Università di Padova – Dipartimento di Matematica

Scuole di Dottorato in Matematica Pura, Matematica Computazionale e Informatica

Seminario Dottorato 2015/16

Preface

This document offers a large overview of the eight months' schedule of Seminario Dottorato 2015/16. Our "Seminario Dottorato" (Graduate Seminar) is a double-aimed activity. At one hand, the speakers (usually Ph.D. students or post-docs, but sometimes also senior researchers) are invited to think how 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.
Let us take this opportunity to warmly thank the speakers once again, in particular 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, July 2nd, 2016

Corrado Marastoni, Tiziano Vargiolu

Abstracts (from Seminario Dottorato's web page)

Wednesday 7 October 2015

Rare events in finance by PDE methods

Daria GHILLI (Padova, Dip. Mat.)

Rare events, or tails events, are events which happen only ?rarely", in other words, they are situated in the tails of the distribution. Take for example the well-known experiment of tossing a coin: our experience (and also the law of large numbers) says that, after a big enough number of tosses, the most probable value for the empirical mean of the outcomes is 1/2. But what about the probability of being far from 1/2? This is a typical rare event.
The theory who deals with the estimation of tails events is called "large deviations theory" and has many applications, for example, in risk management and finance.
After an introduction to the theory, we consider applications to financial mathematics, concerning the estimation of price of particular type of options (out-of the money) near their maturity. These are typical financial objects whose value deteriorates quickly in time and then are considered, in this context, as rare events. Our approach – mainly of analytical nature – is different from the classical probabilistic ones.

Wednesday 18 November 2015

Learning with Kernels

Michele DONINI (Padova, Dip. Mat.)

To solve a problem on a computer, we need an algorithm, which is a sequence of instructions that should be carried out to transform the input into the output. For some tasks, we do not have an algorithm: we know what the input is, we know what the output should be but we do not know how to transform the input into the output. What we lack in knowledge, we make up for in data. We can exploit data to "learn" using a Machine Learning algorithm, that is able to extract automatically the algorithm for the task.
In this talk, we give an introduction to a family of Machine Learning algorithms called Kernel Methods, starting from a general introduction to the Machine Learning problems and its purposes. After building up the fundamental tools of learning with kernels, we will introduce the principal ideas behind this family of algorithms and its ability to learn automatically using data.

Wednesday 2 December 2015

A simple mathematical model for micro-swimmers

MARTA ZOPPELLO (Padova, Dip. Mat.)

What does it mean swimming? How can mathematics treat this problem? What is the best strategy to move in a certain direction? The study of the swimming strategies of micro-organisms is attracting increasing attention in the recent literature. One of the main difficulties is the complexity of the hydrodynamic forces exerted by the fluid on the swimmer as a reaction to its shape changes. We show that there exists an optimal swimming strategy which leads to minimize the time to reach a desired target. Numerical simulations performed are in good agreement with theoretical predictions and suggest that the optimal strategy is periodic, i.e. composed of a sequence of identical strokes.

Thursday 17 December 2015

Computational social choice: between AI and Economics

ANDREA LOREGGIA (Padova, Dip. Mat.)

During the last decades, the trend has been for disciplines to converge on common techniques to be used in similar problems, besides focusing on specific techniques to be used in narrow domains. AI is one of the best examples: the cross-fertilisation process leads to a very fascinating solutions. Consider for example genetic algorithms, which mimic evolutionary mechanisms to solve search and optimization problems. The individualistic approach of problem solving becomes insufficient: concepts, techniques and experts need to collaborate to get a better understanding of the problems they would like to solve. The techniques that AI makes available are being used by many other disciplines. AI nowadays inundates our everyday life with tools and methods that are hidden in our household electrical devices, smartphones and much more. Starting from the field of multi-agent systems, researchers in AI recently considered the use of models and problems from economics. Notable examples are voting systems used to aggregate the results of several search engines, game theoretic methods that analyse the complex interaction of autonomous agents, and matching procedures implemented on large-scale problems such as the coordination of kidneys transplants and the assignment of students to schools. In this scenario, a number of research lines federated under the name of computational social choice. The need for a computational study of collective decision procedures is clear. On the one hand, from crowdsourcing to university admission ranking, many real-life applications apply existing social choice methods to large scale problems. On the other hand, collective decision-making is not a prerogative of human societies, and multi-agent systems can use these methods to coordinate their actions when facing complex situations. In this talk, we would like to focus on two examples that highlight the impact of a computational approach to classical problems of collective choice. First, by studying repeated decisions (think of opinion polls that precede an election) to evaluate the quality of the result, and, second, by devising innovative procedures to predict the preferences of a collection of individuals.

Wednesday 20 January 2016

Quivers, representations of algebras and beyond

Gabriella D'ESTE (Univ. Milano, Dip. Mat.)

I will illustrate some results obtained by using techniques and general ideas coming from representation theory of finite dimensional algebras. These algebras will almost always be "path algebras" given by quivers, that is oriented graphs, with finitely many vertices and arrows. In less technical words, I will describe some results of applied linear algebra.

Wednesday 17 February 2016

On the behavior of membranes and plates upon perturbations of shape and density

Luigi PROVENZANO (Padova, Dip. Mat.)

In this talk we consider eigenvalue problems for second and fourth order partial differential operators. Such problems arise from the study of the transverse vibrations of thin membranes and plates, respectively. We are interested in the behavior of the normal modes of vibration (i.e., the eigenvalues) upon variations of the shape and the density of the membrane/plate. In particular, we shall consider the issue of the optimization of the eigenvalues depending on such parameters, under suitable constraints (of fixed volume or mass, for example). The talk is of introductory type and is intended for a general audience, no matter the field of expertise.

Wednesday 2 March 2016

Introduction to propagation of chaos for mean-field interacting particle systems

Luisa ANDREIS (Padova, Dip. Mat.)

The purpose of this talk is to give an overview on mean-field interacting particle systems. We will focus on the notion of propagation of chaos, which aims to understand the connection between the microscopic and the macroscopic description of phenomena. Usually, an interacting particle system refers to the microscopic level and a corresponding nonlinear process describes the macroscopic one. In a great number of situations, under hypothesis on the symmetry of the system and on the type of interaction, the link between these two levels is precisely given by propagation of chaos. Since the talk is intended for a general audience, we start by recalling basic definitions and results of Probability. Then we introduce the basic concepts of the theory, by means of classical examples as well as recent ones.

Wednesday 16 March 2016

Cosheaves, an introduction

Pietro POLESELLO (Padova, Dip. Mat.)

