, with $(\boldsymbol{q}, \boldsymbol{p}) \in \mathcal{F} \subseteq \mathbb{R}^n \times \mathbb{R}^n$, *Hamilton's equations* are described by the following system of $2n$ differential equations of the first order:

$$\begin{cases} \dot{q}_j = \partial \mathcal{H}(\boldsymbol{q}, \boldsymbol{p})/\partial p_j \\ \dot{p}_j = -\partial \mathcal{H}(\boldsymbol{q}, \boldsymbol{p})/\partial q_j \end{cases} \qquad j = 1, \dots, n \ , \tag{2}$$

where with the dot $\dot{}$ we mean the time derivative d/dt. The variables $(\boldsymbol{q}, \boldsymbol{p})$ are called positions and momenta respectively. The system (2) describes the motion of a *n-degree of freedom* dynamical system (where n is the number of independent coordinates $\boldsymbol{q}$ required to describe the system.)[(2)] The solutions of Hamilton's equations (2) is called *flow* and it maps the inizial values $(\boldsymbol{q}(0), \boldsymbol{p}(0))$ to the solution at time t, i.e. $\phi^t_{\mathcal{H}}(\boldsymbol{q}(0), \boldsymbol{p}(0)) = (\boldsymbol{q}(t), \boldsymbol{p}(t))$. An *orbit* is defined as $\cup_{t \in I} \phi^t_{\mathcal{H}}$, where I is the maximal set of definition of the solution. In general the Hamiltonian can depend also explicitly on the time; nevertheless, through an 'extension of the phase space' it is possible to remove this last dependence. Whenever $\partial \mathcal{H}/\partial t = 0$ the Hamiltonian is said *autonomous*. For an autonomous Hamiltonian the value $\mathcal{E} = \mathcal{H}(\boldsymbol{q}(t), \boldsymbol{p}(t))$ is preserved along any orbit. $\mathcal{E}$ is called the energy of the system. A classical method to visualize the global behaviour of a system is to pass from the continuous flow $\phi^t_{\mathcal{H}}$ to a discrete map $\Pi : \Sigma \to \Sigma$, with $\Sigma \subset \mathbb{R}^{2n}$ a surface of dimension $2n-1$; this is called the Poincaré section method. In particular, taken a surface Σ transversal to the flow, an orbit is followed until crossing, in a given direction, the surface Σ; i.e. let P_0 be the first point of intersection between the flow and Σ, $P_1 := \Pi(P_0)$ be the successive point of intersection, and so on. An orbit is then represented by the sequence of the successive intersections $P_0, P_1 = \Pi(P_0), P_2 = \Pi(P_1), \dots$, as shown in Figure 1.

[(2)] In general, denoting by M the n-dimensional manifold in which the motion takes place, the phase space is the cotangent bundle of M, i.e. $\mathcal{F} = T^*M$ being $(\boldsymbol{q}, \boldsymbol{p}) \in T^*M$. However, we can consider the simple case in which the phase space $\mathcal{F}$ is an open set of $\mathbb{R}^{2n}$.

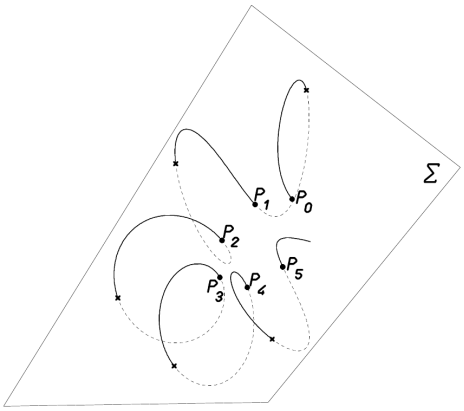


Figure 1: The method of Poincaré section for an orbit in the three dimensional space. This figure is available from [3].

Now we give a series of definitions useful in the sequel.

Definition 2.1 The Poisson bracket between two functions f and g is the bilinear map $\{\cdot,\cdot\} : \mathcal{C}^\infty(\mathcal{F}) \times \mathcal{C}^\infty(\mathcal{F}) \to \mathcal{C}^\infty(\mathcal{F})$ defined by

$$\{f,g\} = \sum_{j=1}^{n} \left(\frac{\partial f}{\partial q_j} \frac{\partial g}{\partial p_j} - \frac{\partial f}{\partial p_j} \frac{\partial g}{\partial q_j} \right) = \nabla f \cdot \mathbb{J} \nabla g \,,$$

where $\mathbb{J} = \begin{pmatrix} \mathbb{O}_n & \mathbb{I}_n \\ -\mathbb{I}_n & \mathbb{O}_n \end{pmatrix}$ is the symplectic matrix, with $\mathbb{O}_n$ and $\mathbb{I}_n$ respectively the $n \times n$ zero and unit matrices.

Definition 2.2 The function $f(\boldsymbol{q}, \boldsymbol{p})$ is called first integral of $\mathcal{H}$ if it is constant under the Hamiltonian flow, i.e.

$$\dot{f} = \sum_{j=1}^{n} \left(\frac{\partial f}{\partial q_j} \frac{\partial \mathcal{H}}{\partial p_j} - \frac{\partial f}{\partial p_j} \frac{\partial \mathcal{H}}{\partial q_j} \right) = \{f, \mathcal{H}\} := L_{\mathcal{H}} f = 0 \,.$$

The operation $L_{\mathcal{H}} f$ is called the Lie derivative of the function f along the Hamiltonian vector field.

Example: an autonomous Hamiltonian is a first integral of its own flow, since $L_{\mathcal{H}} \mathcal{H} = 0$.

Definition 2.3 The functions $f_1, \ldots, f_r \in \mathcal{C}^\infty(\mathcal{F})$ are called

- independent, if $\operatorname{rank}\left(\dfrac{\partial(f_1, \ldots, f_r)}{\partial(q_1, \ldots, q_n, p_1, \ldots, p_n)} \right) = r$;
- in involution, if $\{f_i, f_j\} = 0 \;\; \forall\, i,\, j = 1, \ldots, r$.

Definition 2.4 Consider the change of coordinates $(\boldsymbol{q}, \boldsymbol{p}) \mapsto (\boldsymbol{Q}, \boldsymbol{P})$, given by $\boldsymbol{z} = \boldsymbol{z}(\boldsymbol{y})$, with $\boldsymbol{y} := \begin{pmatrix} \boldsymbol{q} \\ \boldsymbol{p} \end{pmatrix}$ and $\boldsymbol{z} := \begin{pmatrix} \boldsymbol{Q} \\ \boldsymbol{P} \end{pmatrix}$. Let $M = \partial \boldsymbol{z} / \partial \boldsymbol{y}$ be the Jacobian matrix of

the tranformation. The transformation is called canonical if the matrix M satisfies the symplectic property

$$M\mathbb{J}M^T = M^T\mathbb{J}M = \mathbb{J}.$$

If the trasformation $\boldsymbol{z} = \boldsymbol{z}(\boldsymbol{y})$ is canonical, then the new variables $(\boldsymbol{Q}, \boldsymbol{P})$ satisfy Hamilton's equations for $H(\boldsymbol{z}) = \mathcal{H}(\boldsymbol{y}(\boldsymbol{z}))$.
Example: the *Hamiltonian flow* defines a canonical mapping. Namely, given $(\boldsymbol{q}, \boldsymbol{p}) \in \mathcal{F}$ the evolution of the orbit with initial condition $(\boldsymbol{q}, \boldsymbol{p})$ along the Hamiltonian flow leads, after a time t, to the mapping $(\boldsymbol{q}_t, \boldsymbol{p}_t) = \phi^t_{\mathcal{H}}(\boldsymbol{q}, \boldsymbol{p})$ which is canonical.
For the treatment of perturbative systems, there will be crucial the research of canonical transformation, near to the identity. For this reason we need the following definition:

Definition 2.5 Given $\chi \in \mathcal{C}^\infty(\mathcal{F})$ we call *Lie series operator* the exponential operator of εL_χ, i.e.

$$\exp(\varepsilon L_\chi)\cdot = \sum_{j\geq 0} \frac{\varepsilon^j}{j!} L^j_\chi \cdot .$$

Remembering that $\dot{f} = \{f, \mathcal{H}\}$, and performing a Taylor series expansion, it is easy to prove that, for any function $f \in \mathcal{C}^\infty(\mathcal{F})$ we have $f(\boldsymbol{q}(t), \boldsymbol{p}(t)) = \exp(tL_\mathcal{H}) f(\boldsymbol{q}(0), \boldsymbol{p}(0))$; this means that the Lie series map the function from its initial value in $t = 0$ to its value at time t, along the Hamiltonian flow. Then, taking as function f the canonical coordinates, we have that

$$(\boldsymbol{q}(\varepsilon), \boldsymbol{p}(\varepsilon)) = \phi^\varepsilon_{\mathcal{H}}(\boldsymbol{q}(0), \boldsymbol{p}(0)) = \big(\exp(\varepsilon L_\mathcal{H})\,\boldsymbol{q}, \exp(\varepsilon L_\mathcal{H})\,\boldsymbol{q}\big)\big|_{\substack{\boldsymbol{q}=\boldsymbol{q}(0)\\ \boldsymbol{p}=\boldsymbol{p}(0)}}.$$

Thus, the transformation

$$(\boldsymbol{q}, \boldsymbol{p}) \longmapsto \Big(\exp(\varepsilon L_\mathcal{H})\,\boldsymbol{q}, \exp(\varepsilon L_\mathcal{H})\,\boldsymbol{p}\Big)$$

is a canonical transformation near to the identity, i.e.

$$\big(\exp(\varepsilon L_\mathcal{H})\,\boldsymbol{q}\,, \exp(\varepsilon L_\mathcal{H})\,\boldsymbol{p}\big) - (\boldsymbol{q}, \boldsymbol{p}) = \mathcal{O}(\varepsilon)\,.$$

In practice, in Hamiltonian perturbation theory, we use the method of the Lie series to generate changes of coordinates of this form, i.e., given an appropriate generating function χ, we will consider $\exp(\varepsilon L_\chi)$. To this end, the following theorem will be useful:

Theorem 2.6 (Exchange) *Given $\chi, f \in \mathcal{C}^\infty(\mathcal{F})$, then*

$$f(\boldsymbol{q}, \boldsymbol{p})|_{\substack{\boldsymbol{q}=\exp(\varepsilon L_\chi)\widehat{\boldsymbol{q}}\\ \boldsymbol{p}=\exp(\varepsilon L_\chi)\widehat{\boldsymbol{p}}}} = \exp(\varepsilon L_\chi) f|_{\substack{\boldsymbol{q}=\widehat{\boldsymbol{q}}\\ \boldsymbol{p}=\widehat{\boldsymbol{p}}}}.$$

Theorem 2.6 establishes that, given a function f, the act of replacing the 'old' variables with the 'new' ones (obtained through the Lie series canonical transformation) is equivalent

to acting directely on the function with the Lie series, and renaming, at the end, the 'old' variables as the 'new' ones.
Finally, in order to characterize the type of motion, we recall the definition of *quasi-periodic* function:

Definition 2.7 Let $g(\theta_1, \dots, \theta_n)$ be a function periodic in the angles $\boldsymbol{\theta} \in \mathbb{T}^n$, i.e. such that

$$g(\theta_1, \dots, \theta_i + 2\pi, \dots \theta_n) = g(\theta_1, \dots, \theta_i, \dots \theta_n) \quad \forall\, i = 1, \dots\, n.$$

The function $f(t) = g(\omega_1 t, \dots, \omega_n t)$ is called quasi-periodic in t with respect to the frequency vector $\boldsymbol{\omega} \in \mathbb{R}^n$ when $\boldsymbol{\omega}$ is non-resonant, i.e. $\boldsymbol{k} \cdot \boldsymbol{\omega} \neq 0\ \forall\, \boldsymbol{k} \in \mathbb{Z}^n \setminus \{\boldsymbol{0}\}$.

3 Integrability

In the present section we briefly present the concept of integrability for Hamiltonian systems and we characterize the solutions of an integrable Hamiltonian. In order to define the concept of integrability we need the following theorem:

Theorem 3.1 (Liouville) *Let $\mathcal{H} : \mathcal{F} \to \mathbb{R}$ be a n-degrees of freedom Hamiltonian with n first integrals $f_1, \dots, f_n$ independent and in involution. Then the system is integrable by quadrature.*

From now on by *integrability* we mean *Liouville integrability*. In order to characterize the solutions of an integrable system, the following theorem is essential:

Theorem 3.2 (Liouville-Arnold-Jost) *Let $\mathcal{H} : \mathcal{F} \to \mathbb{R}$ be a n-degrees of freedom Hamiltonian admiting n first integrals independent and in involution $f_i : \mathcal{F} \to \mathbb{R}\ \ i = 1, \dots, n$. Assume there exists a compact and connected component $M_{\boldsymbol{c}}$ of the level set $\{(\boldsymbol{q}, \boldsymbol{p}) \in \mathcal{F} : f_i(\boldsymbol{q}, \boldsymbol{p}) = c_i, i = 1, \dots, n\}$, $\boldsymbol{c} = (c_1, \dots, c_n) \in \mathbb{R}^n$. Then:*

- *$M_{\boldsymbol{c}}$ is diffeomorphic to the n-dim torus $\mathbb{T}^n$;*
- *in a neighbourhood U of $M_{\boldsymbol{c}}$ there exists a canonical transformation*

$$\begin{array}{rcl} \mathcal{C} : \mathcal{G} \times \mathbb{T}^n & \to & U \\ (\boldsymbol{I}, \boldsymbol{\varphi}) & \mapsto & (\boldsymbol{p}, \boldsymbol{q}) \end{array}$$

  with $\mathcal{G} \subseteq \mathbb{R}^n$ an open set, such that the Hamiltonian takes the form $\mathcal{H}(\mathcal{C}(\boldsymbol{I}, \boldsymbol{\varphi})) = h(\boldsymbol{I})$.

The canonical variables $(\boldsymbol{I}, \boldsymbol{\varphi})$ are called action-angle variables. The equations of motion in these variables are

$$\begin{cases} \dot{\varphi}_j = \dfrac{\partial h(\boldsymbol{I})}{\partial I_j} := \omega_{0_j}(\boldsymbol{I}) \\ \dot{I}_j = -\dfrac{\partial h(\boldsymbol{I})}{\partial \varphi_j} = 0 \end{cases} \qquad j = 1, \dots, n. \tag{3}$$

Thus the orbits lie on n-dimensional tori, parametrized by the actions, with linear motions on $\mathbb{T}^n$, i.e. $t \mapsto \{(\boldsymbol{I}, \boldsymbol{\varphi}) : \boldsymbol{I}(t) = \boldsymbol{I}_0, \boldsymbol{\varphi(t)} = \boldsymbol{\varphi}_0 + \boldsymbol{\omega}_0(\boldsymbol{I}_0)t\}$ are the solutions of (3). The motions are non-periodic, and are instead dense in $\mathbb{T}^n$, if the frequency vector $\boldsymbol{\omega}_0$ is *non-resonant*, i.e. $\boldsymbol{k}\cdot\boldsymbol{\omega}_0 \neq 0 \,\forall\, \boldsymbol{k} \in \mathbb{Z}^n \setminus \{\boldsymbol{0}\}$. In this case the motion is called quasi-periodic. In particular, if $n = 2$, setting $\boldsymbol{\omega}_0 = (\omega_1, \omega_2)$, the orbit is *periodic* if and only if $\omega_1/\omega_2 \in \mathbb{Q}$; otherwise, the orbit is dense on $\mathbb{T}^2$. Also, it is useful to observe that, in a Poincaré section Σ, the linear flow on a torus yields a finite number of points, if the orbit is periodic, or a closed curve, if the orbit is quasi-periodic.

4 Nearly-integrability

In the previous section we have characterized the solutions of an integrable system. Thus, taken the general problem of dynamics (1), described by

$$\mathcal{H}(\boldsymbol{\varphi}, \boldsymbol{I}) = h(\boldsymbol{I}) + \varepsilon\, f(\boldsymbol{\varphi}, \boldsymbol{I})$$

where $\boldsymbol{I} \in \mathcal{G} \subset \mathbb{R}^n$, $\boldsymbol{\varphi} \in \mathbb{T}^n$, we know that for $\varepsilon = 0$ the Hamiltonian is integrable (by the Liouville-Arnold-Theorem 3.2) and the motions are conjugated to linear flows on $\mathbb{T}^n$. The question now is: if $\varepsilon \neq 0$, do there exist quasi-periodic solutions lying on invariant tori? The answer to this question is given by the celebrated Kolmogorov-Arnold-Moser (KAM) Theorem, which states that, under suitable assumptions and if the size of the perturbation ε is 'small enough', the existence of invariant tori is ensured. Furthermore, these tori are deformations of those of the integrable case.