It is well known that locally defined distributions glue together, that is, they define a sheaf. In fact, this follows immediately from the fact that test functions (i.e. smooth functions with compact support) form a cosheaf, which is the dual notion of a sheaf.
By definition, cosheaves on a space X and with values in category $\mathcal{C}$ are dual to sheaves on X with values in the opposite category $\mathcal{C}^{\mathrm{op}}$. For this reason, cosheaves did not attract much attention, being considered as part of sheaf theory. However, passing from $\mathcal{C}$ to $\mathcal{C}^{\mathrm{op}}$, may cause difficulties, as in general $\mathcal{C}$ and $\mathcal{C}^{\mathrm{op}}$ do not share the same good properties needed for sheaf theory (*e.g.* colimits are not exact in $\mathrm{Ab}^{\mathrm{op}}$, the opposite of category of abelian groups). Moreover, dealing with cosheaves may be more convenient, as they appear naturally in analysis (as the compactly supported sections of c-soft sheaves, such as smooth functions or distributions), in algebraic analysis (*e.g.* as the subanalytic cosheaf of Schwartz functions), in topology (in relation with Fox's theory of topological branched coverings), and in tops theory. Moreover, as sheaves are the natural coefficient spaces for cohomology theories, cosheaves play the same role for homology theories, such as Čech homology, and they are (hidden) ingredients of Poincaré duality (recently, ∞-cosheaves infiltrated Poincaré-Verdier duality in the context of Lurie's *higher topos theory*).
In this seminar, I will give a brief introduction to cosheaves, giving examples and explaining the relation with sheaves, with Lawvere's distributions and with Fox's branched coverings.

Wednesday 13 April 2016

Cheapest Routes with Integer Linear Programming

Michele BARBATO (LIPN, Univ. Paris 13, France)

Combinatorial Optimization deals with the optimization of a function over a finite, but huge, set of elements. It has a great impact on real life, as several problems arising in logistics, scheduling, facility location, to cite a few, can be stated as Combinatorial Optimization problems. Often problems of this kind can be expressed as Integer Linear Programs (ILP), i.e., problems in which the function to be optimized is linear and so are the constraints that define the feasibility set. In the first part of the talk, we provide an introductory presentation of some well-established methods in Integer Linear Programming. These methods are presented through examples that, in several cases, also motivate theoretical questions (e.g., the polyhedral study). We will consider as initial case of study the Traveling Salesman Problem (TSP). The TSP consists in finding the cheapest route that visits a prescribed set of cities exactly once, before returning to the starting point. As such, the TSP is a prototype of several other problems arising in logistics. In the second part of

the presentation we will talk about the Double Traveling Salesman Problem with Multiple Stacks, that combines the construction of a cheapest route with loading constraints. We will reveal links between this problem and the TSP, as well as the limitations that a purely routing-based approach has for this problem.

Wednesday 4 May 2016

Isoperimetric inequalities in Carnot-Carathéodory spaces

Valentina FRANCESCHI (Padova, Dip. Mat.)

One of the most ancient mathematical problems is Dido's problem, appearing in Virgil's *Aeneid*: what is the shape to give to a rope in order to enclose a maximal region of land? The expected solution is of course the circle. Despite the ancient origins, a rigorous mathematical formulation and solution is quite recent, dating back to the 1950s when Caccioppoli and De Giorgi introduced the notion of perimeter in the n-dimensional Euclidean space. The latter notion led to the study of isoperimetric inequalities and to the solution of Dido's problem generalized to n dimensions. Mathematicians then generalized isoperimetric inequalities to different frameworks, such as riemannian manifolds and metric spaces. After an overview of the classical definitions, in this talk, we present isoperimetric inequalities in a class of metric spaces arising from the study of hypoelliptic differential operators, called Carnot-Carathéodory spaces. We conclude presenting the main conjecture in this framework (Pansu's conjecture) ad some related results.

Wednesday 25 May 2016

Fractional Calculus: Numerical Methods and Models

Abdelsheed Ismail GAD AMEEN (Padova, Dip. Mat.)

In this talk, we first give a short introduction of fractional calculus (FC) and its geometrical, physical interpretation. Then, we discuss the differential equations of fractional order (Caputo type) which have recently proved to be valuable tools for modeling of many biological phenomena. Most of fractional ordinary differential equations (FODEs) do not have exact analytic solutions so that numerical techniques must be used. Hence, we present the fractional Euler method to solve systems of nonlinear FODEs and show how to use this method for solving the Susceptible-Infected-Recovered (SIR) model of fractional order.

Wednesday 1 June 2016

Polyhedral structures in algebraic geometry

Stefano URBINATI (Padova, Dip. Mat.)

Algebraic geometry studies the zero locus of polynomial equations connecting the related algebraic and geometrical structures. In several cases, nevertheless the theory is extremely precise and elegant, it is hard to read in a simple way the information behind such structures. A possible way of avoiding this problem is that of associating to polynomials some polyhedral structures that immediately give some of the information connected to the zero locus of the polynomial. In relation to this strategy I will introduce Newton-Okounkov bodies and Tropical Geometry.

Wednesday 15 June 2016

Computed Tomography: a real case example of inverse problem

Elena MOROTTI (Padova, Dip. Mat.)

X-ray computed tomography (CT) is a well known medical imaging technique, that seeks to reveal internal structures hidden by the skin and bones. Mathematically, the CT process can be modelled as a linear system and the image reconstruction is a challenging inverse problem. In this talk I will show both phisical and mathematical basic concepts, to explain the CT process, and the two possible approaches to solve the problem (leading to analitical or iterative numerical methods). Finally, I will shortly introduce the Digital Breast Tomosynthesis (DBT) technology, that is a 3D emerging technique for the diagnosis of breast tumors, together with numerical results for a simulated problem.

Rare events in finance by PDE methods

Daria Ghilli (*)

Abstract. Rare events, or tails events, are events which happen only ?rarely", in other words, they are situated in the tails of the distribution. Take for example the well-known experiment of tossing a coin: our experience (and also the law of large numbers) says that, after a big enough number of tosses, the most probable value for the empirical mean of the outcomes is 1/2. But what about the probability of being far from 1/2? This is a typical rare event.
The theory who deals with the estimation of tails events is called "large deviations theory" and has many applications, for example, in risk management and finance.
After an introduction to the theory, we consider applications to financial mathematics, concerning the estimation of price of particular type of options (out-of the money) near their maturity. These are typical financial objects whose value deteriorates quickly in time and then are considered, in this context, as rare events. Our approach – mainly of analytical nature – is different from the classical probabilistic ones.

Rare events, or tails events, are events which happens only "rarely", in other words, they are situated in the tails of the distribution. The theory who deals with the estimation of tails events is called *large deviations theory*. Roughly speaking, large deviations theory concerns itself with the exponential decline of the probability measures of certain kinds of extreme or tail events. Some basic ideas of the theory can be traced back to Laplace and Cramér, but a clear and unified formal definition was only introduced in 1966, in a paper by Varadhan.

What is precisely a rare event? Take for example the well-known experiment of tossing a coin. Our experience (and also the law of large numbers) says that, after a big enough number of tosses, the most probable value for the empirical mean of the outcomes is $\frac{1}{2}$. But what about the probability of being far from $\frac{1}{2}$? This is a typical rare event.

Let us desribe the example more precisely. Consider a sequence of independent tosses of a fair coin. The possible outcomes could be heads or tails. Let us denote the possible outcome of the i-th trial by X_i, where we encode head as 1 and tail as 0. Now let M_N denote the mean value after N trials, namely $M_N = \frac{1}{N}\sum_{i=1}^{N} X_i$. Then M_N lies between 0 and 1. From the law of large numbers (and also from our experience) we know that as N grows, the distribution of M_N converges to $0.5 = \mathrm{E}[X_1]$ (the expectation value of a single coin toss) almost surely. Moreover, by the central limit theorem, we know that M_N is

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: daria.ghilli@gmail.com . Seminar held on October 7th, 2015.

approximately normally distributed for large N. The central limit theorem can provide more detailed information about the behavior of M_N than the law of large numbers. For example, we can approximately find a tail probability of $M_N, P(M_N > x)$, that M_N is greater than x, for a fixed value of N. However, the approximation by the CLT may not be accurate if x is far from $\mathrm{E}[X_1]$. Also, it does not provide information about the convergence of the tail probabilities as $N \to \infty$. However, the large deviation theory can provide answers for such problems.

Indeed, for a given value $0.5 < x < 1$, let us compute the tail probability $P(M_N > x)$. Define $I(x) = x\ln x + (1-x)\ln(1-x) + \ln 2$. Note that the function $I(x)$ is a convex, nonnegative function that is zero at $x = \frac{1}{2}$ and increases as you move to $x = 1$. By Chernoff's inequality, it can be shown that $P(M_N > x) < e^{-NI(x)}$. This bound is rather sharp, in the sense that $I(x)$ cannot be replaced with a larger number which would yield a strict inequality for all positive N.

Hence, we obtain the following result: $P(M_N > x) \approx e^{-NI(x)}$. The probability $P(M_N > x)$ decays exponentially as N grows to infinity, at a rate depending on x. This is an example of the typical results provided by large deviation theory.

The theory of large deviations has many application, for example in risk management and finance

An interesting application to risk management goes back directly to the birth of the theory. Indeed, the first rigorous results concerning large deviations are due to the Swedish mathematician Harald Cramér, who applied them to model the insurance business. From the point of view of an insurance company, the earning is at a constant rate per month (the monthly premium) but the claims come randomly. For the company to be successful over a certain period of time (preferably many months), the total earning should exceed the total claim. Thus to estimate the premium you have to ask the following question : "What should we choose as the premium q such that over N months the total claim $C = \sum_i X_i$ should be less than Nq?" This is clearly the same question asked by the large deviations theory.

Cramér gave a solution to this question, proving the so-called Cramér's theorem, which can be considered as the first basic result of large deviations theory.

Let $X_1, X_2, \cdots$ be a sequence of bounded independent and identically distributed (i.i.d) random variables. Then Cramér's theorem states that the following limit exists: $\lim_{N\to\infty} \frac{1}{N} \ln P(M_N > x) = -I(x)$. (or equivalently $P(M_N > x) \approx e^{-NI(x)}$. for large N), where the function $I(\cdot)$ is called the "rate function" or "Cramér function". Also, if we know the probability distribution of X, an explicit expression for the rate function can be obtained. For example, if X follows a normal distribution, the rate function becomes a parabola with its apex at the mean of the normal distribution.

As already mentioned, large deviation theory has many applications also to mathematical finance. In the following we present some applications to the estimation of price of particular type of options, so-called out-of-the-money, near their maturity. An option, say a call option, on a stock S with maturity T and stock price K is a sort of contract which gives you the right to buy a certain amount of the stock S in the date T at a fixed price K. We consider a particular kind of call options, namely out-of the money call options, meaning that $S_0 < K$. We also consider their near-maturity value meaning that we take

$T = \varepsilon t$ for small values of ε. These are tipically financial objects whose value tends to deteriorate quickly in time. In this context they can be considered as "rare events". In particular their value can be estimated as a consequence of a large deviation principle for the stochastic process, say S_t, which models the dynamic of the stock price.

These results have been carried out in collaboration with Martino Bardi and Annalisa Cesaroni of the Univeristy of Padua and represent the first part of my phD thesis. We remark that our approach is different from the classical probabilistic one and we use methods taken mainly from the theory of nonlinear partial differential equations. The techniques are mainly from viscosity solutions theory and homogenization of Hamilton-Jacobi-Bellman equations.

First we introduce our approach and explain our main result. Then we describe the main application to large deviations and asymptotic estimates of option prices near maturity. We refer to the end of this report for more details and the precise statement of the estimate.

We are interested in stochastic differential equations with two small parameters $\varepsilon > 0$ and $\delta > 0$ of the form

$$\begin{cases} dX_t = \varepsilon\phi(X_t, Y_t)dt + \sqrt{2\varepsilon}\sigma(X_t, Y_t)dW_t & X_0 = x \in \mathbb{R}^n, \\ dY_t = \frac{\varepsilon}{\delta}b(Y_t)dt + \sqrt{\frac{2\varepsilon}{\delta}}\tau(Y_t)dW_t & Y_0 = y \in \mathbb{R}^m, \end{cases} \tag{1}$$

where W_t is a standard r-dimensional Brownian motion, and the matrix τ is non-degenerate.

We remark that this is a model of systems where the variables Y_t evolve at a much faster time scale, namely $s = \frac{t}{\delta}$, than the other variables X_t. Note also that the second parameter ε is added in order to study the small time behavior of the system, by this we mean that the time has been rescaled in (1) as $t \mapsto \varepsilon t$.

Passing to the limit as $\delta \to 0$, with ε fixed, is a classical singular perturbation problem. This means that the solution of the limit problem leads to the elimination of the state variable Y_t and the reduction of the dimension of the system from $n + m$ to n and to the definition of an averaged limit system defined in $\mathbb{R}^n$ only. Of course the limit problem keeps some informations on the fast part of the system.

There is a large mathematical literature on singular perturbation problems, both in the deterministic ($\sigma \equiv 0, \tau \equiv 0$) and in the stochastic case. In particular we mention some mathematical contributions most related to our work and methods, such as [9] and [2].

Our first motivation for the study of systems of the form (1) comes from financial models with stochastic volatility. In such models the vector X_t represents the log-prices of n assets (under a risk-neutral probability measure) whose volatility σ is affected by a process Y_t driven by another Brownian motion, which is often negatively correlated with the one driving the stock prices (this is the empirically observed leverage effect, i.e., asset prices tend to go down as volatility goes up).