Theorem 4.1 (KAM (according to Kolmogorov)) *Consider a Hamiltonian function $\mathcal{H} : \mathbb{T}^n \times \mathcal{G} \to \mathbb{R}$ (where $\mathcal{G} \subseteq \mathbb{R}^n$ open) of the form $\mathcal{H}(\boldsymbol{\varphi}, \boldsymbol{I}) = \boldsymbol{\omega}_0 \cdot \boldsymbol{I} + h(\boldsymbol{I}) + \varepsilon\, f(\boldsymbol{\varphi}, \boldsymbol{I})$ where $h(\boldsymbol{I}) = \mathcal{O}(||\boldsymbol{I}||^2)$ for $\boldsymbol{I} \to 0$.*
Let us assume the following hypotheses:

1. *$\boldsymbol{\omega}_0$ is Diophantine, i.e. $\exists$ two constants $\gamma > 0$ and $\tau \geq n-1$ s.t. $|\boldsymbol{k}\cdot\boldsymbol{\omega}_0| \geq \gamma|\boldsymbol{k}|^{-\boldsymbol{\tau}} \quad \forall\, \boldsymbol{k} \in \mathbb{Z}^n \setminus \{\boldsymbol{0}\}$;*
2. *$\mathcal{H}$ is analytic on $\mathcal{G} \times \mathbb{T}^n$;*
3. *$h(\boldsymbol{I})$ is non-degenerate, i.e. $\det\left(\partial^2 h(\boldsymbol{I})/(\partial I_i \partial I_j)\right)_{i,j} \neq 0 \;\forall\, \boldsymbol{I} \in \mathcal{G}$;*
4. *ε is a small parameter, i.e. $\exists\, \varepsilon_\star > 0$ s.t. $|\varepsilon| \leq \varepsilon_\star$.*[3]

Then, there exists a canonical transformation $(\boldsymbol{\varphi}, \boldsymbol{I}) = \mathcal{C}_\varepsilon(\tilde{\boldsymbol{\varphi}}, \tilde{\boldsymbol{I}})$ leading $\mathcal{H}$ in the so called Kolmogorov normal form, i.e. $\mathcal{K}(\tilde{\boldsymbol{\varphi}}, \tilde{\boldsymbol{I}}) = \mathcal{H}\left(\mathcal{C}_\varepsilon(\tilde{\boldsymbol{\varphi}}, \tilde{\boldsymbol{I}})\right)$, where $\mathcal{K}(\tilde{\boldsymbol{\varphi}}, \tilde{\boldsymbol{I}}) = \boldsymbol{\omega}_0 \cdot \tilde{\boldsymbol{I}} + \mathcal{O}(||\tilde{\boldsymbol{I}}||^2)$.

We can easily verify that if the Hamiltonian is in Kolmogorov normal form $\mathcal{K}(\tilde{\boldsymbol{\varphi}}, \tilde{\boldsymbol{I}}) =$

[3] The definition of this ε is quite complicated and it depends on different parameters. It's definition is explicitly given during the proof of the theorem.

$\boldsymbol{\omega}_0 \cdot \tilde{\boldsymbol{I}} + \mathcal{O}(||\tilde{\boldsymbol{I}}||^2)$, a solution for Hamilton's equations

$$\begin{cases} \dot{\tilde{\boldsymbol{\varphi}}} = \dfrac{\partial \mathcal{K}}{\partial \tilde{\boldsymbol{I}}} = \boldsymbol{\omega}_0 + \mathcal{O}(||\tilde{\boldsymbol{I}}||) \\ \dot{\tilde{\boldsymbol{I}}} = -\dfrac{\partial \mathcal{K}}{\partial \tilde{\boldsymbol{\varphi}}} = \mathcal{O}(||\tilde{\boldsymbol{I}}||^2) \end{cases}$$

is given by $t \to (\tilde{\boldsymbol{I}}(t) = \boldsymbol{0}, \tilde{\boldsymbol{\varphi}}(t) = \tilde{\boldsymbol{\varphi}}(0) + \boldsymbol{\omega}_0 t)$.

Despite the fact that this formulation of KAM theorem gives the existence of a single invariant torus, it can be extended, ensuring the existence, for the perturbed Hamiltonian, of a set of invariant tori of large measure. In particular, remembering that if $\tau > n-1$ almost all the n-dimentional vectors $\boldsymbol{\omega} \in \mathbb{R}^n$ belong to the set $\mathcal{D}_\gamma = \cup_{\gamma>0}\{\boldsymbol{\omega} \in \mathbb{R}^n : |\boldsymbol{k}\cdot\boldsymbol{\omega}| \geq \gamma|\boldsymbol{k}|^{-\tau} \, \forall \boldsymbol{k} \neq \boldsymbol{0}\}$, i.e., they are diophantine, and that the so-called action-frequency map $\tilde{\boldsymbol{I}} \mapsto \boldsymbol{\omega_0}(\tilde{\boldsymbol{I}})$ is a local bijection (from the non-degeneracy of h), it is possible to prove the following:

Corollary 4.2 (KAM (according to Arnold)) *Consider a Hamiltonian $\mathcal{H} : \mathbb{T}^n \times \mathcal{G} \longrightarrow \mathbb{R}$ (where $\mathcal{G} \subseteq \mathbb{R}^n$ open) of the form $\mathcal{H}(\boldsymbol{\varphi}, \boldsymbol{I}) = h(\boldsymbol{I}) + \varepsilon f(\boldsymbol{\varphi}, \boldsymbol{I})$. Assume the hypotheses (2)-(4) of the previous Theorem 4.1. Then there is a set $\mathcal{S}_\varepsilon$ that is made by invariant tori and s.t. its Lebesque measure $\mu(\mathcal{S}_\varepsilon) > 0$. Moreover, $\lim_{\varepsilon \to 0} \mu\left((\mathcal{G} \times \mathbb{T}^n) \setminus \mathcal{S}_\varepsilon\right) = 0$.*

This means that a set of large measure in the phase space $\mathcal{G} \times \mathbb{T}^n$ is filled by invariant tori hosting quasi-periodic motions. In fact, the measure of the set increases as ε decreases. For further details see [5], [7] (and the references theirein) and [1].
The KAM theorem is an example of *convergent normalization procedure*; in particular, we pass from a pertubed Hamiltonian, of which we do not control the dynamics, to a 'normal form' (the Kolmogorov one), of which we can characterize a particular solution. This is the goal of Hamiltonian perturbation theory and the spirit to do it is described in the next subsection.

4.1 Normalization procedure

Let us start from a Hamiltonian of the form

$$\mathcal{H}(\boldsymbol{\varphi}, \boldsymbol{I}) = Z_0(\boldsymbol{\varphi}, \boldsymbol{I}) + \varepsilon f(\boldsymbol{\varphi}, \boldsymbol{I}),$$

where Z_0, called the *normal-form* term, is the term of which we have a control of the dynamics (for example, it can be integrable $Z_0(\boldsymbol{\varphi}, \boldsymbol{I}) = Z_0(\boldsymbol{I})$ (and we know that the solutions are linear on invariant tori), or it can be in Kolmogorov normal form (and we know a particular solution that is the torus $\{(\boldsymbol{\varphi}, \boldsymbol{I}) : \boldsymbol{I}(t) = \boldsymbol{0}, \boldsymbol{\varphi}(t) = \boldsymbol{\varphi}(0) + \boldsymbol{\omega} t)\}$), and so on). Now, applying a normalization procedure means to apply a sequence of r canonical transformations

$$\tilde{\mathcal{C}}^{(r)} : (\boldsymbol{\varphi}, \boldsymbol{I}) := (\boldsymbol{\varphi}^{(0)}, \boldsymbol{I}^{(0)}) \overset{\mathcal{C}^{(1)}}{\to} (\boldsymbol{\varphi}^{(1)}, \boldsymbol{I}^{(1)}) \overset{\mathcal{C}^{(2)}}{\to} \ldots\ldots \overset{\mathcal{C}^{(r)}}{\to} (\boldsymbol{\varphi}^{(r)}, \boldsymbol{I}^{(r)}),$$

such that, after r steps we arrive at new coordinates $(\boldsymbol{\varphi}^{(r)}, \boldsymbol{I}^{(r)})$ and at a 'new' Hamiltonian $\mathcal{H}^{(r)}$

$$\mathcal{H}^{(r)}(\boldsymbol{\varphi}^{(r)}, \boldsymbol{I}^{(r)}) = \mathcal{H}\left(\boldsymbol{\varphi}(\boldsymbol{\varphi}^{(r)}, \boldsymbol{I}^{(r)}), \boldsymbol{I}(\boldsymbol{\varphi}^{(r)}, \boldsymbol{I}^{(r)})\right) = \mathcal{H}\left(\tilde{\mathcal{C}}^{(r)^{-1}}(\boldsymbol{\varphi}^{(r)}, \boldsymbol{I}^{(r)})\right)$$

of the form

$$\mathcal{H}^{(r)}(\boldsymbol{\varphi}^{(r)}, \boldsymbol{I}^{(r)}) = \underbrace{Z^{(r)}(\boldsymbol{\varphi}^{(r)}, \boldsymbol{I}^{(r)})}_{\textit{Normal form term}} + \underbrace{R^{(r)}(\boldsymbol{\varphi}^{(r)}, \boldsymbol{I}^{(r)})}_{\textit{Remainder}}.$$

As before, $Z^{(r)}(\boldsymbol{\varphi}^{(r)}, \boldsymbol{I}^{(r)})$ is the term of which we control the dynamics, while $R^{(r)}(\boldsymbol{\varphi}^{(r)}, \boldsymbol{I}^{(r)})$ is the remainder, and tell us how the real dynamics differs from the one of $Z^{(r)}$.
The normalization procedure can be of different types:[4]

- If $\lim_{r\to\infty} ||R^{(r)}|| = 0 \Rightarrow$ we have a **convergent** normalization procedure (e.g. KAM);
- If $\lim_{r\mapsto\infty} ||R^{(r)}|| \neq 0$ and $\exists\, r_{ott}$ (optimal r) s.t. minimize $||R^{(r)}||$ (i.e. $||R^{(r_{ott})}||$ is the min.) $\Rightarrow$ we have an **asymptotic** normalization procedure (e.g. Birkhoff normal form).

By these two types of methods we can establish various results concerning the stability of the orbits. For example, in the case of KAM theorem, we can take the initial conditions (for the actions) in a Cantor set and we can establish 'perpetual' stability, in the sense that orbits with initial conditions on the torus remain always on that torus. Instead, in the case of the *Birkhoff normal form* (see [2]), the initial values for the initial actions are taken in an open set and we can state only an asymptotic result of stability, such as:

Theorem 4.3 (Nekhoroshev) *Let $\mathcal{H}(\boldsymbol{\varphi}, \boldsymbol{I}) = h(\boldsymbol{I}) + \varepsilon f(\boldsymbol{\varphi}, \boldsymbol{I})$ analytic on the domain $\mathcal{G} \times \mathbb{R}^n$ where $\mathcal{G} \subset \mathbb{R}^n$ open and s.t. the imperturbed part $h(\boldsymbol{I})$ is convex[5], i.e.*

$$|\, \langle C(\boldsymbol{I})\, \boldsymbol{v}, \boldsymbol{v} \rangle \,| \geq m||\boldsymbol{v}||^2 \quad \forall\, \boldsymbol{v} \in \mathbb{R}^n, \quad C_{jk} = \frac{\partial^2 h}{\partial \varphi_j \partial \varphi_k}\,.$$

Then, for ε sufficiently small, it holds the following: for every orbit with initial value $(\boldsymbol{I}_0, \boldsymbol{\varphi}_0) \in \mathcal{G} \times \mathbb{T}^n$, one has

$$|\boldsymbol{I}(t) - \boldsymbol{I}_0| \leq \varepsilon^b \quad \forall\, t \quad \text{s.t.} \quad |t| \leq T(\varepsilon) \sim \exp(1/\varepsilon^a)\,,$$

for suitable positive values of $a < 1$ and $b < 1$.

This means that the time scale in which the stability of the actions in guaranteed in not 'infinite', but exponentially long in the inverse of the small parameter ε. However, this type of result is good and meaningful for the applications.
Now, having in mind which is the spirit of normalization procedures, we will present an example, seeing, in practise, how it is possible to compute normal forms that lead the Hamiltonian to the desired form.

[4] The norm used to compute $||R^{(r)}||$ is a suitable norm depending on the domain in which we define the normalization method. For example, in the KAM Theorem, the domains are a complexification of $\mathcal{G} \times \mathbb{T}^n$ and the used norms are the so called *weighted Fourier norms*. See, for example, [2].

[5] There is the possibility to substitute the convexity condition with other conditions, as the quasi-convexity, or the 'steepness' condition. For further details see [12].

4.1.1 Example of normalization procedure: one step of the KAM algorithm

Let us start from the Hamiltonian described in the KAM Theorem 4.1; we can express it as

$$\mathcal{H}^{(0)}(\boldsymbol{\varphi},\boldsymbol{I}) = \underbrace{\boldsymbol{\omega}_0\cdot\boldsymbol{I}+h(\boldsymbol{I})}_{Z_0}+\sum_{i\geq 1}\varepsilon^i f_i^{(0)}(\boldsymbol{\varphi},\boldsymbol{I})\,,\quad h(\boldsymbol{I})\in\mathcal{O}(||\boldsymbol{I}||^2)\quad\text{for}\quad \boldsymbol{I}\mapsto 0\,.$$

At the first normalization step, we are interested in acting on the perturbation terms of size ε, i.e. $\varepsilon\, f_1^{(0)}(\boldsymbol{\varphi},\boldsymbol{I})$. It is particularly useful to split this term as:

$$\varepsilon\, f_1^{(0)}(\boldsymbol{\varphi},\boldsymbol{I}) = \varepsilon\left(f_{1,0}^{(0)}(\boldsymbol{\varphi})+f_{1,1}^{(0)}(\boldsymbol{\varphi},\boldsymbol{I})+\sum_{j\geq 2} f_{1,j}^{(0)}(\boldsymbol{\varphi},\boldsymbol{I})\right),$$

where in $f_{1,j}^{(0)}$ the subindex 1 is the degree of the correspondent ε, j represents the degree of $\boldsymbol{I}$ and (0) is the step of the algorithm. This allows to see the dependence of the perturbed term on the action-angle variables. With this notation, it is easy to understand that, at the end of the first normalization step, we would like to remain only with the quadratic terms in the actions of order ε. This means that our goal is to delete $f_{1,0}^{(0)}(\boldsymbol{\varphi})$ and $f_{1,1}^{(0)}(\boldsymbol{\varphi},\boldsymbol{I})$. The first normalization step is divided in two substeps.
I substep. We find the generating function $\chi_1^{(1)}(\boldsymbol{\varphi}) = X^{(1)}(\boldsymbol{\varphi})+\boldsymbol{\xi}^{(1)}\cdot\boldsymbol{\varphi}$. The part $X^{(1)}(\boldsymbol{\varphi})$ is given by the homological equation[6]

$$\{\boldsymbol{\omega}_0\cdot\boldsymbol{I}, X^{(1)}(\boldsymbol{\varphi})\}+f_{1,0}^{(0)}(\boldsymbol{\varphi}) = \left\langle f_{1,0}^{(0)}(\boldsymbol{\varphi})\right\rangle_{\boldsymbol{\varphi}}; \tag{4}$$

this allows to eliminate the term $f_{1,0}^{(0)}(\boldsymbol{\varphi})$. Instead, the generating function $\boldsymbol{\xi}^{(1)}\cdot\boldsymbol{\varphi}$ is required to keep fixed the initial frequency $\boldsymbol{\omega}_0$; i.e. it allows to arrive at a Kolmogorov normal form with the same frequency $\boldsymbol{\omega}_0$ of the initial problem. It is possible to find $\boldsymbol{\xi}^{(1)}\cdot\boldsymbol{\varphi}$ from the equation $\left\langle f_{1,1}^{(0)}(\boldsymbol{\varphi},\boldsymbol{I})+L_{\varepsilon\chi_1^{(1)}}h(\boldsymbol{I})\right\rangle_{\boldsymbol{\varphi}} = 0$.[7] At the end of the first substep we determine 'new intermediate coordinates' $(\widehat{\boldsymbol{\varphi}},\widehat{\boldsymbol{I}})$ s.t. $\boldsymbol{\varphi} = \exp\left(L_{\varepsilon\chi_1^{(1)}}\widehat{\boldsymbol{\varphi}}\right)$ and $\boldsymbol{I} = \exp\left(L_{\varepsilon\chi_1^{(1)}}\widehat{\boldsymbol{I}}\right)$; thus,

[6] It is easy to solve this homological equation expanding $f_{1,0}^{(0)}(\boldsymbol{\varphi})$ and $X(\boldsymbol{\varphi})$ in Fourier series; in fact, given $f_{1,0}^{(0)}(\boldsymbol{\varphi}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^n} c_{\boldsymbol{k}}^{(0)} e^{i\boldsymbol{k}\cdot\boldsymbol{\varphi}}$ and $X^{(1)}(\boldsymbol{\varphi}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^n}\alpha_{\boldsymbol{k}}^{(1)} e^{i\boldsymbol{k}\cdot\boldsymbol{\varphi}}$ (where $\alpha_{\boldsymbol{k}}^{(1)}$ are unknown), from the homological equation (4) we find $X^{(1)}(\boldsymbol{\varphi}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^n\setminus\{\boldsymbol{0}\}}\frac{c_{\boldsymbol{k}}^{(0)}}{i\,\boldsymbol{k}\cdot\boldsymbol{\omega}_0}e^{i\boldsymbol{k}\cdot\boldsymbol{\varphi}}$. Observe also that the diophantine condition ensures that the denominator cannot be equal to zero.