An important extension of the stochastic volatility approach was introduced recently by Fouque, Papanicolaou, and Sircar in the book [6]. The idea is trying to describe the bursty behavior of volatility: in empirical observations volatility often tends to fluctuate to a high level for a while, then to a low level for another small time period, then again at high level, and so on, for several times during the life of a derivative contract. These

phenomena are also related to another feature of stochastic volatility, which is mean reversion. A mathematical framework which takes into account both bursting and mean reverting behavior of the volatility is that of multiple time scale systems and singular perturbations. In this setting volatility is modeled as a process which evolves on a faster time scale than the asset prices and which is ergodic, in the sense that it has a unique invariant distribution (the long-run distribution) and asymptotically decorellates (in the sense that it becomes independent of the initial distribution). We refer the reader to the book [6] and to the references therein for a detailed presentation of these models and for their empirical justification. Several extensions, applications to a variety of financial problems, and rigorous justifications of the asymptotics can be found in [8, 3] and may others .

According to the previous discussion, stochastic systems of the form (1), under some suitable assumptions implying ergodicity of the Y_t process, are appropriate for studying financial problems in this setting. Indeed, here the slow variables represent prices of assets or the wealth of the investor, whereas Y_t is an ergodic process representing the volatility and evolving on a faster time scale for δ small (and ε fixed).

Let us enter more into details and describe our main results. Our aim is to study the asymptotic behavior of the system (1) as both the parameters go to 0 and we expect different limit behaviors depending on the rate ε/δ. Therefore we put $\delta = \varepsilon^\alpha$, with $\alpha > 1$. We consider a functional of the trajectories of (1) of the form

$$v^\varepsilon(t,x,y) := \varepsilon \log E\Big[e^{h(X_t)/\varepsilon} | (X.,Y.) \text{ satisfy (1)}\Big],$$

where h is a bounded continuous function. We observe that the logarithmic form of this payoff is motivated by the applications to large deviations that we want to give.

The starting point of our methods is the well-known fact that v^ε solves the Cauchy problem with initial data $v^\varepsilon(0,x,y) = h(x)$ for a fully nonlinear parabolic equation. Letting $\varepsilon \to 0$ in this PDE is a regular perturbation of a singular perturbation problem for an nonlinear PDE of Hamilton-Jacobi-Bellman type.

The main result we carried out is a convergence theorem for the functionals v^ε. We proved that, under suitable assumptions,the functions $v^\varepsilon(t,x,y)$ converge to a function $v(t,x)$ characterised as the solution of the Cauchy problem for a first order Hamilton-Jacobi equation

$$v_t - \bar{H}(x, Dv) = 0 \text{ in }]0,T[\times\mathbb{R}^n, \quad v(0,x) = h(x). \tag{2}$$

We remark again that this kind of problem is a singular perturbation problem where the fast variable y lives in $\mathbb{R}^m$. The resolution of this limit problem is not banal and the main difficulties is the unboundedness of the fast variable. The existing techniques to treat this kind of problems have been developped so far mainly under assumptions implying some kind of boundness of the fast variable. The methods stem from Evans' perturbed test function method for homogenization [5] and its extensions to singular perturbations, see mainly [1, 2].

A classical hypothesisis is the periodicity with respect to Y_t of the coefficients of the stochastic system, which in particular implies the periodicity in y of the solutions v^ε. In the paper [4] we carry out our analysis under this main assumption, treating what we call the *periodic case*.

In the periodic case, the convergence is quite standard once the effective problem is identified and a comparison principle is proved. The most significant part of paper [4] is the identification of the *effective Hamiltonian* $\bar{H}$, for which we provide also representation formulas.

Then, we considered a particular type of fast mean-reverting stochastic volatility systems as in (1) where the fast variable is unbounded and actually lives in $\mathbb{R}^m$. Mainly because of the difficulties caused by the unboundedness of the y, we managed to trat processed mainly of Orstein-Uhlenbeck type, that is $Y_t = (\mu - Y_t)dt + \tau(Y_t)dw_t$ where $\mu \in \mathbb{R}^m$ is a vector, and τ is bounded and uniformly non degenerate. Note that the non compactness is replaced by ergodicity, i.e that the Y_t process has a unique invariant distribution (the long-run distribution) and that in the long term it becomes independent of the initial distribution.

The motivation behind the analysis of such kind of systems relies in the fact that the assumption of periodicity of the paper [4] seems a bit restrictive for the financial applications we have in mind, in particular it does not appear natural in order to model volatility in financial markets, according to the empirical data and in the discussion presented in [6] and the references therein.

The main application of the convergence results is a large deviations analysis of (1). We prove that the measures associated to the process X_t in (1) satisfy a Large Deviation Principle (briefly, LDP) with good rate function

$$I(x;x_0,t) := \inf\left[\int_0^t \bar{L}\left(\xi(s),\xi'(s)\right)\,ds \;\middle|\; \xi \in AC(0,t),\ \xi(0)=x_0, \xi(t)=x\right], \tag{3}$$

where $\bar{L}$ is the *effective Lagrangian* associated to $\bar{H}$ via convex duality. In particular we get that

$$P(X_t^\varepsilon \in B) = e^{-\inf_{x\in B}\frac{I(x;x_0,t)}{\varepsilon}+o(\frac{1}{\varepsilon})}, \quad \text{as } \varepsilon \to 0 \tag{4}$$

for any open set $B \subseteq \mathbb{R}^n$.

Then, the asymptotic estimate of the price of out of the money call option with strike price K and short maturity time $T = \varepsilon t$ comes as a direct consequence of the large deviation analysis. Note that this applications are derived following mainly the approach of [7], where similar kind of problems have been investigated under different scalings and usign different methods.

Let S_t^ε be the asset price evolving accordingly the following differential system

$$\begin{cases} dS_t^\varepsilon = \varepsilon\xi(S_t^\varepsilon, Y_t)dt + \sqrt{2\varepsilon}\zeta(S_t^\varepsilon, Y_t)S_t^\varepsilon dW_t & S_0^\varepsilon = S_0 \in \mathbb{R}^n, \\ dY_t = \frac{\varepsilon}{\delta}b(Y_t)dt + \sqrt{\frac{2\varepsilon}{\delta}}\tau(Y_t)dW_t & Y_0 = y \in \mathbb{R}^m, \end{cases}$$

where $\alpha > 1$, τ, b are as in (1) and $\xi : \mathbb{R}^+ \times \mathbb{R}^m \to \mathbb{R}, \zeta : \mathbb{R}^+ \times \mathbb{R}^m \to \mathbf{M}^r$ are Lipschitz continuous bounded functions.

We define $X_t^\varepsilon = \log S_t^\varepsilon$. Then $(X_t^\varepsilon, Y_t^\varepsilon)$ satisfies (1) with $\phi(x,y) = \xi(e^x,y) - \zeta(e^x,y)\zeta^T(e^x,y)$ and $\sigma(x,y) = \zeta(e^x,y)$. We consider out-of-the-money call option by taking $S_0 < K$ or $x_0 < logK$.

Then, as an application of the large deviation principle (4), we have the following asymptotic estimate

$$\lim_{\varepsilon\to 0} \varepsilon \log E\left[(S_t^\varepsilon - K)^+\right] = -\inf_{y>\log K} I(y; x_0, t),$$

where I is the (positive and continuous) rate function defined in (3).

References

[1] O. Alvarez, M. Bardi, *Singular Perturbation of Nonlinear Degenerate Parabolic PDEs: a General Convergence Result.* Arch. Rational Mech. Anal. 170 (2003), 17–61.

[2] O. Alvarez, M. Bardi, "Ergodicity, stabilization, and singular perturbations for Bellman-Isaacs equations". Mem. Amer. Math. Soc. 960, 2010, vi+77 pp..

[3] M. Bardi, A. Cesaroni, L. Manca, *Convergence by viscosity methods in multiscale financial models with stochastic volatility.* Siam J. Financial Math. 1 (2010), no.1, 230–265.

[4] M. Bardi, A. Cesaroni, D. Ghilli, *Large deviations for some fast stochastic volatility models by viscosity methods.* Discrete Contin. Dyn. Syst. A. 35 (2015), n. 9, 3965–3988.

[5] L. C. Evans, *The perturbed test function method for viscosity solutions of nonlinear PDE.* Proc. Roy. Soc. Edinburgh Sect. A 111 (1989), no. 3-4, 359–375.

[6] J.-P. Fouque, G. Papanicolaou, K. R. Sircar, "Derivatives in financial markets with stochastic volatility". Cambridge University Press, Cambridge, 2000.

[7] J. Feng, J.-P. Fouque, R. Kumar, "Small time asymptotic for fats mean-reverting stochstic volatility models". The Annals of Applied Probability 22 (2012), No. 4, 1541–1575.

[8] J.-P. Fouque, G. Papanicolaou, R. Sircar, K. Solna, *Singular perturbations in option pricing.* SIAM J. Appl. Math. 63 (2003), no. 5, 1648–1665.

[9] Y. Kabanov, S. Pergamenshchikov, "Two-scale stochastic systems. Asymptotic analysis and control". Springer-Verlag, Berlin, 2003.

Learning with Kernels

Michele Donini [(*)]

Abstract. To solve a problem on a computer, we need an algorithm, which is a sequence of instructions that should be carried out to transform the input into the output. For some tasks, we do not have an algorithm: we know what the input is, we know what the output should be but we do not know how to transform the input into the output. What we lack in knowledge, we make up for in data. We can exploit data to "learn" using a Machine Learning algorithm, that is able to extract automatically the algorithm for the task.
In this article, we give an introduction to a family of Machine Learning algorithms called Kernel Methods, starting from a general introduction to the Machine Learning problems and its purposes. After building up the fundamental tools of learning with kernels, we will introduce the principal ideas behind this family of algorithms and its ability to learn automatically using data.

1 Learning from data

Over the past three decades, research on machine learning and data mining has led to a wide variety of algorithms that induce general functions from examples. As machine learning is maturing, it has begun to make the successful transition from academic research to various practical applications. Generic techniques are now being used in various commercial and industrial applications.[(†)]

Specifically, machine learning's purpose is to create algorithms able to optimize a performance criterion using examples from data or from past experience [2]. Given a model defined by some parameters, learning is the execution of a method to optimize them exploiting the training data. The model obtained should be able to make predictions in the future or to extract knowledge from data.

The theory of statistics is the most important tool exploited by machine learning in order to build mathematical models, since one of machine learning crucial tasks is making inference from a sample. Inference is the process of deriving logical conclusions from premises known, or assumed to be true (i.e. training examples). On the other hand, machine learning is more than only inference from a set of past experiences.

In fact, computer science is the second tool exploited by machine learning with two principal roles. Firstly, in training, where the efficiency of the algorithms is fundamental

[(*)]Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: m.donini@ucl.ac.uk . Seminar held on November 18th, 2015.

[(†)]Sebastian Thrun, founder of Google X Research Laboratory

to solve the optimization problem and to store the solutions. Secondly, once we have a model, its representation and solution for inference needs to be efficient as well.

When we have to deal with a machine learning problem, the first step is to define a representation of the task, i.e. we have to select how to describe the known information. Typically, this pre-training step is performed to describe the problem that we have to solve in the best way possible by defining a set of features (explicitly or implicitly). In real world applications, the efficiency of the learning and the space and time complexity required by the methods have the same importance as to the predictive accuracy of the model.

1.2 Binary classification task

Supervised classification task has the goal to learn how to clissify an object (selecting a solution from a finite set of possible *labels*). The learning has to be performed using a set of examples called training set (i.e. a set of pairs example-label). Considering the binary classification task, we define the training set as

$$\{(x_i, y_i)\}_{i=1}^{l}, \tag{1}$$

and test set as

$$\{(x_i, y_i)\}_{i=l+1}^{L}, \tag{2}$$

where the example x_i is in a generic set $\mathcal{X}$ and y_i with values +1 or −1. Finally, the test set is used for a fairly evaluation of the performance of a specific algorithm.

2 Kernels

In this section will be presented a brief introduction concerning kernel functions in machine learning. In a general space $\mathcal{X}$, a kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive semi-definite function and represents a dot product in an implicitly defined Hilbert space X (a.k.a., feature space). K represents a similarity measure between the elements in $\mathcal{X}$. Given a kernel K, its feature mapping is a (typically non linear) embedding $\boldsymbol{\phi} : \mathcal{X} \to X$. The kernel K can be written in the form:

$$K(x, y) = \boldsymbol{\phi}(x) \cdot \boldsymbol{\phi}(y), \quad \forall x, y \in \mathcal{X}. \tag{3}$$

So, given a kernel, the explicit evaluation of the vector $\boldsymbol{\phi}(x)$ is avoidable and we are able to obtain a significant improvement in the performance of the kernel methods.

A formal definition of a kernel function requires the finitely positive semi-definite property.

Definition 1 A function

$$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

satisfies the finitely positive semi-definite property if it is a symmetric function for which the matrices formed by restriction to any finite subset of the space X are positive semi-definite.

Note that this definition does not require the set $\mathcal{X}$ to be a vector space. Now, we are ready to introduce the characterisation theorem for kernels.

Theorem 1 *A function*

$$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

which is either continuous or has a finite domain, can be decomposed

$$K(x, y) = \boldsymbol{\phi}(x) \cdot \boldsymbol{\phi}(y), \quad \forall x, y \in \mathcal{X}.$$

into a feature map $\boldsymbol{\phi}$ into a Hilbert space F applied to both its arguments followed by the evaluation of the inner product in F if and only if it satisfies the finitely positive semi-definite property.

A proof of this theorem can be found in [10]. The space F is called the Reproducing Kernel Hilbert Space (RKHS) of the kernel function K. Given a set of vectors, $S = \{x_1, ..., x_L\} \subset \mathcal{X}$, the Gram matrix is defined as the $L \times L$ matrix G whose entries are $G_{ij} = \mathbf{x}_i \cdot \mathbf{x}_j$. If we are using a kernel function K to evaluate the inner products in a feature space with feature map $\boldsymbol{\phi}$, the associated Gram matrix has entries $G_{ij} = \boldsymbol{\phi}(\mathbf{x}_i) \cdot \boldsymbol{\phi}(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j)$. In this case the matrix is often referred to as the kernel matrix.