[7] It is important to observe that, in order to solve this equation, it is necessary to assume the non degeneracy of $h(\boldsymbol{I})$ required in the KAM theorem 4.1. It is possible to substitute that hypothesis with weaker hypothesis, but we have to renounce to take the frequency fixed. For example, in the 'isochronous' case, i.e. when $h(\boldsymbol{I}) = 0$, it is possible to perform a 'Kolmogorov normal form' but we have to 'detune' the frequencies; this means that the lynear flow on the torus is with frequency $\boldsymbol{\omega}\neq\boldsymbol{\omega}_0$. For further details see [9].

by the Exchange Theorem 2.6, it is possible to compute the 'intermediate Hamiltonian' as

$$\begin{aligned}\widehat{\mathcal{H}}^{(1)}(\widehat{\boldsymbol{\varphi}},\widehat{\boldsymbol{I}}) &= \mathcal{H}^{(0)}(\boldsymbol{\varphi},\boldsymbol{I})\big|_{\substack{\boldsymbol{\varphi}=\exp(L_{\varepsilon\chi_1^{(1)}})\widehat{\boldsymbol{\varphi}}\\ \boldsymbol{I}=\exp(L_{\varepsilon\chi_1^{(1)}})\widehat{\boldsymbol{I}}}} = \exp L_{\varepsilon\chi_1^{(1)}}\mathcal{H}^{(0)}\big|_{\substack{\boldsymbol{\varphi}=\widehat{\boldsymbol{\varphi}}\\ \boldsymbol{I}=\widehat{\boldsymbol{I}}}}\\ &= \boldsymbol{\omega}_0\cdot\boldsymbol{I} + h(\boldsymbol{I}) + \sum_{i\geq 1}\varepsilon^i\,\widehat{f}_i^{(1)}(\boldsymbol{\varphi},\boldsymbol{I})\,,\end{aligned}$$

with (as before)

$$\varepsilon\,\widehat{f}_1^{(1)}(\boldsymbol{\varphi},\boldsymbol{I}) = \varepsilon\left(\underbrace{\widehat{f}_{1,0}^{(1)}(\boldsymbol{\varphi})}_{=0} + \underbrace{\widehat{f}_{1,1}^{(1)}(\boldsymbol{\varphi},\boldsymbol{I})}_{\langle\cdot\rangle_{\boldsymbol{\varphi}}=0} + \sum_{j\geq 2}\widehat{f}_{1,j}^{(1)}(\boldsymbol{\varphi},\boldsymbol{I})\right).$$

II substep. We compute the generating function $\chi_2^{(1)}(\boldsymbol{\varphi},\boldsymbol{I})$ through the following homological equation

$$\{\boldsymbol{\omega}_0\cdot\boldsymbol{I},\chi_2^{(1)}(\boldsymbol{\varphi},\boldsymbol{I})\} + \widehat{f}_{1,1}^{(1)}(\boldsymbol{\varphi},\boldsymbol{I}) = 0\,;$$

this allows to completely eliminate the terms linear in the action at order ε. Thus, after the first normalization step, we arrive at a 'new' Hamiltonian given by (apart from a constant)

$$\begin{aligned}\mathcal{H}^{(1)} &= \exp L_{\varepsilon\chi_2^{(1)}}\widehat{\mathcal{H}}^{(1)} = \exp L_{\varepsilon\chi_2^{(1)}}\exp L_{\varepsilon\chi_1^{(1)}}\mathcal{H}^{(0)}\\ &= \boldsymbol{\omega}_0\cdot\boldsymbol{I} + h(\boldsymbol{I}) + \underbrace{\varepsilon\,f_1^{(1)}}_{\in\mathcal{O}(||\boldsymbol{I}||^2)} + \sum_{i\geq 2}\varepsilon^i\,f_i^{(1)}(\boldsymbol{\varphi},\boldsymbol{I})\end{aligned}$$

concluding that, up to order ε, the Hamiltonian is in Kolmogorov normal form. The algorithm can procede iteratively, repeating it r times and proving the convergence in the limit $r\to+\infty$. For more details on these arguments an a more detailed exposition, see [2] and [1].

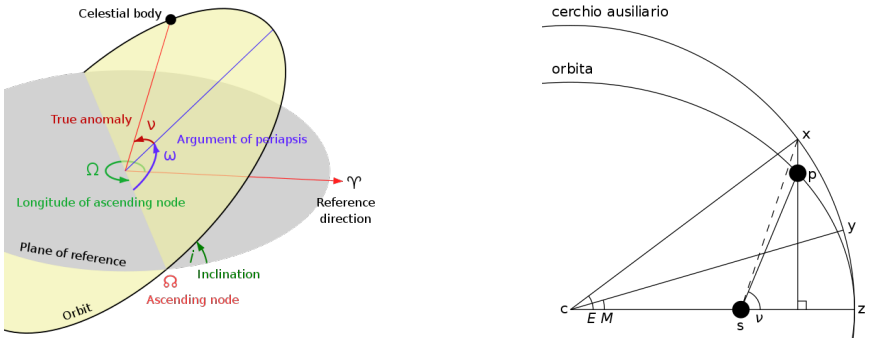
5 Applications

In this section we are interested in applications of the theory presented above in celestial mechanics; we want to study the perturbed Hamiltonian that describes a three body problem of a multi-planetary mutually inclined system. In particular we have to deal with exoplanetary systems; these systems are very different from our Solar system, having masses, eccentricities and mutual inclinations bigger than those of our own solar system.

The Hamiltonian of the three-body problem in Poincaré heliocentric canonical variables takes the form:

$$\mathcal{H} = \underbrace{\frac{\boldsymbol{p_2}^2}{2\,m_2} - \frac{\mathcal{G}\,m_0\,m_2}{r_2} + \frac{\boldsymbol{p_3}^2}{2\,m_3} - \frac{\mathcal{G}\,m_0\,m_3}{r_3}}_{\textbf{\textit{Keplerian part}}} + \underbrace{\frac{(\boldsymbol{p_2}+\boldsymbol{p_3})^2}{2\,m_0}}_{\textbf{\textit{"Indirect" part}}} - \underbrace{\frac{\mathcal{G}\,m_2\,m_3}{|\boldsymbol{r_2}-\boldsymbol{r_3}|}}_{\textbf{\textit{"Direct" part}}}\,, \tag{5}$$

where m_0 = mass of the star, m_i , $\boldsymbol{p_i}$, $\boldsymbol{r_i}$, $i = 2, 3$ are the masses, barycentric momenta and heliocentric position vectors of the planets and $\mathcal{G}$ is the gravitational constant.



(a) Orbital elements.

(b) Anomalies. M is related to E through the Kepler law $M = E - e\sin(E)$.

Figure 2: In order to locate a celestial body in the space we use the 'orbital parameters' (on the left): e = eccentricity of the orbit, a = semimajor axes, ω = argument of periapsis, ν = true anomaly, Ω = longitude of the ascending node, i = inclination. Moreover we use also the angles (on the right) E = eccentric anomaly and the 'virtual angle' M = mean anomaly. For further details on this topic, see [10].

Starting from (5), it is possible to express the Hamiltonian in the so called *orbital parameters* (see Figure 2); then, a secular Hamiltonian is arrived at by averaging the above Hamiltonian with respect to all short period terms, i.e.

$$\mathcal{H}_{sec} = <\mathcal{H}>_{\lambda_2,\lambda_3} = -\frac{\mathcal{G}m_0m_2}{2a_2} - \frac{\mathcal{G}m_0m_3}{2a_3} + \mathcal{R}_{sec}(m_0, m_j, a_j, e_j, i_j, \omega_j, \Omega_j)\,,$$

where $\lambda_j = M_j + \omega_j + \Omega_j$, $j = 2, 3$ are called the 'fast' angles. In fact, in astronomy, there exist different time scales for the evolutions of the orbital parameters; the angle λ (related to the revolution of the planet around the star) is 'fast' with respect to the 'secular' time required from other quantities (such as the eccentricities, inclinations, etc.) to vary their behaviour. Moreover, it can be proved that the total angular momentum vector $\boldsymbol{L} = \boldsymbol{r_2} \times \boldsymbol{p_2} + \boldsymbol{r_3} \times \boldsymbol{p_3}$ is an exact first integral of the Hamiltonian (5), a fact implying that the dependence of the Hamiltonian on the angles Ω_2, Ω_3 is only through the difference $\Omega_2 - \Omega_3$. However, the existence of two independent integrals in involution (i.e. the components L_z and L_{plane} of the total angular momentum $\boldsymbol{L}$) allows to reduce the number of degrees of freedom by two, a process known as "Jacobi's reduction of the nodes" (see [4]). In particular it is possible to pass to the so called *Laplace reference frame*: it is an invariant reference frame, orthogonal to the total angular momentum vector $\boldsymbol{L}$ where we have the invariance $\Omega_3 - \Omega_2 = \pi$. In this reference frame we can express the secular Hamiltonian in the following form (see [8]):

$$\begin{aligned}(6)\qquad \mathcal{H}_{sec} = \mathcal{H}_{plane}(m_0, m_2, m_3, a_2, a_3, e_2, e_3, \omega_2 - \omega_3) \\ + \varepsilon\,\mathcal{H}_{space}(m_0, m_2, m_3, a_2, a_3, e_2, e_3, \omega_2, \omega_3; \mathrm{AMD})\,,\end{aligned}$$

with AMD the 'angular momentum deficit', defined as

$$\text{AMD} = \sum_{j=2}^{3} m_j \sqrt{\mathcal{G} m_0 a_j} \left(1 - \sqrt{1-e_j^2}\cos(i_j)\right). \tag{7}$$

It can be also expressed in the canonical Poincaré variables (positions-momenta)

$$X_j = -\sqrt{2\Gamma_j}\cos(\omega_j), \qquad Y_j = -\sqrt{2\Gamma_j}\sin(\omega_j), \tag{8}$$

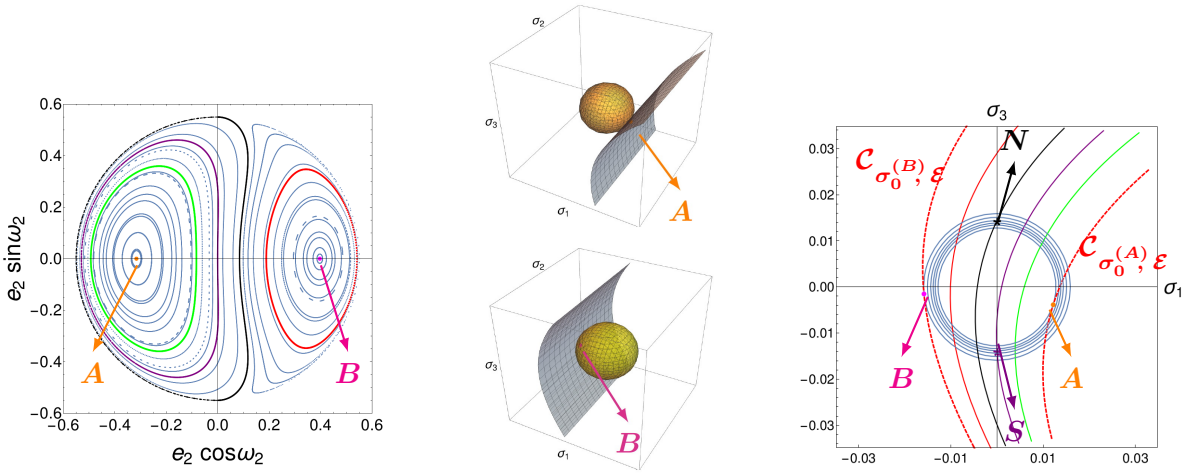
with $\Lambda_j = m_j\sqrt{\mathcal{G} m_0 a_j}$, $\Gamma_j = \Lambda_j\left(1-\sqrt{1-e_j^2}\right)$, $j = 2, 3$ (modified Delaunay variables).

5.1 Integrability: application to the planar case

If we restrict ourselves to the planar part of the Hamiltonian (6)

$$\mathcal{H}_{plane} = \mathcal{H}_{plane}(\Gamma_2, \Gamma_3, \omega_2 - \omega_3) = \mathcal{H}_{plane}(X_2, X_3, Y_2, Y_3)$$

we are in an integrable case. This is a 2 degrees of freedom Hamiltonian with 2 independent first integrals in involution, that are $\mathcal{H}_{plane}$, that represents the energy, and the angular momentum $||\boldsymbol{L}|| = L_z = \Lambda_2 + \Lambda_3 - \Gamma_2 - \Gamma_3$. Thus, by the Liouville-Arnold-Jost Theorem 3.2, the dynamics takes places on 2-dimensional tori and, depending on the frequencies, we can have periodic or quasi-periodic orbits. Moreover, remembering the observation done in section 3, we know that in a Poincaré section Σ the linear flow on a torus yields in a finite number of points if the orbit is periodic, or a closed curve otherwise. An example of Poincaré section for $\mathcal{H}_{plane}$ (with section $Y_3 = 0$, direction $\dot{Y}_3 \geq 0$ and for a fixed level of energy $\mathcal{E}$) is reported in Figure 3a.



(a) (b) (c)

Figure 3: Example of two equilibria computed through the tangency method fixed $\mathcal{E} \sim 6.6 \cdot 10^{-5}$. Figure 3a: Poincaré section in the representative plane $(e_2\cos(\omega_2), e_2\sin(\omega_2))$ (considering only $\mathcal{H}_{int}$). Figure 3b: tangencies between the sphere and the energy surface. Figure 3c: projection of the 3D-surfaces on the plane $\sigma_2 = 0$ for different values of the radius σ_0.