In the following, the matrix $\mathbf{K} \in \mathbb{R}^{L \times L}$ is the complete kernel (Gram) matrix containing the values of the kernel of each (training and test) data pair. Further, we indicate with an hat, like for example $\hat{\mathbf{y}} \in \mathbb{R}^l$ or $\hat{\mathbf{K}} \in \mathbb{R}^{l \times l}$, the submatrices (or subvectors) obtained considering training examples only.

2.1 KOMD: a kernel machine

The kernel functions can be used in different learning algorithms. In this section we present one of them, starting from the following definition:

Definition 2 Given a training set, we consider the domain $\hat{\Gamma}$ of probability distributions $\boldsymbol{\gamma} \in \mathbb{R}^l_+$ defined over the sets of positive and negative training examples as $\hat{\Gamma} = \{\boldsymbol{\gamma} \in \mathbb{R}^l_+ \mid \sum_{i \in \oplus} \gamma_i = 1, \sum_{i \in \ominus} \gamma_i = 1\}$.

Note that any element $\boldsymbol{\gamma} \in \hat{\Gamma}$ corresponds to a pair of points, the first in the convex hull of positive training examples and the second in the convex hull of negative training examples.

In [1] a game theoretic interpretation of the problem of margin maximization and an algorithm called "Kernel method for the Optimization of the Margin Distribution" (KOMD) have been proposed. The classification task is posed as a two-player zero-sum game and it is shown how this zero-sum game can be solved efficiently by optimizing a simple linearly constrained convex function on variables $\boldsymbol{\gamma} \in \hat{\Gamma}$, namely,

$$\text{minimize}_{\boldsymbol{\gamma} \in \hat{\Gamma}} D(\gamma) := \boldsymbol{\gamma}^\top \hat{\mathbf{Y}} \hat{\mathbf{K}} \hat{\mathbf{Y}} \boldsymbol{\gamma}.$$

The vector $\boldsymbol{\gamma}^* \in \hat{\Gamma}$ that minimizes $D(\boldsymbol{\gamma})$ identifies the two nearest points in the convex hulls of positive and negative examples, in the feature space of kernel. Moreover, a quadratic regularization over γ, namely $R_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}) = \boldsymbol{\gamma} \cdot \boldsymbol{\gamma}$, is introduced, that makes the player to prefer

optimal distributions (strategies) with low variance. The final best strategy for $\boldsymbol{\gamma}$ will be given by solving the optimization problem $\min_{\boldsymbol{\gamma}\in\hat{\Gamma}} (1-\lambda)D(\boldsymbol{\gamma}) + \lambda R_{\boldsymbol{\gamma}}(\boldsymbol{\gamma})$. A correct selection of the parameter $\lambda \in [0,1]$ (usually made by validating on training data) is fundamental. In KOMD, the evaluation on a new generic example $\mathbf{x}$ is obtained by: $f(\mathbf{x}) = \sum_i y_i \gamma_i \mathbf{K}(\mathbf{x}_i, \mathbf{x}) = \mathbf{K}_{tr}(\mathbf{x})\hat{\mathbf{Y}}\boldsymbol{\gamma}$, where $\mathbf{K}_{tr}(\mathbf{x}) = [\mathbf{K}(\mathbf{x}_1, \mathbf{x}), \ldots, \mathbf{K}(\mathbf{x}_l, \mathbf{x})]^\top$.

Once the model is learned from training data, the evaluation on a new generic example $\mathbf{x}$ is obtained by:

$$f(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x}) = \sum_i y_i \boldsymbol{\gamma}_i \mathbf{K}(\mathbf{x}_i, \mathbf{x}) = \mathbf{K}_{tr}(\mathbf{x})\hat{\mathbf{Y}}\boldsymbol{\gamma},$$

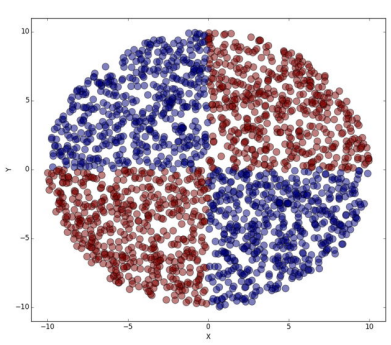
where $\mathbf{K}_{tr}(\mathbf{x}) = [\mathbf{K}(\mathbf{x}_1, \mathbf{x}), \ldots, \mathbf{K}(\mathbf{x}_l, \mathbf{x})]^\top$, i.e. the vector containing the kernel values with the training examples for $\mathbf{x}$.

When a pure binary classification is required, which is not the case in our work, then the second phase of the algorithm is also performed. The threshold is set corresponding to the point standing in the middle between the optimal point in the convex hull of positive examples and the one in the convex hull of negative examples, that is

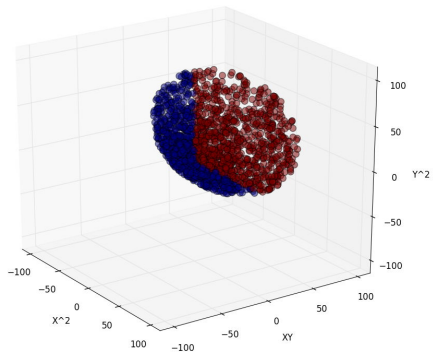
$$\theta = \frac{1}{2}\left(\sum_{i\in\oplus} \boldsymbol{\gamma}_i f(\mathbf{x}_i) + \sum_{i\in\ominus} \boldsymbol{\gamma}_i f(\mathbf{x}_i)\right) = \frac{1}{2}\sum_i \boldsymbol{\gamma}_i f(\mathbf{x}_i) = \frac{1}{2}\boldsymbol{\gamma}^\top \hat{\mathbf{K}}\hat{\mathbf{Y}}\boldsymbol{\gamma}.$$

2.2 Famous kernels

In this section we exploit the KOMD kernel machine with different famous kernel functions in a binary classification task. The task is an artificial task created with the cardinality of the training set equals to $l = 10$ and the cardinality of the whole dataset equals to $L = 1000$ examples. The examples are generated uniformly over the area of the circle of radius 10, centerd in $(0,0) \in \mathbb{R}^2$. The labels are assigned in order to simulate a logic XOR, as presented in Figure 1a. In Table 1, linear, polynomial, and RBF kernels are presented. The RKHS of the polynomial kernel with the degree $d = 2$ and the coefficient $c = 0$ is depicted in Figure 1b.



(a) The original data.



(b) The data in the RKHS of a polynomial kernel with $d = 2$ and $c = 0$.

Figure 1. Artificial XOR binary classification task. In red the positive examples, in blue the negative examples.

Kernel	Formula	Parameters
Linear	$\mathbf{x} \cdot \mathbf{z}$	-
Polynomial ($K_{D,c}$)	$(\mathbf{x} \cdot \mathbf{z} + c)^D$	$c \geq 0, D \in \mathbb{N}$
RBF (K^{γ}_{RBF})	$e^{-\gamma \|\mathbf{x}-\mathbf{z}\|^2}$	$\gamma \geq 0$

Table 1. Classical kernels with formula and parameters.

Finally, in Table 2 the results of the accuracy of the presented kernels are presented for the XOR binary classification task. The results are the average among several different selections of training and test sets. From these results, we can claim that the linear kernel is not able to learn a correct model for this task. Note that the polynomial kernel with degree $d = 1$ is equivalent to the linear kernel. These two kernels generate models that have too low expressiveness to solve our problem. On the other hand, when the degree d of the polynomial kernel is too high the generated model is too complex and is not able to generalize (i.e. to transfer the information from the training set to the test set in the correct way). The same behavior is obtained for value of γ that are too small (i.e. too simple) or too high (i.e. too complex). Finding the best parameter is a key step in order to use a correct kernel function and, consequently, to represent correctly a specific task.

Kernel	Parameters	Accuracy $\in [0, 1]$
Linear	-	0.53
Polynomial	$c = 0, D = 1$	0.53
Polynomial	$c = 1, D = 1$	0.53
Polynomial	$c = 0, D = 2$	**0.98**
Polynomial	$c = 1, D = 2$	**0.95**
Polynomial	$c = 1, D = 10$	0.69
RBF	$\gamma = 1$	0.63
RBF	$\gamma = 10$	**0.91**
RBF	$\gamma = 100$	0.56

Table 2. Accuracy of different kernels tackling the XOR classification task..

3 The representation problem

As seen in the previous section, the data representation plays a key role in the success of machine learning methods. Due to the current data growth in size, heterogeneity and structure, the new generation of algorithms are expected to solve increasingly challenging problems. This must be done under growing constraints such as computational resources, memory budget and energy consumption. The representation should ideally distill the

relevant information about a learning problem in a compact manner, such that it becomes possible to learn the model from a small number of examples.

In the past, the research community has been focused on investigating new algorithms to obtain a model from a fixed a *priori* representation. In fact, the learning process was considered mainly the process of choosing an appropriate function from a given set of functions [12].

A new point of view is arising in this last decade, and the problem of learning the optimal representation has become a hot topic in the most important conferences and journals. When dealing with the representation of a task, a plethora of questions arise. Some questions are simple to be formulated but the answers are not easy to be found. The first questions are: how can we find automatically a good representation for a specific task? How can we compare two representations to assert that one is better than another?

3.1 Kernels as representations

Kernels have consistently outperformed previous generations of learning techniques because they provided a flexible and expressive learning framework that has been successfully applied to a wide range of real world problems. For example, kernel methods are widely applied in machine learning for structured data because, unlike the majority of machine learning techniques, their application to any type of data is painless as long as a kernel function for such data is defined. Kernel methods offer an elegant framework that decouples learning algorithms from data representations. On the other hand, kernels have lost some of their initial appealing in the research community, due to some of their weaknesses:

- The scaling problem and the high computational complexity: dealing with kernel based methods imposes the storage in memory of a kernel matrix, i.e. a matrix with a number of entries quadratic with respect to the number of examples. So, kernel methods are not able (in general) to tackle a task when the number of examples becomes huge;
- The shallowness of the representation: the representation implicitly defined by a shallow kernel does not take into account several layer of abstraction and is fixed *a priori*.

Moreover, the so called local kernels suffer the *curse of dimensionality* [4], i.e. the problem of learning a model in a high dimensional space. This concept was coined by Bellman [3] and it refers to the exponential growth of hyper-volume as a function of dimensionality. Formally, we consider a local kernel K a kernel function with the following behavior:

$$\lim_{\|\mathbf{x}-\mathbf{x}_i\|_2 \to +\infty} K(\mathbf{x}, \mathbf{x}_i) = C_i, \tag{4}$$

where $\mathbf{x}$ is a test example, $\mathbf{x}_i$ is a training example and C_i is a constant that does not depend on $\mathbf{x}$. For example, in the RBF kernel this constant equals 0. In the binary classification problem, as consequence of this behavior of the local kernels, the models generated by a kernel machine collapse to constant models (i.e. classifiers with the same

output for all the examples) or to nearest neighbor models (i.e. the predicted class of an example is defined only by the nearest neighbor example in the high-dimensional space). In both cases, the obtained models have poor prediction (constant or highly local). Moreover, when $\mathbf{x}$ is a high-dimensional vector, the nearest neighbor example is not much closer than the other examples, due to the geometrical properties of the high-dimensional spaces highlighted by Bellman.

New solutions have been recently presented with the aim to circumvent the difficulties concerning the scaling and the computational efficiency issues. For example, the random approximations of the kernels in order to avoid computational and memory issues are presented in [6, 11, 13]. These techniques consist in the approximation of the features of the Reproducing Kernel Hilbert Space (RKHS) to generate a linear kernel that approximates the original one. This new kernel can be used as a linear kernel in the original space and can be easily scaled-up. Using a linear kernel, the computational efficiency arises in a natural way exploiting the very efficient and distributed linear kernel machines (e.g. the Pegasos algorithm [9]).

As pointed out before, the second critical issue about kernels is that they bring shallow and local information. In theory, kernels learn non-linear functions ϕ in the input space $\mathcal{X}$. However, traditionally, kernels were used to implement only a linear function in a pre-defined RKHS (the feature space). Starting from a fixed *a priori* kernel on a space $\mathcal{X}$, the entire learning phase is performed in a single step by exploiting the implicit representation $\boldsymbol{\phi}(\mathbf{x})$ of the examples $\mathbf{x} \in \mathcal{X}$, as summarized in the following scheme:

$$\mathbf{x} \in \mathcal{X} \rightarrow \mathbf{K}_{fixed} \sim \boldsymbol{\phi}(\mathbf{x}) \rightarrow \mathbf{w}_{learned} \cdot \boldsymbol{\phi}(\mathbf{x}).$$

where $\mathbf{w}_{learned}$ is the learaned model in the RKHS.

4 Next step: Learning the kernel

Learning a new implicit representation defined by a kernel is one the current challenges in the machine learning research community.

Kernel learning has the goal to learn the optimal kernel (and then the best implicit feature map $\boldsymbol{\phi}$) given a specific task or set of tasks. Multiple Kernel Learning (MKL) [7] is one of the most popular approach to kernel learning. MKL algorithms are designed to combine a set of *weak* kernels to obtain a better one. MKL has the support of a theoretical framework [5] that claims a fundamental rule to overcome the shallowness of the single kernel representation. In particular, it has been proved that combining a large number of different kernels produces just a minor penalty in the generalization bounds. This result suggest that it is possible to combine thousands or millions of kernels without falling in the overfitting problem. A kernel can been seen as a different point of view of a task and then, the combination of millions of different point of views can be a solution to create a sort of *deeper* kernel, i.e. a kernel that is not shallow.