Concerning this picture, the two points, called **A** and **B** represent periodic orbits. They are called the *anti-aligned* and *aligned apsidal corotations* and they have a particular phisical meaning; they correpond, respectively, to the orbital configurations with the anti-alignment and aligment of the pericenters of the two planets. However, the Poincaré section creates the impression that there is a sort of 'separation' between the left part of the picture (that surround **A**) and the right one (that surround **B**). Actually, there are no separatrices between $\boldsymbol{A}$ and $\boldsymbol{B}$ and the singularities are due only to the choice of variables. In fact the system's reduced phase space has the topology of the 3D-sphere and it can be seen used the following variables:

$$\sigma_0 = \frac{1}{2}(X_2^2 + Y_2^2 + X_3^2 + Y_3^2)\,, \qquad \sigma_1 = X_2X_3 + Y_2Y_3\,,$$
$$\sigma_2 = Y_2X_3 - Y_3X_2\,, \qquad \sigma_3 = \frac{1}{2}(X_2^2 + Y_2^2 - X_3^2 - Y_3^2)\,,$$

where X_j, Y_j, $j = 2, 3$ are the canonical Poincaré variables described in (8). In particular, we can prove the following:

Lemma 5.1 (Poisson algebra relation) $\{\sigma_i, \sigma_j\} = -2\epsilon_{ijk}\sigma_k$ *, where* ϵ_{ijk} *is the Levi-Civita symbol, with* $i, j, k = 1, 2, 3$ *and* $\{\sigma_i, \sigma_j\} = 0$ *if* i *or* j *is* 0 *.*

Moreover, they satisfy the condition of sphere:

$$\sigma_1^2 + \sigma_2^2 + \sigma_3^2 = \sigma_0^2\,.$$

Thus, it is possible to define the following surfaces:

$$\mathcal{S}_{\sigma_0} = \{(\sigma_1, \sigma_2, \sigma_3) \in \mathbb{R}^3 \,:\, \sigma_1^2 + \sigma_2^2 + \sigma_3^2 = \sigma_0^2\}\,,$$
$$\mathcal{C}_{\sigma_0,\mathcal{E}} = \{(\sigma_1, \sigma_2, \sigma_3) \in \mathbb{R}^3 \,:\, \mathcal{H}_{int}(\sigma_0, \sigma_1, \sigma_3) = \mathcal{E}\}$$

and study, varying the values of the radius σ_0 (at a fixed level of energy $\mathcal{E}$), the intersections between the two surfaces.[(8)] Thus, the tangency points represent periodic orbits, while the curves of intersection represent quasi-periodic orbits. This is shown in Figure 3c, illustrating the projection of the two surfaces $\mathcal{S}_{\sigma_0}$ and $\mathcal{C}_{\sigma_0,\mathcal{E}}$ on the plane $\sigma_2 = 0$. It is possible to move, with continuity, from the smallest to the greatest value of the radius, called respectively $\sigma_0^{(A)}$ and $\sigma_0^{(B)}$; for such values of σ_0 the two surfaces are tangent and the tangency points correspond to the periodic orbits in the Poincaré section (in Figure 3a). Instead, increasing σ_0 from $\sigma_0^{(A)}$ to $\sigma_0^{(B)}$, the two surfaces intersect each other in closed curves. Finally, the closed curve passing through a singularity (indicated as N, that physically indicates $e_3 = 0$) corresponds, in the Poincaré section, to the black curve (where N corresponds to the dotted one).

[(8)] It is possible to prove that the integrable Hamiltonian (thus, the planar part in our case), does not depend on σ_2 ; being a cylindric simmetry, the study of the intersections is reduced to the projection in the plane $\sigma_2 = 0$.

5.2 Nearly-Integrability: application to the spatial case

If we now consider the complete Hamiltonian (6)

$$\mathcal{H}_{sec} = \mathcal{H}_{plane}(m_0, m_2, m_3, a_2, a_3, e_2, e_3, \omega_2 - \omega_3) + \varepsilon\, \mathcal{H}_{space}(m_0, m_2, m_3, a_2, a_3, e_2, e_3, \omega_2, \omega_3; \mathrm{AMD})\,,$$

we have no longer an integrable Hamiltonian. In this case the perturbation is related to the mutual inclination; the more mutually-inclined is the system, the further we are from the integrability of the problem (i.e. in this case the planar part).

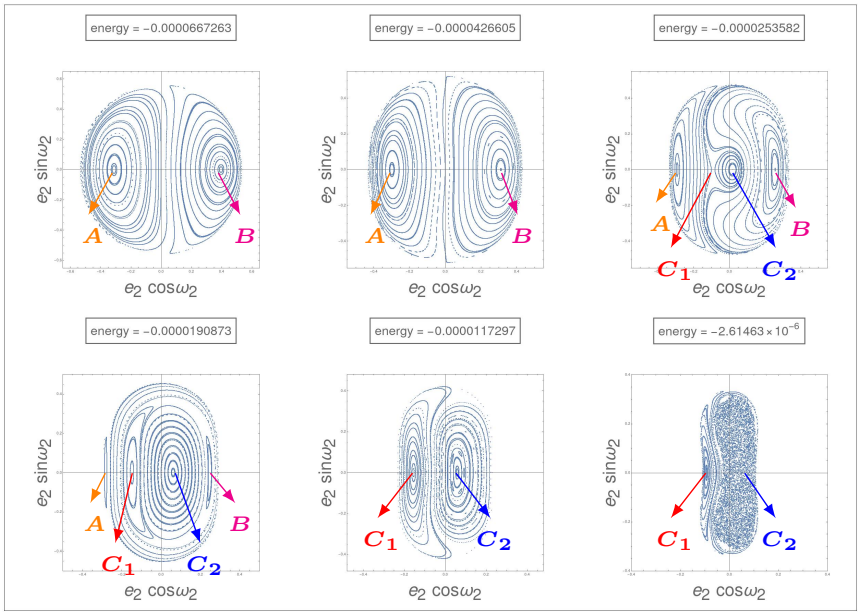


Figure 4: Poincaré surfaces of section in the representative plane $(e_2\cos(\omega_2), e_2\sin(\omega_2))$ with AMD fixed and for different values of the energy.

This is evident from the analysis of the Poincaré section of the complete Hamiltonian, shown in Figure 4; these pictures have been done changing the value of e_2 and e_3 (that means to change the energy $\mathcal{E}$ of the system) while keeping fixed the value of the AMD. Since the AMD is fixed, altering the values of the planets' eccentricities implies that the inclinations also change to keep the AMD constant to its pre-selected value (see eq. (7)). From the first two pictures we observe a similar behaviour to the planar case (see Figure 3a), characterized from the two apsidal corotations; we call it nearly-planar regime. As the energy increases (from left to right), the maximum allowed mutual inclination between the planets also increases; thus, we recognise a sequence of bifurcations. In particular a saddle-mode bifurcation generates the orbits C_1, C_2 which correspond to an orbital configuration with non-zero mutual inclination. Furthermore, as the mutual inclination increases, the orbit C_2 becomes unstable by the "Kozai mechanism" (see [6]), as shown in

the last picture. Thus, we can observe that the bigger is the mutual inclination (and then the size of the perturbation), the more chaotic the motions became. On the other hand, the computations of some periodic orbits (around the apsidal corotations) are still possible through some normalization procedure, like the Birkhoff normal form.

References

[1] G. Benettin, "Appunti per il corso di Meccanica Analitica". Università di Padova, `https://www.math.unipd.it/∼benettin/links-MA/ma-17_10_25.pdf`, 2014.

[2] A. Giorgilli, "Notes on Hamiltonian Dynamical Systems". Cambridge University Press 102, 2022.

[3] A. Giorgilli, "Metodi e Modelli Matematici per le Applicazioni, Cap. 6". Available at `http://www.mat.unimi.it/users/antonio/metmod/Note_6.pdf`.

[4] M. Jacobi, "Sur l'élimination des noeuds dans le problème des trois corps. Par M. Jacobi". Astronomische Nachrichten, 20, 81 (1842).

[5] A.N. Kolmogorov, *Preservation of conditionally periodic movements with small change in the Hamilton function.* Dokl. Akad. Nauk SSSR 98, 527–530 (1954). Engl.transl. in: Los Alamos Scientific Laboratory translation LA-TR-71-67; reprinted in: Lecture Notes in Physics, 93, 51–56, Springer (1979).

[6] Y. Kozai, *Secular perturbations of asteroids with high inclination and eccentricity.* The Astronomical Journal, 67, 591–598 (1962).

[7] U. Locatelli, C. Caracciolo, M. Sansottera, M. Volpi, *Invariant KAM tori: from theory to applications to exoplanetary systems.* G. Baù, S. Di Ruzza, R.I. Páez, T. Penati, M. Sansottera (eds.), I-CELMECH Training School - New frontiers of Celestial Mechanics: theory and applications, Springer PROMS (in press) (or `arXiv:2202.06572 2022`).

[8] R. Mastroianni, C. Efthymiopoulos, *Secular dynamics in extrasolar systems with two planets in mutually inclined orbits.* Proceedings of the International Astronomical Union, 15(S364), 191–196 (2021). `https://www.doi.org/10.1017/S1743921321001368`.

[9] R. Mastroianni, C. Efthymiopoulos, *Kolmogorov algorithm for isochronous Hamiltonian systems.* Mathematics in Engineering, 5 (2): 1–35 (2023). DOI: `10.3934/mine.2023035`.

[10] C. Murray, F. Stanley, "Solar system dynamics". Cambridge University Press, 1999.

[11] H. Poincaré, *Les méthodes nouvelles de la mécanique céleste.* Gauthier-Villars, Paris, 1892.

[12] J. Pöschel, *Nekhoroshev estimates for quasi-convex Hamiltonian systems.* Math. Z., 213 (2): 187–216, 1993.

Chaotic dynamical systems and applications to the Solar System dynamics

Mattia Rossi (*)

Abstract. Dynamical systems are an essential tool to model physical phenomena in applied sciences whose state changes over time according to either differential or discrete difference equations. In this context two concepts are in opposition: "order", or "integrability", versus "chaos". Integrable systems, for which all the solutions can be explicitly analytically determined, are special and represent only a crude approximation of the real dynamics. On the other hand, more accurate models are usually represented by non-integrable differential equations, whose solutions exhibit a highly sensitive dependence on initial conditions, termed as chaotic. In these notes we discuss some of the main geometric and topological properties of deterministic chaos in connection with orbital stability of small objects in our solar system. After a short recap of the theory of non-linear dynamical systems, we present a modern approach of detecting and quantifying chaotic behaviours using finite time chaos indicators, a numerical strategy capable to capture the dynamical structure of the phase space. In the last part, we introduce the restricted N-body problem in Hamiltonian mechanics and implement the above technique to discriminate between the realms of regular and chaotic motions of asteroids.

1 Non-linear dynamical systems

1.1 General definitions

We start by recalling some basic definitions of the theory of dynamical systems.

Definition 1 A *(smooth) dynamical system* is a triple (G, M, Φ) s.t.:

- M is a smooth manifold called *phase space*;
- $G = \mathbb{R}$ (continuous time) or $G = \mathbb{Z}$ (discrete time);
- Φ is a free differentiable action of G on M.

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: mattia.rossi@math.unipd.it . Seminar held on 25 May 2022.

Specifically, in the continuous case a dynamical system is represented by an ordinary differential equation of the form

$$\dot{x} = X(x) \,, \quad x \in M \,, \tag{1}$$

where $X : M \to TM$ is a differentiable vector field; the map $\Phi : \mathbb{R} \times M \to M$ is called the *flow* of X, such that $\Phi(t, x_0)$ symbolizes the value at current time t of the solution which at initial time $t = 0$ is equal to x_0.
On the other hand, it is also possible to consider phenomena evolving in discrete time, whose associated mathematical law, that is a function, is iteratively repeated. The resulting dynamical system is thus obtained by the relationship

$$x_n = \Psi^n(x_0) \,, \quad x_0, x_n \in M \,, \quad n \in \mathbb{Z} \,, \tag{2}$$

where $\Psi : M \to M$ is a diffeomorphism and

$$\Psi^n = \begin{cases} \Psi \circ \ldots \circ \Psi & n \text{ times}, \ n > 0 \\ \Psi^{-1} \circ \ldots \circ \Psi^{-1} & |n| \text{ times}, \ n < 0 \end{cases} .$$

Analogously, we can define now $\Phi : \mathbb{Z} \times M \to M$ by setting $\Phi(n, x_0) \coloneqq \Psi^n(x_0)$.
Finally, we remind that Φ is a differentiable action in the sense that it satisfies the following properties:

(i) $\Phi(0, \cdot) = \mathrm{id}_M$;

(ii) $x \mapsto \Phi(t, x)$ is a diffeomorphism $\forall t \in G$;

(iii) $\Phi(t, \Phi(s, \cdot)) = \Phi(t + s, \cdot) \ \forall t, s \in G$.

For practical purposes, we shall assume from now on that $M = D \subseteq \mathbb{R}^d$ and mostly $G = \mathbb{R}$.

Definition 2 The *orbit* (or *integral curve*) of a point $x \in D$ is the set $\mathscr{O}_x = \{\Phi(t, x) : t \in \mathbb{R}\}$. The set of all orbits is called *phase portrait*.

Remark 1 Given $\Phi \in \mathcal{C}^\infty(\mathbb{R} \times D)$, Cauchy theorem about the existence and uniqueness of the solution in a neighbourhood of $t = 0$ to $\dot{x} = X(x)$, $x(0) = x_0$, holds. In addition, as already understood, we assume to extend this $\forall t \in \mathbb{R}$, so that the phase portrait forms a partition of D.

1.2 Non-integrability and chaotic systems

In general (1) is highly non-linear and the corresponding solutions cannot be provided easily. Usually it is not possible to find an analytic formula at all.
These observations about the solvability of a system are encompassed by the notion of *non-integrability*, i.e. a property of the system itself, consisting in the incapability of determining all the solutions explicitly. To better understand its implications, let us consider the opposite situation of integrability, whose meaning is twofold: of operative nature and geometric nature.

Operative $\longrightarrow$ The solutions can be expressed in terms of quadratures, that is simple arithmetic operations $(+, -, \cdot, :)$, radicals $(\sqrt[\cdot]{\cdot})$ or integration and inversion of elementary functions $(\int \cdot,\ \cdot^{-1})$.

Geometric $\longrightarrow$ There exists an embedded Φ-invariant foliation in the phase space (e.g. invariant tori in integrable Hamiltonian systems).

The second character is intimately related to the existence of conserved quantities along the flow, that act as constraints for the orbits (Remark 2).

Definition 3 A *first integral* of $\dot{x} = X(x)$ is a smooth function $f : D \to \mathbb{R}$ which is constant on all the solutions of (1), i.e. $f(\Phi(t, x)) = f(x)\ \forall t \in \mathbb{R}$ and $x \in D$.

Remark 2 Every time we provide a first integral, the problem is reduced from dimension d to dimension $d-1$. So if $\exists\ f_1, \dots, f_{d-1}$ independent first integrals, the level sets have dimension $d - (d-1) = 1$, which are the orbits of the ODE.[9]

Non-integrability is a necessary condition for the central topic in question of *chaos*.

Definition 4 We call a dynamical system **chaotic** [1] whenever

(i) it is *topologically transitive*, namely $\forall U, V \subseteq D$ open, $\exists\ t \in \mathbb{R}$ s.t. $\Phi(t, U) \cap V \neq \varnothing$;

(ii) it has a dense set of periodic orbits;

(iii) it is sensitive to initial conditions.

Banks et al. in [1] proved that (iii) is redundant, that is the first two conditions (i) and (ii) imply the third one. However, it is precisely the picture that we practically bear in mind: depending on the rate of separation $\lambda > 0$ (Lipschitz constant), given initial conditions s.t. $\|x_1 - x_0\| < \varepsilon$, we have

$$\|\Phi(t, x_1) - \Phi(t, x_0)\| \leq e^{\lambda |t|} \|x_1 - x_0\| \,, \tag{3}$$

so in principle the dynamics becomes totally unpredictable for $|t|$ large enough. Then, we customarily associate overt chaos to the worst case: *exponential separation* in time of orbits of close initial data.

2 Phase space analysis

2.1 Poincaré surface of section

An effective technique to explore the phase space of a dynamical system consists in reducing the study of the flow of the differential equation (1) to the study of the iterations of a discrete map, called *Poincaré section*, defined in a subspace of dimension $d-1$. Poincaré sections are used, in particular, to visualize the complicate topology of non-integrable

[9] Ordinary Differential Equation

systems. Basically, we consider $\Sigma^{d-1} \subseteq D$: $\forall x \in \Sigma\ \exists\ t(x)$ such that $\Phi(t(x), x) \in \Sigma$ returns "on the same side" (Fig. 1), viz. $\bigcup_x \mathscr{O}_x$ is replaced by $\bigcup_{n\in\mathbb{Z}} \Psi^n(x)$, $\Psi(x) \coloneqq \Phi(t(x), x)$.

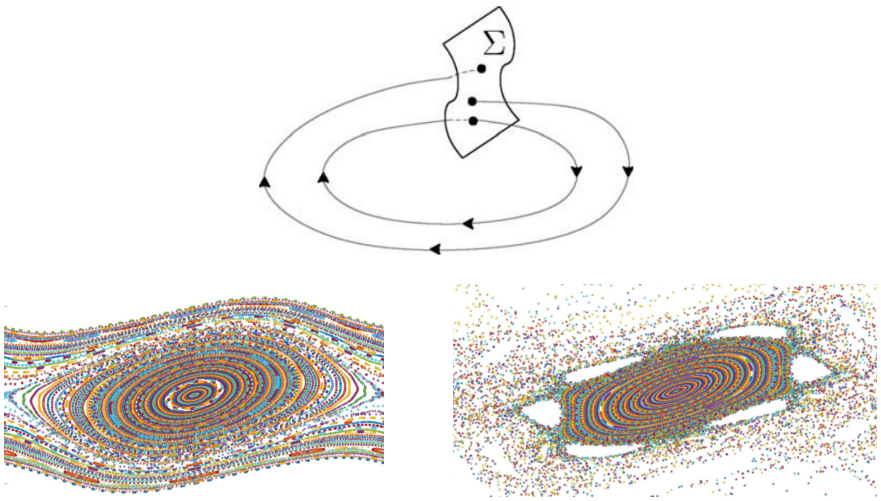


Figure 1. Exemplification of the Poincaré surface of section. **Top panel**: pictogram showing the transversal intersection of Φ with Σ. **Bottom panels**: example of numerical computation for the classic standard map [6], with Σ given by a coordinate plane, for two different values of the perturbing parameter; in the picture on the right we appreciate more chaos, represented by unevenly distributed dots, as opposed to regular motions (recognizable as libration or circulation like).

2.2 Hyperbolic dynamics

Equilibrium points are the simplest solutions of an ODE and are found as critical points of the vector field $X(x)$. Among all of them, those of hyperbolic origin exhibit the richest and most interesting local dynamics, characterized by smooth intricate structures stemming from such equilibria.

Definition 5 An *equilibrium point* for (1) (that is $x = c \in D$ s.t. $X(c) = 0$) is called *hyperbolic* if all the eigenvalues of the Jacobian matrix $J(x) = \partial X(x)/\partial x$ have non-zero real part.

As a standard approach, around equilibrium points one performs the linearization of the system in order to make the investigation easier and try to deduce relevant properties of the full dynamics or at least of the approximate one. So we look at

$$\dot{y} = J(c)y\ , \quad y = x - c\ . \tag{4}$$

This strategy turns out to be extremely powerful with hyperbolic equilibrium points thanks to the two following fundamental results.

Theorem 1 (Grobman-Hartman theorem) *Let $c \in D$ be an hyperbolic equilibrium point. Then there exists a neighbourhood U of c in D, a neighbourhood V of $0 \in \mathbb{R}^d$ and a homeomorphism $h : U \to V$ such that*

- $h(c) = 0$;
- $h(\Phi(t, c)) = e^{tJ(c)}h(c)$

for all $x \in U_0 \subseteq U$ *neighbourhood of* c *and* $t \in (-\varepsilon, \varepsilon)$.

The upshot is that the phase portraits around hyperbolic equilibrium points appear as deformations of the portraits of their linearized system. Nevertheless, h is only continuous and out of U there is no information about the global evolution. Luckily, we can benefit from the so-called *stable manifold theorem* about the existence and regularity of the sets

$$W^s = \{x \in D : \lim_{t\to+\infty} \Phi(t, x) = c\}$$
$$W^u = \{x \in D : \lim_{t\to-\infty} \Phi(t, x) = c\} \tag{5}$$

called respectively *stable* and *unstable manifold*.

Theorem 2 (Local stable manifold theorem) *Given* c *hyperbolic equilibrium point, let* E^s, E^u *be the stable and unstable spaces of the linearization (4) at* c. *Then* $\exists$ U *neighbourhood of* c *s.t.:*

(i) $W^s_{loc} \coloneqq W^s|_U$, $W^u_{loc} \coloneqq W^u|_U$ *are smooth connected submanifolds of* D *with* $T_cW^s_{loc} = E^s$, $T_cW^u_{loc} = E^u$;

(ii) $\exists\ a_s, a_u \in (0, 1)$ *and* $b > 0$ *s.t. for all* $t > 0$

$$\begin{aligned} x \in W^s_{loc},\ \xi \in T_xW^s_{loc} &\implies \|D\Phi(t, x)\xi\| \le ba_s^t\|\xi\| \\ x \in W^u_{loc},\ \xi \in T_xW^u_{loc} &\implies \|D\Phi(-t, x)\xi\| \le ba_u^{-t}\|\xi\| \end{aligned}.$$

Remark 3

(i) From the knowledge of $W^s_{\text{loc}}, W^u_{\text{loc}}$ we can recover the global W^s, W^u considering respectively $\bigcup_{t\le0} \Phi(t, W^s_{\text{loc}})$, $\bigcup_{t\ge0} \Phi(t, W^u_{\text{loc}})$.

(ii) The topology of the stable and unstable manifolds becomes even more complex when they have transverse intersections, especially *homoclinic points* (Fig. 2, middle panel) or *heteroclinic points* (Fig. 2, right panel). As outcome, this underlying web becomes the main hidden structure we can look at to give a first classification to chaotic orbits.

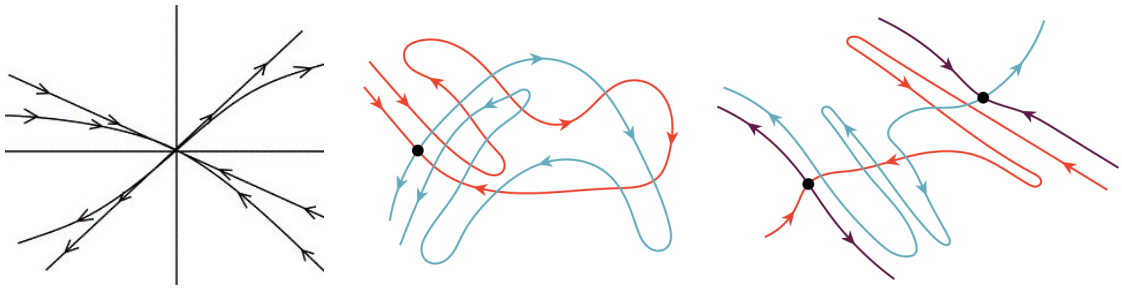


Figure 2. Dynamics originating at hyperbolic equilibria. **Left panel**: Illustration of statement (i) of Theorem 2, where the tangent straight lines stand for E^s, E^u. **Middle panel**: intersection between W^s and W^u spreading from an hyperbolic equilibrium point (homoclinic). **Right panel**: intersection between W^s and W^u emanating from two different hyperbolic equilibrium points (heteroclinic).

3 Finite time chaos indicators

3.1 Variational dynamics

We have seen so far where (§2.2) and how (§2.1) to inspect chaotic behaviours in phase space. It is time now to quantify, by means of unavoidable numerical methods, the amount of chaos, in other words the growth of exponential separation of two trajectories starting at close initial points.
For this purpose, instead of working with the system (1), we pass to the *variational dynamics*, that is we focus on the rate of change of tangent vectors along the flow. Let us introduce the *tangent map*

$$T\Phi(t,x) : T_xD \to T_{\Phi(t,x)}D \ , \quad v_0 \mapsto v_t = \frac{\partial \Phi}{\partial x}(t,x)v_0 \ ,$$

where $v_t \coloneqq v(t)$ is the time evolution of $v_0 \coloneqq v(0)$ according to the *variational equation*:

$$\dot{v} = \frac{\partial X}{\partial x}(\Phi(t,x))v \ . \tag{6}$$

From (6) we have again the simple estimate similar to (3)

$$\|v(t)\| \leq a(t)\|v(0)\| \ ,$$

with $a(t) \leq e^{\Lambda|t|}$, $\Lambda > 0$ and particularly $a(t) \sim e^{\Lambda|t|}$ for chaotic systems. The issue resides on the fact that, as before, this estimate may be very inaccurate. Nonetheless, the variational dynamics allows to do better by defining a new quantity, always well-defined, capable to capture the exact asymptotic growth of the tangent vectors.

Definition 6 Let $v(t)$ be the solution of (6) with initial conditions x_0, v_0. The *characteristic Lyapunov exponent* (CLE) of an initial condition x_0 and initial tangent vector v_0 is the limit:

$$\chi(x_0, v_0) = \lim_{t\to\infty} \frac{\ln \|v(t)\|}{t} \ . \tag{7}$$

The anticipated well-definedness of (7) descends from a classic theorem in multiplicative ergodic theory.

Theorem 3 (Oseledets theorem) *Let $\mathscr{M}$ be a probability measure on D. $\chi(x_0, v_0)$ is a real number $\forall v_0 \in T_{x_0}D$, $v_0 \neq 0$, and for $\mathscr{M}$-almost every $x_0 \in D$. Moreover:*

(i) *the CLE is a constant of motion for (6);*

(ii) *if c is an equilibrium point and $lim_{t\to\infty}\Phi(t, x_0) = c$, then, called $L(x) \coloneqq \{\chi(x, v_0) : v_0 \in T_xD\}$, $L(x_0) = L(c)$;*

(iii) *$L(x_0)$ for any x_0 is discrete, with at most d different elements and for $\mathscr{M}$-almost all $v_0 \in T_{x_0}D$, $\chi(x_0, v_0) = \max L(x_0)$.*

Remark 4 Property (iii) of Theorem 3 has important consequences for actual computations: a random choice of the initial tangent vector provides the largest characteristic Lyapunov exponent (LCLE), that we shall denote by χ_L.

3.2 Chaos indicators

About the necessity to find numerical approximations of the flow of (1), we emphasize that

- initial data of real systems are affected by errors, so we need to compute the time evolution of a set of "compatible" initial conditions;
- there are strong limitations in terms of reliability of the approximate solution when the dynamics separates exponentially the orbits.

We thereby overcome these inconveniences by making use of suitable chaos markers inspired by χ_L. Clearly, chaotic orbits are detected by $\chi_L > 0$, hence $T_L = 1/\chi_L$, called *Lyapunov time*, represents the time scale needed to observe the exponential separation: $\|v(t)\| \sim \|v(0)\| e^{t/T_L}$. This suggests to construct *finite time chaos indicators*, so they can be practically computed.

We outline in the following some of the most popular chaos indicators that can be found in the literature (for further reading cf. [5, 6]) and their application to simple models from classical and fluid mechanics (Fig. 3).

Fast Lyapunov Indicator:

$$FLI(x_0, v_0; \tau) = \max_{t\in[0,\tau]} \ln \|v(t)\| \, . \tag{8}$$

Given $e_j(t)$ time evolution of $\mathcal{B} = \{e_j(0)\}_{j=1,\dots,d}$ orthonormal basis of $\mathbb{R}^d$,

Fast Lyapunov Indicator of the Basis:

$$FLIB(x, \mathcal{B}; \tau) = \max_{t\in[0,\tau]} \max_{j=1,\dots,d} \|e_j(t)\| \, ,$$

Finite Time Lyapunov Exponent:

$$FTLE(x; \tau) = \frac{1}{\tau} \max_{e_j(0)\in\mathcal{B}} \frac{\|D\Phi(\tau, x) e_j(0)\|}{\|e_j(0)\|} \, .$$

As one should expect, these numbers are theoretically consistent, because

- $FLI \approx FLIB \approx FTLE$;
- $\chi(x_0, v_0) = \lim_{\tau\to+\infty} \dfrac{FLI(x_0, v_0; \tau)}{\tau}$.

Typically, due to their substantial equivalence, (8) is computationally preferable for its simpler formulation and for this reason adopted in the subsequent examples. Finally,

we stress that τ depends on the problem and is determined heuristically: essentially one computes FLI for various τ and retains that according to which the portrait stabilizes.

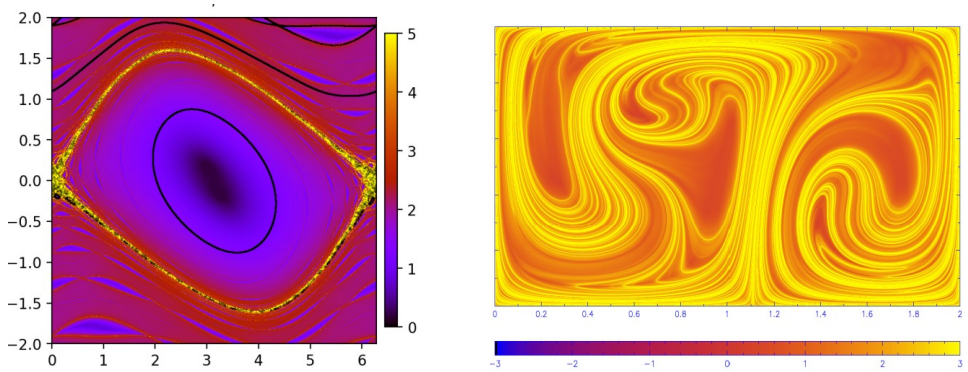


Figure 3. Colour plots from [6] of the FLI indicator in the case of the standard map (**left panel**) and the non-autonomous double gyre (**right panel**). Lighter nuances correspond to more chaotic regions, in agreement with Fig. 1 for the picture on the left. The effectiveness of the FLI map is manifested also though the capture of more sophisticated details in phase space in the image on the right, like the various folds of the time-dependent generalization of W^s, W^u (often termed as Lagrangian coherent structures).

4 Modelling the dynamics of small bodies in the Solar System

4.1 Solar System in a nutshell

Assuming a certain familiarity of the reader, at least at an elementary level, for the sake of completeness let us quickly summarize below some of the mean features of the bodies populating our solar system, especially to fix orders of magnitudes. We specify that in the following the symbol AU stands for Astronomic Unit and is the average distance of the Earth from the Sun, roughly equal to 150 million kilometres.
The Solar System is made up of

- four *inner rocky planets* + two *gas giant planets* + two *ice giant planets*, respectively:
  - Mercury (0.39 AU), Venus (0.72 AU), Earth (1.00 AU), Mars (1.52 AU),
  - Jupiter (5.2 AU), Saturn (9.6 AU),
  - Uranus (19.2 AU), Neptune (30.1 AU);
- one *asteroid belt* (called also *main belt*) located after the orbit of Mars;
- one external disk-like distribution of small objects called *Kuiper belt*, extending from about 30 AU to 50 AU and containing Pluto (≈ 40 AU);
- one outermost spherical reservoir of long-period comets called *Oort cloud*, ranging from 2000 to 100000 AU.

4.2 The Sun-Jupiter restricted three-body problem

Let us build up a simple, but sufficiently accurate, model for the dynamics of minor bodies moving under the effect of the gravitational pull of more massive ones. In this regard, let us begin with the easiest integrable system concerning just two interacting particles, say a planet m_1 and an asteroid m_2, known as *Kepler problem* (or *2-body problem*), as a basis for a successive non-integrable refinement. Denoting by $\mathcal{G}$ the gravitational constant, the equation of the relative motion reads

$$\ddot{r} = -\frac{\mathcal{G}(m_1 + m_2)}{\|r\|^3} r \,, \tag{9}$$

whose solution is a conic section expressed in polar form by $r(t) = r(a, e, i, f(t), \omega, \Omega)$,

$$\|r(t)\| = \frac{p}{1 + e\cos f(t)} \,.$$

In our case, we are interested in elliptic trajectories: the angles $i \in [0, \pi[$, $f, \omega, \Omega \in \mathbb{T}$ give the orientation of the ellipse in space and are called respectively *inclination*, *true anomaly*, *argument of pericenter* and *argument of the ascending node*; $p = a(1 - e^2)$, $a > 0$ is the *semi-major axis* and $0 \leq e < 1$ is the *eccentricity*.

Since our aim is to treat the situation $m_2 \ll m_1$, it is reasonable to assume m_2 vanishing in (9): the asteroid then moves in the gravitational field generated by the planet without affecting the motion of this one. Such approximation is precisely the idea behind the first nearly integrable extension we can introduce, named *restricted N-body problem* (RNBP hereafter), in which we study the dynamics of a massless point $\mathcal{P}$ under the influence of $N - 2$ uncoupled Kepler problems.
Specifically, the case $N = 3$ with the Sun ($\mathcal{P}_0$) and Jupiter ($\mathcal{P}_1$) as major bodies (Sun–Jupiter restricted 3-body problem) is already sufficiently representative in the Solar System, at least at first glance, since it involves the two most massive objects, and shows a highly non-trivial dynamics which is still not exhaustively understood. Furthermore, we assume that $\mathcal{P}_0 - \mathcal{P}_1$ move on circular orbits around their common center of mass. Thence, $\mathcal{P}$ is expected to describe a non-integrable near-Keplerian orbit where now all the orbital elements depend on time: $a = a(t), e = e(t), i = i(t), f = f(t), \omega = \omega(t), \Omega = \Omega(t)$.

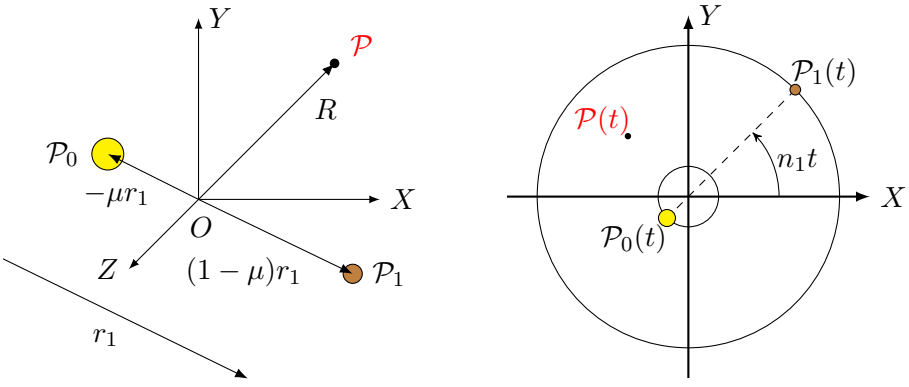


Figure 4. Sketch of the Sun–Jupiter circular R3BP.

The dynamical system at issue comes from the Hamiltonian setting:[10]

$$\mathcal{H} = \frac{\|P\|^2}{2} - \frac{\mathcal{G}m_0}{\|R + \mu r_1\|} - \frac{\mathcal{G}m_1}{\|R - (1-\mu)r_1\|} + n_1 J_1 \,, \tag{10}$$

where $\mu = \dfrac{m_1}{m_0 + m_1} \approx 1 \cdot 10^{-3}$ is the mass parameter, $r_1(M_1) = \|r_1\| (\cos M_1, \sin M_1)$, $M_1 = n_1 t$, $(R, M_1, P, J_1) \in T^*(\mathbb{R}^3 \setminus \{-\mu r_1, (1-\mu)r_1\} \times \mathbb{T})$ are symplectic variables.
So

$$\dot{x} = X(x) \,, \quad x = (R, M_1, P, J_1) \,, \; X(x) = \left(\frac{\partial \mathcal{H}}{\partial P}, \frac{\partial \mathcal{H}}{\partial J_1}, -\frac{\partial \mathcal{H}}{\partial R}, -\frac{\partial \mathcal{H}}{\partial M_1} \right) \,. \tag{11}$$

In a suitable frame $Oxyz$ rotating at frequency n_1 around the z-axis, the system possesses a constant of motion, called *Jacobi integral* $\mathcal{J}$, and five equilibrium points L_1, L_2, L_3, L_4, L_5, called *Lagrangian points*. In particular, L_1, L_2, L_3 are of (partially) hyperbolic nature: we can then apply the machinery in §3 to identify regular and chaotic motions in phase space, together with hyperbolic invariant manifolds (Fig. 5).

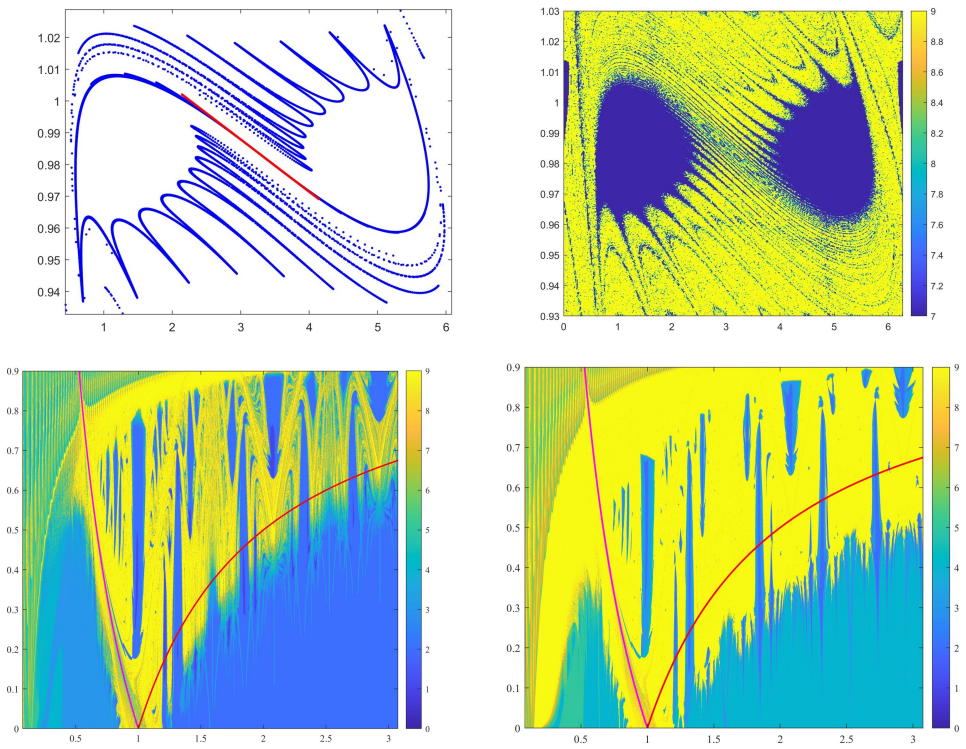


Figure 5. FLI for the Sun-Jupiter circular R3BP. **Top panels**: W^u and E^u stemming from L_3 numerically propagated (left) and corresponding plot using the FLI map (right). **Bottom panels**: FLI cartographies for two different exposition times: $\tau = 50T_1$ on the left and $\tau = 1000T_1$ on the right, where T_1 is Jupiter's revolution period. The Poincaré surface of section is the plane (a, e) and the two curves represent the loci of points $\{\|r_1\| = a(1-e)\}$ (red), $\{\|r_1\| = a(1+e)\}$ (magenta).

[10] In case of unfamiliarity with the Hamiltonian context or for a more detailed reading, cf. [7].

5 Long-term stability of asteroid families

We conclude with a brief argument about long-term stability of families of asteroid looking at Fig. 5, taking advantage once more of the power of chaos indicators. Besides highlighting the complexity of the unstable manifold originating at L_3 (top panels), the FLI reveals several other remarkable structures if we compare the illustrations on bottom panels: we can see how regions of regular orbits (deep blue colour) permeate the whole phase space, even above the red line of *pericenter crossing*, in which an asteroid is located whenever its minimum distance from the Sun is equal to Jupiter's orbital radius. It is worth noticing that such curve overestimates the boundary of the lower stability region. Moreover this boundary has a fractal shape whose form becomes clearer increasing the integration time, as displayed in the image on the right. Also, slightly chaotic spikes penetrate the regular regions, that shall be attributed to other dynamical phenomena (like resonances). Lastly, we can also notice in the plot on the left arch-like structures created by the projection of invariant manifolds of various equilibria or periodic orbits (the "Arches of Chaos" [9], whose detailed classification has not been carried out yet).

References

[1] J. Banks, J. Brooks, G. Cairns, G. Davis and P. Stacey., *On Devaney's definition of chaos.* The American Mathematical Monthly 99.4 (1992), pp. 332–334.

[2] G. Benettin, L. Galgani, A. Giorgilli and J.M. Strelcyn, *Lyapunov characteristic exponents for smooth dynamical systems and for Hamiltonian systems; a method for computing all of them. Part 1: Theory.* Meccanica 15.1 (1980), pp. 9–20.

[3] A. Celletti, "Stability and chaos in celestial mechanics". Springer Science & Business Media, 2010.

[4] F. Fassò, "Sistemi dinamici differenziabili". Lecture notes, 2017.

[5] M. Guzzo, E. Lega and C. Froeschlé, *On the numerical detection of the effective stability of chaotic motions in quasi-integrable systems.* Physica D: Nonlinear Phenomena 163.1–2 (2002), pp. 1–25.

[6] E. Lega, M. Guzzo and C. Froeschlé, *Theory and applications of the Fast Lyapunov Indicator (FLI) method.* Chaos Detection and Predictability (2016), pp. 35–54.

[7] A. Morbidelli, "Modern celestial mechanics: aspects of solar system dynamics". 2022.

[8] M. R. and C. Efthymiopoulos, *Characterization of the stability for trajectories exterior to Jupiter in the restricted three-body problem via closed-form perturbation theory.* Proceeding IAU: Multi-scale (time and mass) dynamics of space objects (2021).

[9] N. Todorović, W. Di and A.J. Rosengren, *The arches of chaos in the Solar System.* Science Advances 6.48 (2020), pp. eabd1313.

Introduction to sub-Riemannian geometry

Alessandro Socionovo (*)

Abstract. The purpose of this short article is to introduce the reader to the world of sub-Riemannian geometry, starting from the basics up to one of the main open problems of sub-Riemannian geometry: the regularity of length minimization curves.
We start from the study of a very simple model problem taken from real life. After having understand the key properties which characterize this toy model, we give the general and basic definitions, up to the one of the end point map.
The end-point map is the central object of the paper: the difficulty of the afromentioned regularity problem is due to the singularities of the end-point map, i.e., to the presence of points where its differential is not surjective. Curves corresponding to such points are called abnormal.
Finally, after a first order analysis of the end-point map, we make an overview about active reasearch about lengt-minimizing sub-Riemannian curves, citing and briefly explaining some new regularity results.

1 Introduction

The purpose of this short article is to introduce the reader to the world of sub-Riemannian geometry, starting from the basics up to one of the main open problems of sub-Riemannian geometry: the regularity of length minimization curves.

We start from the study of a very simple model problem taken from real life. After having understand the key properties which characterize this toy model, we give the general and basic definitions, up to the one of the end point map.

The end-point map is the central object of the paper: the difficulty of the afromentioned regularity problem is due to the singularities of the end-point map, i.e., to the presence of points where its differential is not surjective. Curves corresponding to such points are called abnormal.

Finally, after a first order analysis of the end-point map, we make an overview about active reasearch about lengt-minimizing sub-Riemannian curves, citing and briefly explaining some new regularity results.

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: alessandro.socionovo@math.unipd.it. Seminar held on 15 June 2022.

2 Preliminary notions

In this section, we spend a very few words to fix the following notations concerning differential geometry.

1. M denotes a smooth and connected manifold of dimension n and $q \in M$ ia s point of the manifold.

2. If $\varphi : M \to M$ is a smooth map, we denote the differential (or pushforward) of φ at a point $q \in M$ as $d_q\varphi = \varphi_{*,q} : T_qM \to T_{\varphi(q)}N$. The differential of φ is the map $d\varphi = \varphi_* : TM \to TM$ with $d\varphi|_{T_qM} = d_q\varphi$. We also denote $\varphi^* : T^*M \to T^*M$ the dual map of the differential, namely

$$\langle \omega, \varphi_* v \rangle = \langle \varphi^* \omega, v \rangle, \quad \forall \omega \in T^*M, v \in TM.$$

   Here $\langle \cdot, \cdot \rangle$ denotes the usual action of 1-forms over vectors, and $\langle \omega, v \rangle|_q = \langle \omega(q), v(q) \rangle$.

3. We denote the set of smooth vector fields of M with $\mathrm{Vec}(M)$. If $X \in \mathrm{Vec}(M)$, we denote with $P_{s,t} = e^{(t-s)X}$ its flow. When $\{X_t\}_{t\in\mathbb{R}}$ is a nonautonomus vector field, we denote its flow adopting the notation of chronological calculus

$$P_{s,t} = \overrightarrow{\exp} \int_s^t X_\tau d\tau.$$

   We point out that, even if this is just a notation, it respects in a certain sense the integral properties. We will use the chronological calculus in Section 5.

4. Let $X_1, \dots, X_k \in \mathrm{Vec}(M)$. With $[X_1, X_2]$ we denote the Lie bracket of two vector fields X_1, X_2. We also use the compact notation

$$[X_1, \dots, X_k] = [X_1, [X_2, [\dots [X_{k-1}, X_k] \dots].$$

   When $\mathcal{D} \subset \mathrm{Vec}(M)$ we define

$$\mathcal{D}^k(q) = [\underbrace{\mathcal{D}, \dots, \mathcal{D}}_{(h-1)\text{-times}}](q) = \{[X_1, \dots, X_k](q) | X_1, \dots, X_k \in \mathcal{D}\}.$$

We assume the reader of this paper to be familiar with these arguments. If it is not the case, we refer the reader to [1]. In particular, one can find all these notions in Chapter 2, except for the chronological calculus, which is developed in Chapter 6.

3 A toy problem

In this section, we study the driving trajectories of a car moving on the road as a model problem (taken from real life) for sub-Riemannian geometry.

This toy model can be viewed mathematically in $\mathbb{R}^3 = (x_1, x_2, \theta)$, where the first two coordinates (x_1, x_2) determine the position of the car on the road (actually, for us the road

is the whole plane), and the third coordinate θ define the angle of the car's wheels, namely it defines the moving direction of the the car. When θ is fixed, our car moves in the plane following the direction $(\cos\theta, \sin\theta)$, i.e. it solves the following PDE

$$\begin{cases} \dot{x}(t) = \cos\theta \\ \dot{y}(t) = \sin\theta \\ \dot{\theta}(t) = 0. \end{cases}$$

In other words we are moving along the vector field $X_1 := \cos\theta \frac{\partial}{\partial x_1} + \sin\theta \frac{\partial}{\partial x_2}$. Otherwise, when the car is stopped, we can steer the wheels, namely we are changing the value of θ, solving this second equation

$$\begin{cases} \dot{x}(t) = 0 \\ \dot{y}(t) = 0 \\ \dot{\theta}(t) = 1. \end{cases}$$

In this case we are following the vector field $X_2 := \frac{\partial}{\partial \theta}$. Finally, what it is not allowed for us, is to move the car orthogonally with respect to the vector $(\cos\theta, \sin\theta)$

$$\begin{cases} \dot{x}(t) = -\sin\theta \\ \dot{y}(t) = \cos\theta \\ \dot{\theta}(t) = 0. \end{cases}$$

In terms of vector fields, the non-allowed direction is $X_3 := -\sin\theta \frac{\partial}{\partial x_1} + \cos\theta \frac{\partial}{\partial x_2}$.

Remark 3.1 The vector fields X_1 and X_2 do not commute and, in particular, their Lie bracket is $[X_1, X_2] = X_3$. Moreover, the set $\{X_1(p), X_2(p), X_3(p)\}$ is a basis of $\mathbb{R}^3$ for any point $p = (x, y, \theta)$. Hence, the distribution $\mathcal{D}_p = \text{span}\{X_1(p), X_2(p)\}$ is bracket generating.

Clearly, in all this construction, we are supposing that both the velocity of the car and the velocity of the angle θ is constant and unit. To make our model complete, we have to make this parameter dynamic, namely we can move in $\mathbb{R}^3$ only along curves $\gamma : [0, 1] \to \mathbb{R}^3$ that satisfy the following PDE

$$\dot{\gamma}(t) = u_1(t)X_1(\gamma(t)) + u_2(t)X_2(\gamma(t)).$$

Here $\gamma = (x, y, \theta)$ is the curve denoting the configuration of the car at the time t, while u_1 and u_2 are generic functions from $[0, 1]$ to $\mathbb{R}$, called controls.

Remark 3.2 In spite of we restrict the set of admissible curves (i.e., there is a nonallowed direction for the movement of the car), it is easy to see that we can join anyway any couple of point of $\mathbb{R}^3$.

If we define a metric g for vectors that are linear combinations of $X_1(p)$ and $X_2(p)$, we can also measure the length of every admissible curve in this model as the integral of its derivative norm. This allow us to measure the distance between any couple of points

as the infimum of the length of admissible curves joining them. This makes our space a metric space.

We will se that the set $\mathbb{R}^3$ with the vector fields X_1 and X_2, and with any metric g as described above, is an example of a sub-Riemannian manifold.

4 Sub-Riemannian geometry

In this section we generalize the construction of the previous section. We start with the general definition of a sub-Riemannian manifold

Definition 4.1 A sub-Riemannian manifold of dimension n and constant-rank m is a triple $(M, \mathcal{D}, g)$ where

i) M is a smooth manifold of dimension n.

ii) $\mathcal{D}$ is a smooth distribution of rank m satisfying the Hormander condition, namely there exists $k \in \mathbb{N}$ such that

$$\mathcal{D}_q^{(k)} = T_q M, \quad \forall\, q \in M.$$

iii) g_q is a scalar product on $\mathcal{D}_q$, smoothly depending on $q \in M$. Sometimes, we use the symbol $\langle \cdot, \cdot \rangle$ instead of g for the sub-Riemannian metric.

We call n the dimension and m the rank of the sub-Riemannian manifold.

Let us analyze the model presented in the previous section. There, $\mathbb{R}^3$ plays the role of the manifold M, and $\mathcal{D}$ is the smooth distribution $D_q = \operatorname{span}\{X_1(q), X_2(q)\}$. Thus, with any metric defined on $\mathcal{D}$, $(\mathbb{R}^3, D, \langle \cdot, \cdot \rangle)$ is a sub-Riemannian manifold of dimension 3 and rank 2. Moreover, here we have the property that X_1 and X_2 are a moving frame for $\mathcal{D}$, i.e. they globally generate the distribution $\mathcal{D}$.

Remark 4.2 From now on, we assume that $\mathcal{D}$ admits a globally generating family of vector fields $\{f_1, \dots, f_m\}$, namely

$$\mathcal{D}_q = \operatorname{span}\{f_1(q), \dots, f_m(q)\}, \quad \forall\, q \in M.$$

We also assume g to be the metric that makes $f_1, \dots, f_m$ orthonormal. These assumptions are not restrictive, as explained in [1, Chapter 3].

If $u \in \mathbb{R}^m$ we denote with f_u the vector field

$$f_u(q) = \sum_{i=1}^{m} u_i f_i(q), \quad q \in M.$$

Definition 4.3 Let $(M, \mathcal{D}, g)$ be a sub-Riemannian manifold. A Lipschitz curve $\gamma : [0, T] \to M$ is said to be admissible or horizontal if

$$\dot\gamma(t) \in \mathcal{D}_{\gamma(t)}, \quad \text{for a.e. } t \in [0, T].$$

Notice that, in definition 4.3 we are asking that

$$\dot{\gamma}(t) = f_{u(t)}(\gamma(t)) = \sum_{i=1}^{m} u_i(t) f_i(\gamma(t)), \quad \text{for a.e. } t \in [0,T],$$

for some (unique) $u = (u_1, \dots, u_m) \in L^\infty(I, \mathbb{R}^m)$. The function u is called the control associated to γ. The reason why we are asking that u is essentially bounded is explained in Remark 4.6 below.

Definition 4.4 Let $\gamma : [0,T] \to M$ be an admissible curve. We define the length of γ as the integral of the sub-Riemannian norm of its tangent vector, namely

$$\ell(\gamma) = \int_0^T g(\dot{\gamma}(t), \dot{\gamma}(t))^{1/2} dt.$$

Notice that, when g makes $f_1, \dots, f_m$ orthonormal, we have

$$\ell(\gamma) = \int_0^T \sqrt{\sum_{i=1}^{m} |u_i(t)|^2} dt = \|u\|_{L^1([0,T],\mathbb{R}^m)}. \tag{4.1}$$

Then, by Remark 4.2, the definition of length of an admissible curve is well posed, since $L^\infty([0,T],\mathbb{R}^m) \subset L^1([0,T],\mathbb{R}^m)$.

Definition 4.5 Let $(M, \mathcal{D}, g)$ be a sub-Riemannian manifold and let $q_0, q_1 \in M$. We define the sub-Riemannian distance between q_0 and q_1 as

$$d(q_0,q_1) = \inf\{\ell(\gamma) \,|\, \gamma : [0,T] \to M \text{ admissible }, T \in \mathbb{R}\,, \gamma(0) = q_0\,, \gamma(T) = q_1\} \tag{4.2}$$

Remark 4.6 We defined horizontal curves to be Lipschitz. As clearly explained in [1, Chapter 3], this choice corresponds to essentially boundness of the controls. Since $L^\infty(I,\mathbb{R}^m) \subset L^2(I,\mathbb{R}^m) \subset L^1(I,\mathbb{R}^m)$, by (4.1) we can define horizontal curves with corresponding controls in L^1 or L^2, without changing the length definition. In these cases, we should ask in definition 4.3 curves to be absolutely continuos.

Even if $\mathrm{Lip}([0,1],M) \subset AC([0,1],M)$, the infimum in (4.2) does not change. Thus, from now on, we can assume controls associated with horizontal curves to be in L^1 or L^2 if we prefer.

Remark 4.7 We emphasize that when $\mathcal{D} = TM$, we have in fact a Riemannian manifold, and all the definitions given by now are the same as in the classical Riemannian geometry.

Theorem 4.8 (Chow-Raschevskii) *Let M be a sub-Riemannian manifold. Then*

i) *(M,d) is a metric space;*

ii) *The topology induced by* $(M; d)$ *is equivalent to the manifold topology.*

Remark 4.9 One of the main consequences contained in the proof of the Chow-Raschevskii theorem is that, thanks to the bracket-generating condition, every couple of points in M is always joined by an admissible curve. Hence $d(q_0, q_1) < +\infty$, for every $q_0, q_1 \in M$.

The proof of this theorem is very long and it is given with all the details in [1, Chapter 3]. The key idea is to use the Hormander condition in such a way that, following the flow of n vector fields choosing among $\{f_1, \dots, f_m\}$, one can always cover a neighborhood of every point.

We conclude this section introducing the objects we are interested in studying the regularity of, namely length-minimizing admissible curves. We refer the reader to [1, Chapter 3] for the proofs of the statements given in this final part of this section.

Definition 4.10 Let $\gamma : [0, T] \to M$ be an admissible curve. We say that γ is a length-minimizer (or a length-minimizing curve) if

$$\ell(\gamma) = d(\gamma(0), \gamma(T)).$$

In other words, a length minimizer minimize the length among admissible curves with same endpoints.

The (local) existence of length-minimizers is guaranteed by the following Theorem.

Theorem 4.11 (Existence of length-minimizers) *Let* M *be a sub-Riemannian manifold and* $q_0 \in M$*. Assume that the closed ball* $\overline{B}_{q_0}(r)$ *is compact, for some* $r > 0$*. Then for all* $q \in B_{q_0}(r)$ *there exists a length minimizer joining* q_0 *and* q*.*

Remark 4.12 The compactness assumption in Theorem 4.11 is completely natural and cannot be removed. In fact, the existence of length-minimizers between two points is not true in general, as it happens, for example, for two symmetric points with respect to the origin in $M = \mathbb{R}^n \setminus \{0\}$, endowed with the Euclidean metric.

On the other hand, when length-minimizers exist between two fixed, they may not be unique, as in the case of two antipodal points on the sphere $\mathbb{S}^2$.

The existence of length-minimizing curves leads to the following characterization of the metric completness of sub-Riemannian distance.

Proposition 4.13 *Let* M *be a sub-Riemannian manifold. Then the three following properties are equivalent*

i) (M, d) *is complete;*

ii) $\overline{B}_q(r)$ *is compact for every* $q \in M$ *and* $r > 0$*;*

iii) *There exists* $\varepsilon > 0$ *such that* $\overline{B}_q(r)$ *is compact for every* $q \in M$*.*

Combining Theorem 4.11 and Proposition 4.13 we obtain the total existence of length minimizers.

Corollary 4.14 *Let (M, d) be a complete sub-Riemannian manifold. Then for every $q_0, q_1 \in M$ there exists a length minimizer joining q_0 and q_1.*

5 End-point map: first-order analysis and Pontryagin extremals

In this section we define the end-point map and develop its first-order analysis to get necessary conditions for the minimality of admissible curves. From now on, $(M, \mathcal{D}, g)$ is a n-dimensional sub-Riemannian manifold. We assume $\mathcal{D}$ is globally generated by the vector fields $f_1, \dots, f_m$, and g is the metric making the generating family orthonormal.

Fix $q_0 \in M$. For $u \in L^2(I, \mathbb{R}^m)$, the corresponding horizontal trajectory γ based at q_0 satisfies the Cauchy problem

$$\dot\gamma(t) = f_{u(t)}(\gamma(t)), \quad \gamma(0) = q_0. \tag{5.1}$$

For semplicity (and without loss of generality, see [1, Chapter 8], we assume that γ is dedfined on the whole interval I. Here and hereafter, $I = [0, 1]$ denotes the closed unit interval.

Definition 5.1 The end-point map based at q_0 is the map

$$E_{q_0} : L^2(I, \mathbb{R}^m) \to M, \quad E_{q_0}(u) = \gamma_u(1),$$

where $\gamma_u : I \to M$ is the unique solution of the Cauchy problem (5.1).

A first property of the end-point map, which is a consequence of the Chow-Raschevskii theorem, is its openness.

Theorem 5.2 *Fix $q_0 \in M$. Then the end-point map E_{q_0} is open at every $u \in L^2(I, \mathbb{R}^m)$.*

The key result in the derivation of Pontryagin extramal conditions for the minimality is the smoothness of the end-point map with the computation of its differential.

Theorem 5.3 *Let $q_0 \in M$. Then, the end-point map E_{q_0} is smooth. In particulat, its Fréchet differential $D_u E_{q_0} : L^2(I; \mathbb{R}^m) \to T_{\gamma_u(1)} M$ is given by*

$$D_u E_{q_0}(v) = \int_0^1 (P^u_{t,1})_* f_{v(t)}\big|_{\gamma_u(1)}\, dt. \tag{5.2}$$

Here $P^u_{t,1}$ denotes the nonautonomus flow generated by the vector field f_u.

The proof of both theorems 5.2 and 5.3 are given in [1, Chapter 8].

Remark 5.4 The geometrical meaning of (5.2) is that the differential of the end-point map computed at u acts on an element v by integrating the vector field f_v along the whole curve γ, with the pushforward which is carried out through the flow of u.

The proof of the Pontryagin conditions relies upon the Lagrange multipliers rule.

Theorem 5.5 (Lagrange multipliers rule) *Let $\mathcal{H}$ be an Hilbert space and let M be a smooth n-dimensional manifold. Let $\mathcal{U} \subset \mathcal{H}$ be an open subset of $\mathcal{H}$. Consider two smooth maps $\varphi : \mathcal{U} \to \mathbb{R}$ and $F : \mathcal{U} \to M$. Fix a point $q \in M$ and assume that $u \in \mathcal{U}$ is a solution to the minimization problem*

$$\min\{\varphi(v) \,|\, F(v) = q\} = \min \varphi\big|_{F^{-1}(q)}. \tag{5.3}$$

Then there exists a non-null $(\lambda, \nu) \in T_q^*M \times \mathbb{R}$, i.e. $(\lambda, \nu) \neq (0, 0)$, such that

$$\lambda d_u F + \nu d_u \varphi = 0. \tag{5.4}$$

Remark 5.6 We explicitly puntualize that formula (5.4) means that for every $v \in \mathcal{H} = T_u\mathcal{U}$ one has

$$\langle \lambda, d_u F(v) \rangle + \nu d_u \varphi(v) = 0.$$

The compact notation in (5.4) will be used also in the sequel, with the same meaning.

The idea of the proof of Theorem 5.5 is the following. If u solves (5.3), then the extended map

$$\bar{F} : \mathcal{U} :\to M \times \mathbb{R}, \quad \bar{F}(v) = (F(v), \varphi(v)),$$

is not open at $u \in \mathcal{U}$. Then, the differential $d_u\bar{F} = (d_u F, d_u\varphi)$ is not surjective, namely its image is annilhated by a certain $0 \neq (\lambda, \nu) \in T_q^*M \times \mathbb{R}$, and this is exactly (5.4).

As a consequence of theorem 5.5, up to scalar multiplications, there may happen the following situations

(N) $\nu = -1$. Then from (5.4) we deduce

$$\lambda d_u F = d_u \varphi; \tag{5.5}$$

(A) $\nu = 0$. Then $\lambda \neq 0$, and from (5.4) we deduce

$$\lambda d_u F = 0.$$

The case (N) is called normal, while the case (A) is called abnormal or singular.

Remark 5.7 Classically, Theorem 5.5 is presented when $\mathcal{H} = \mathbb{R}^n = (x_1, \dots, x_n)$, $M = \mathbb{R}^m = (x_1, \dots, x_m)$, $u = \bar{x} = \operatorname{argmin}\{\varphi(x) | F(x) = 0\} \in \mathbb{R}^n$, and one adds the assumption $J_{\bar{x}}F \neq 0$, removing the possibility for $\bar{x}$ to be abnormal. In this case, equation (5.5) reads

$$\nabla_x(\varphi - \lambda \cdot F)(\bar{x}) = 0. \tag{5.6}$$

Denoting $L(x, \lambda) = \varphi(x) - \lambda \cdot F(x)$ (the Lagrangian of the problem), the bound given by F reads

$$\nabla_\lambda(\varphi - \lambda \cdot F)(\bar{x}) = 0 \tag{5.7}$$

Combining (5.6) and (5.7), one obtains the classical version of the Lagrange multipliers theorem, namely, a necessary condition for the minimality is the existence of a $\lambda \in \mathbb{R}^m$ such that

$$\nabla(\varphi - \lambda \cdot F)(\bar{x}) = \nabla L(\bar{x}, \lambda) = 0.$$

The vector λ is called Lagrange multiplier.

We apply the result of Theorem 5.5 to the end-point map based at a certain point $q_0 \in M$. Thus, we consider the case when $F = E_{q_0}$ and $\varphi = J$ is the sub-Riemannian energy, namely

$$J : L^2(I, \mathbb{R}^m) \to \mathbb{R}, \quad J(u) = \frac{1}{2}\int_0^1 |u(t)|^2 dt = \|u\|^2_{L^2(I,\mathbb{R}^m)}. \tag{5.8}$$

Since γ is a length minimizer if and only if the corresponding control u minimize the energy functional (5.8), one immediately obtains the following result.

Proposition 5.8 *Let $\gamma : I \to M$ be a length minimizing curve with correspond control u. Then there exists $(\lambda, \nu) \in T^*_q M \times \mathbb{R}$ such that $(\lambda, \nu) \neq (0, 0)$, and*

$$\lambda d_u E_{q_0} + \nu d_u J = 0. \tag{5.9}$$

Remark 5.9 Since $J(v) = \frac{1}{2}\|v\|^2_{L^2}$, then $d_u J(v) = (u, v)_{L^2}$ and, identifying L^2 with its dual, we have $d_u J = u$.

Remark 5.10 Even if the end point map is always open, the extended one $\bar{E}_{q_0} = (E_{q_0}, J)$ is not. Actually, if γ (with associated control u) is a length minimizer, then the extended end point map is not open at u, leading to (5.9).

Theorem 5.11 (Pontryagin maximum principle) *Let $q_0 \in M$. Let u be the minimal control corresponding to a curve $\gamma : [0, 1] \to M$ based at q_0 which is length-minimizing (or, equivalently, energy-minimizing) and parametrized with constant speed, namely*

$$\dot{\gamma}(t) = f_{u(t)}(\gamma(t)) \text{ a.e. on } [0, 1], \quad \gamma(0) = q_0 \in M.$$

Let $q_1 = E_{q_0}(u) = \gamma(1)$. Then there exists $\lambda_1 \in T^*_{q_1} M$ such that, defining $\lambda(t) = (P^u_{t,1})^* \lambda_1$, one of the following condition is satisfied

(N) *For every $i = 1, \dots, m$ we have*

$$u_i(t) = \langle \lambda(t), f_i(\gamma(t)) \rangle, \text{ for a.e. } t \in [0, 1], \tag{5.10}$$

and this occours if and only if u satisfies (5.9) with $(\lambda, \nu) = (\lambda_1, -1)$, namely

$$\lambda_1 d_u E_{q_0} = u \tag{5.11}$$

(A) *For every $i = 1, \dots, m$ we have*

$$0 = \langle \lambda(t), f_i(\gamma(t)) \rangle, \text{ for a.e. } t \in [0,1], \tag{5.12}$$

and this occours if and only if u satisfies (5.9) with $(\lambda, \nu) = (\lambda_1, 0)$, namely

$$\lambda_1 d_u E_{q_0} = 0$$

Remark 5.12 Since $q_1 = \gamma(1)$, $\lambda_1 \in T^*_{q_1} M$ and $P^u_{s,t}$ is the flow of γ in M, then $\lambda(t) = (P^u_{t,1})^* \lambda_1$ is a lift of $\gamma(t)$, namely

$$\pi(\lambda(t)) = \gamma(t), \quad \forall t \in [0,1].$$

Proof. Let us prove (N). The proof of (A) is analogous.

Assume that u satisfies (5.10) for some λ_1, and let us prove that the curve defined by $\lambda(t) = (P^u_{t,1})^* \lambda_1$ satisfies (5.11), which means that for every $v \in L^2(I, \mathbb{R}^m)$ we have

$$\langle \lambda_1, d_u E_{q_0}(v) \rangle = (u, v)_{L^2}. \tag{5.13}$$

Using (5.2), the left hand side in (5.13) reads

$$\begin{aligned}
\langle \lambda_1, d_u E_{q_0}(v) \rangle &= \int_0^1 \langle \lambda_1, \big((P^u_{t,1})_* f_{v(t)} \big)(q_1) \rangle dt = \int_0^1 \langle \lambda_1, (P^u_{t,1})_* \big(f_{v(t)}((P^u_{t,1})^{-1} q_1) \big) \rangle dt \\
&= \int_0^1 \langle (P^u_{t,1})^* \lambda_1, f_{v(t)}(\gamma(t)) \rangle dt = \int_0^1 \langle \lambda(t), f_{v(t)}(\gamma(t)) \rangle dt \\
&= \int_0^1 \sum_{i=1}^m \langle \lambda(t), f_i(\gamma(t)) \rangle v_i(t) dt.
\end{aligned}$$

Then (5.13) becomes

$$\int_0^1 \sum_{i=1}^m \langle \lambda(t), f_i(\gamma(t)) \rangle v_i(t) dt = \int_0^1 \sum_{i=1}^m u_i(t) v_i(t) dt.$$

Since v is arbitrary, this implies (5.10).

Conversely, let us assume there exists $\lambda_1 \in T_{q_1} M$ such that the curve defined by $\lambda(t) = (P^u_{t,1})^* \lambda_1$ satisfies (5.10). Then, following the above computations in the opposite direction, one obtains exactly (5.11). □

The Pontryagin maximum principle provides a first-order necessary condition for a curve γ to be length-minimizing, leading to the following definition

Definition 5.13 Let $\gamma : [0,1] \to M$ be a horizontal curve based at q_0 with corresponding control u. Assume that γ and u satisfy at least one condition between (N) and (A) of Theorem 5.11. Then we give the following definitions

i) The curve $\lambda : [0,1] \to T^*M$ is called adjoint curve;

ii) The couple $(u(t), \lambda(t))$ is called sub-Riemannian extremal. When the extremal satisfies (N) it is called normal extremal. Otherwise, it is said to be an abnormal or a singular extremal.

Remark 5.14 A length minimizer is necessarily an extremal (normal, abnormal or both), as a consequence of the Lagrange multipliers rule. Conversely, there are many examples of extremals which are not length minizers.

Remark 5.15 It is a well known fact that if γ is normal, then by (5.10) it follows that γ is smooth, namely C^∞. In particular, in classical Riemannian geometry (i.e. $\mathcal{D} = TM$) the problem of the regularity of geodesics have already been solved, since the abnormal case could not appear.

Remarks 5.14 and 5.15 emphasize the difficulty of the regularity problem, which we underline once again to be related with the presence of the abnormal extremals. In the next section, we analyze this problem, concluding this paper with the statements of some new regularity results.

6 An overview of the regularity problem

This final section is devoted both to make an hystorical overview of the regularity problem and to present some new regularity results.

The story of the problem of the regularity of geodesics in sub-Riemannian geometry starts with the ground-breaking work [14] of R. Montgomery, where it is proved for the first time that singular curves can be as a matter of fact length-minimizing. Also nice abnormal extremals, see [13], are locally length minimizing. An algorithm for producing many new examples of abnormal extremals is proposed in [7]. The length-minimality property of all these examples is not yet well-understood.

The main recent approachs to the regularity problem are the following

A1) A first approach is based on the analysis of specific singularities such as corners. In this optics, the recent, spiral-like curves or curves with no straight tangent line. This approach does not use general open mapping theorems but it rather relies on the ad hoc construction of shorter competitors, see [3, 8, 9, 12, 15, 16, 17].

A2) On the other hand, necessary conditions for the minimality of singular extremals can be obtained from the differential study of the end-point map. The theory is well-known till the second order and was initiated by Goh [6] and developed by Agrachev and Sachkov in [2].

In the following, we explain better this different kind of approachs.

6.1 About approach (A1)

The most elementary kind of singularity for a Lipschitz curve is of the corner-type: at a given point, the curve has a left and a right tangent that are linearly independent. In [12] and [8] it was proved that length minimizers cannot have singular points of this kind, namely

Theorem 6.1 (Hakavuori, Le Donne 2016) *Let $\gamma : [0,1] \to M$ be a horizontal curve. If γ has corner-type singularity, then it cannot be length-minimizing.*

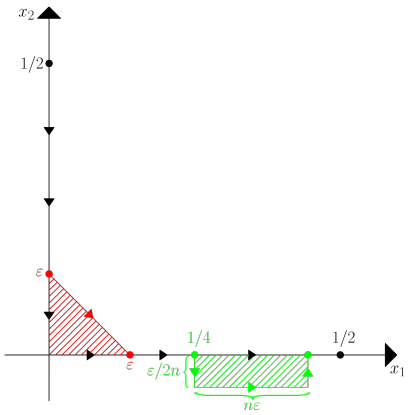
A very basic idea of the proof can be given when $M = \mathbb{R}^3$ is a sub-Riemannian manifold of dimension 3 and rank 2 (as in the example of section 2). Then, a horizontal curve is $\gamma(t) = (\gamma_1(t), \gamma_2(t), \gamma_3(t))$, where the third coordinate of γ is the area of the graph of its horizontal projection with $\kappa(t) = (\gamma_1(t), \gamma_2(t))$.

When γ has a corner, it is not restrictive to suppose that κ is a path walking through the coordinate axes, with corner singularity at the origin.

One builds a competitor curve $\bar{\gamma}$ depending on a fixed poistive parameter ε and on an integer n, modifying the horizontal projection κ. This modification on κ also modifies the third coordinate of γ. Since the sub-Riemannian length of γ is the Euclidean length of κ, such a modification of γ is made into the following two steps

i) First, one cuts the original curve near the singularity, as in the red path of the figure below. This implies a gain of length for $\bar{\gamma}$, but it modifies the end point.

ii) In order to restore the end point, one has to correct the curve in such a way to have a positive gain of length. This step is realized by the green path in the figure below.



$$\kappa(t) = \begin{cases} (0, -t), & t \in [-1, 2, 0), \\ (t, 0), & t \in [0, 1/2]. \end{cases}$$

In this case, we have for n large enough a gain of length given by

$$\ell(\gamma) - \ell(\bar{\gamma}) = (2 - \sqrt{2} - \frac{1}{n})\varepsilon > 0.$$

This proves that the curve γ cannot be a length minimizer, because of the presence of a corner. When M has dimension n and rank m, one needs $n - m$ integer parameters $k_1, \ldots, k_{n-m}$ and $n - m$ associated correction paths to restore the end point, solving a sistem of linear equations.

These results have been improved in [16], where it is proved the following statement.

Theorem 6.2 (Monti, Pigati, Vittone 2018) *Let $\gamma : [0,1] \to M$ be a horizontal length-minimizing curve. Then there is at least one horizontal line in the tangent cone of γ.*

In other words, at any point, the tangent cone to a length-minimizing curve contains at least one line (a half line, for extreme points).

The uniqueness of this tangent line for length minimizers is an open problem. Indeed, there exist other types of singularities related to the non-uniqueness of the tangent. In particular, there exist spiral-like curves whose tangent cone at the center contains many and in fact all tangent lines. These curves may appear as Pontryagin extremals in sub-Riemannian geometry and theorem 6.2 are not enough to prove the nonminimality of spiral-like extremals.

The problem of the minimality of spiral-like curves have been solved by me and R. Monti.

Theorem 6.3 (Monti, S. 2021) *Let $(M, \mathcal{D}, g)$ be an analytic sub-Riemmanian manifold of rank 2 satisfying the following condition*

$$\big[[f_{i_1}, \dots, f_{i_j}], [f_{i_1}, \dots, f_{i_k}]\big] = 0, \quad j, k \geq 2.$$

Then any horizontal spiral $\gamma \in AC([0,1]; M)$ is not length-minimizing near its center.

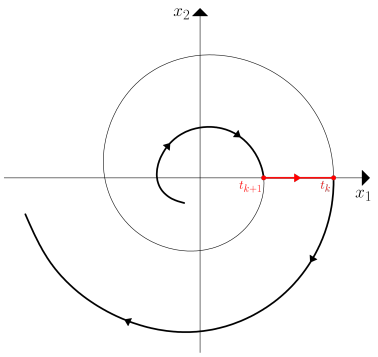
For the notion of horizontal spirals and the details of the proof, which needs very long and detailed computations, we refer the reader to our paper [17].

We just give an idea of strategy we used for cutting and then correcting the spiral.

The cutted horizontal curve is

$$\mathrm{k}_k^{\mathrm{cut}}(t) = \begin{cases} \mathrm{k}(t), & t \in [0, t_{k+1}] \\ (t, 0), & t \in (t_{k+1}, t_k) \\ \mathrm{k}(t), & t \in [t_k, 1]. \end{cases}$$

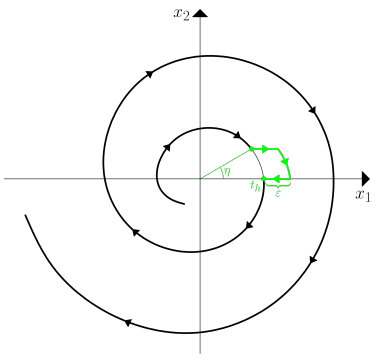
It depends on the integer k.



The correction path is defined by

$$\begin{array}{ll} \kappa(t), & t \in [0, t_{h\eta}], \\ \kappa(t_{h\eta}) + (\mathrm{sgn}(\varepsilon)(t - t_{h\eta}), 0), & t \in [t_{h\eta}, t_{h\eta} + |\varepsilon|], \\ \kappa(t - |\varepsilon|) + (\varepsilon, 0), & t \in [t_{h\eta} + |\varepsilon|, t_h + |\varepsilon|], \\ \kappa(t_h) + (2\varepsilon + \mathrm{sgn}(\varepsilon)(t_h - t), 0), & t \in [t_h + |\varepsilon|, t_h + 2|\varepsilon|], \\ \kappa(t - 2|\varepsilon|), & t \in [t_h + 2|\varepsilon|, 1 + 2|\varepsilon|]. \end{array}$$

It depends on the triple of parameters $\mathscr{E} = (\varepsilon, \eta, h)$, where $\varepsilon, \eta > 0$ and $h \in \mathbb{N}$.



After some computations and the solution (this is really not trivial) of a linear system, we are free to send $k \to \infty$ without modifying the end point but having a gain of length.

6.2 About approach (A2)

Necessary conditions for the minimality of singular extremals can be obtained from the differential study of the end-point map. The theory is well-known till the second order and was initiated by Goh [6] and developed by Agrachev and Sachkov in [2]. Using second order open mapping theorems (index theory), for a strictly singular length minimizing curve γ and for any adjoint curve λ they prove the validity of the following Goh conditions:

$$\langle \lambda(t), [f_i, f_j](\gamma(t)) \rangle = 0, \quad i, j = 1, \dots, m, \tag{6.1}$$

This generalize the first order conditions (5.12) coming from the Pontryagin maximum principle. Partial necessary conditions of the third order are obtained in [4].

Recently, F. Boarotto, R. Monti and me have proved a result generalizing (6.1) to any order n, under a couple of both natural and restrictive assumptions. The statement is the following.

Theorem 6.4 (Boarotto, Monti, S. 2022) *Let $(M, \mathcal{D}, g)$ be a sub-Riemannian manifold, $\gamma = \gamma_u \in AC(I; M)$ be a strictly singular length minimizing curve of corank 1, and assume that*

$$\mathbf{D}_u^h E_{q_0} = 0, \quad h = 2, \dots, n-1. \tag{6.2}$$

Then any adjoint curve $\lambda \in AC(I; T^*M)$ satisfies

$$\langle \lambda(t), [f_{j_n}, [\dots [f_{j_2}, f_{j_1}] \dots]](\gamma(t)) \rangle = 0,$$

for all $t \in I$ and for all $j_1, \dots, j_n = 1, \dots, d$.

Here, the differentials $\mathbf{D}_u^h E_{q_0}$ appearing in (6.2) are the intrinsic differentials of the end point map. In particular, when $n = 1$ the intrinsic differential coincides with the classical one. For $h \geq 2$ it is necessary to restrict the domain $\mathbf{D}_u^h E_{q_0}$ to get the correct definition.

The proof of Theorem 6.4 relies on an open mapping argument applied to the extended end-point map $\bar{E}_{q_0} = (E_{q_0}, J)$, similarly to what happens for the Pontryagin maximum principle. Also in this case, we refer the reader to our preprint article [5] for all the details.

Remark 6.5 We emphasize that, the open mapping argument used in the proof of theorem 6.4 comes from an open mapping theorem of order n, and applies similarly as in Remark 5.10. Also this new open mapping theorem is proved by Boarotto, Monti and me in [5], in a section of indipendent interest.

References

[1] Andrei Agrachev, Davide Barilari, and Ugo Boscain, "A comprehensive introduction to sub-Riemannian geometry". Volume 181 of Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, 2020. With an appendix by Igor Zelenko.

[2] Andrei A. Agrachev and Yuri L. Sachkov, "Control theory from the geometric viewpoint". Volume 87 of Encyclopaedia of Mathematical Sciences. Springer-Verlag, Berlin, 2004. Control Theory and Optimization, II.

[3] D. Barilari, Y. Chitour, F. Jean, D. Prandi, and M. Sigalotti, *On the regularity of abnormal minimizers for rank 2 sub-Riemannian structures.* J. Math. Pures Appl. (9), 133: 118–138, 2020.

[4] Francesco Boarotto, Roberto Monti, and Francesco Palmurella, *Third order open mapping theorems and applications to the end-point map.* Nonlinearity, 33(9):4539–4567, 2020.

[5] Francesco Boarotto, Roberto Monti, and Alessandro Socionovo, *Higher order goh conditions for singular extremals of corank 1.* `https://arxiv.org/abs/2202.00300`, 2022.

[6] B.S. Goh, *Necessary conditions for singular extremals involving multiple control variables.* SIAM J. Control, 4: 716–731, 1966.

[7] Eero Hakavuori, *ODE trajectories as abnormal curves in Carnot groups.* J. Differential Equations, 300:458–486, 2021.

[8] Eero Hakavuori and Enrico Le Donne, *Non-minimality of corners in subriemannian geometry.* Invent. Math., 206(3): 693–704, 2016.

[9] Eero Hakavuori and Enrico Le Donne, *Blowups and blowdowns of geodesics in Carnot groups.* `https://arxiv.org/abs/1806.09375`, 2019.

[10] Enrico Le Donne, Gian Paolo Leonardi, Roberto Monti, and Davide Vittone, *Extremal curves in nilpotent Lie groups.* Geom. Funct. Anal., 23(4): 1371–1401, 2013.

[11] Enrico Le Donne, Gian Paolo Leonardi, Roberto Monti, and Davide Vittone, *Extremal polynomials in stratified groups.* Comm. Anal. Geom., 26(4): 723–757, 2018.

[12] Gian Paolo Leonardi and Roberto Monti, *End-point equations and regularity of sub-Riemannian geodesics.* Geom. Funct. Anal., 18(2): 552–582, 2008.

[13] Wensheng Liu and Héctor J. Sussman, "Shortest paths for sub-Riemannian metrics on rank-two distributions". Mem. Amer. Math. Soc., 118(564): x+104, 1995.

[14] Richard Montgomery, *Abnormal minimizers.* SIAM J. Control Optim., 32(6): 1605–1620, 1994.

[15] Roberto Monti, *Regularity results for sub-riemannian geodesics.* Calc. Var. Partial Differential Equations, 49, 2014.

[16] Roberto Monti, Alessandro Pigati, and Davide Vittone, *Existence of tangent lines to Carnot-Carathéodory geodesics.* Calc. Var. Partial Differential Equations, 57(3): Paper No. 75, 18, 2018.

[17] Roberto Monti and Alessandro Socionovo, *Non-minimality of spirals in sub-Riemannian manifolds.* Calc. Var. Partial Differential Equations, 60(6): Paper No. 218, 20, 2021.