References

[1] Fabio Aiolli, Giovanni Da San Martino, and Alessandro Sperduti, *A kernel method for the optimization of the margin distribution*. In ICANN (1) (2008), 305–314.

[2] Ethem Alpaydin, "Introduction to machine learning". MIT press, 2014.

[3] R. Bellman, "Adaptive control processes: A guided tour". Princeton University Press, New Jersey, 1961.

[4] Yoshua Bengio, Olivier Delalleau, Nicolas Le Roux,and Downtown Branch, *The Curse of Dimensionality for Local Kernel Machines*. 2(2) (2005), 1–17.

[5] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh, *Generalization bounds for learning kernels*. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel (2010), 247–254.

[6] Andrew Cotter, Joseph Keshet, and Nathan Srebro, *Explicit Approximations of the Gaussian Kernel*. Preprint arXiv:1109.4603, p. 11.

[7] Mehmet Gonen and Ethem Alpaydin, *Multiple kernel learning algorithms*. Journal of Machine Learning Research, 12 (2011), 2211–2268.

[8] Gregoire Montavon, Mikio Braun, and Klaus-Robert Muller, *Kernel Analysis of Deep Networks*. 2563–2581.

[9] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter, *Pegasos: Primal estimated sub-gradient solver for svm*. Mathematical Programming, 127(1) (2011), 3–30.

[10] John Shawe-Taylor and Nello Cristianini, "Kernel Methods for Pattern Analysis". Cambridge University Press, 2004.

[11] Si Si, Cho-Jui Hsieh, and Inderjit Dhillon, *Memory Efficient Kernel Approximation*. Proceedings of The 31st International Conference on Machine Learning, 701–709.

[12] Vladimir Naumovich Vapnik and Vlamimir Vapnik, "Statistical learning theory, volume 1". Wiley New York, 1998.

[13] Xiao-tong Yuan, ZhenzhenWang, Jiankang Deng, Qingshan Liu, and Senior Member, *Efficient χ 2 Kernel Linearization via Random Feature Maps*. (2015), 1–6.

A simple mathematical model for micro-swimmers

Marta Zoppello [(*)]

Abstract. What does it mean swimming? How can mathematics treat this problem? What is the best strategy to move in a certain direction? The study of the swimming strategies of micro-organisms is attracting increasing attention in the recent literature. One of the main difficulties is the complexity of the hydrodynamic forces exerted by the fluid on the swimmer as a reaction to its shape changes. We show that there exists an optimal swimming strategy which leads to minimize the time to reach a desired target. Numerical simulations performed are in good agreement with theoretical predictions and suggest that the optimal strategy is periodic, i.e. composed of a sequence of identical strokes.

1 Introduction

Swimming at a micro scale is a subject of growing interest, with potential applications for example in medicine or micro and nano technology. The swimming strategy of micro-organisms in low Reynolds number fluids is attracting increasing attention in the recent literature, see for instance [19] for an extensive list of references. One of the pioneering works is probably [22] by Taylor in 1951, presenting a model of swimmer as an infinite sheet shaped as a sinusoidal traveling wave, with a mathematical setting for the self-propulsion of this thin undulating filament. Later in 1977, Purcell proved in [20] that the swimming strategies must change the shape of the swimmer in a non-reciprocal way, in order to permit a displacement through the fluid, and introduced a 3-link swimmer model along with a stroke that allows it to move. More recently, several works have studied in more detail the physical characteristic of this "Purcell swimmer", see for instance [21], [6], [1], [18]. Another crucial development for our analysis is the recent emergence of the connection between swimming and Control Theory ([17], [3], [4], [10], [16], [2]). One of the difficulties is the study of the swimmer-fluid coupling which gives the dynamics of the swimmer. At a micro scale, the non local hydrodynamic forces exerted by the fluid on the swimmer can be approximated with local drag forces depending linearly on the velocity of each point (see [15], [13]). This technique called Resistive Force Theory provides a

[(*)]Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: marta.zoppello@gmail.com . Seminar held on December 2nd, 2015.

simplified dynamics that matches well those obtained by the full hydrodynamic model, see [1], [13]. We use here this technique for the N-link swimmer, as in [1].

In this work we present a controllability result for the N-link swimmer, and numerical simulations that suggest a new optimal stroke for displacement in minimum time. First, we prove by geometric control techniques that for $N \geq 3$ sticks, the N-link swimmer can reach any configuration in the plane. More precisely, we show that for almost any swimmer (i.e. for almost every set of stick lengths) and for any initial configuration, the swimmer can reach any shape and position. This result shows the existence of a suitable shape deformation which steers the swimmer to the desired final state. As a direct consequence, we show that the optimal control problem to reach a configuration in minimum time is well posed. Therefore, there exists an optimal strategy leading to the final position and configuration in minimum time. Finally, we present some numerical simulations for the Purcell swimmer ($N = 3$) with a direct method (BOCOP[(†)]). Without making any assumptions on the structure of the optimal strategy, our results suggest that the optimal swimming motion is indeed periodic, with a sequence of identical strokes. We observe that the stroke we obtain is different from the Purcell one, and gives a speed greater by about 20%. Then we address the optimal design issue, namely finding the optimal length ratio between the three links which maximizes displacement of the swimmer. A similar issue has been studied in [21] where Fourier transform expansion is used to derive an optimal design. Here, the control theory is used to approximate the leading order term of the swimmer displacement. The advantage of this approach is that this term is easy to be optimized and and fits the numerical results. As far as we know, this procedure is original in that context.

2 Setting of the problem

We recall the N-link swimmer introduced in [1], and present its motion as a system of three ODEs. The system is linear with respect to the deformation rate, and has no drift.

2.1 The N-link swimmer

The swimmer consists of $N \in \mathbf{N}$ rigid links with joints at their ends, see Fig. 1. Motion is expressed in the laboratory-frame, defined by the vectors $(\mathbf{e}_x, \mathbf{e}_y)$. We set $\mathbf{e}_z := \mathbf{e}_x \times \mathbf{e}_y$. The i-th link is the segment with end points $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$. We note $L_i > 0$ its length and θ_i its angle with the horizontal x-axis. We define (x_i, y_i) the coordinates of each point $\mathbf{x}_i$ For $i \in \{2 \cdots N\}$, the coordinates $\mathbf{x}_i$ can be expressed as:

$$\mathbf{x}_i := \mathbf{x}_1 + \sum_{k=1}^{i-1} L_k \begin{pmatrix} \cos(\theta_k) \\ \sin(\theta_k) \end{pmatrix}. \tag{1}$$

The swimmer is described by two sets of variables:

- the position and orientation of the first link, associated with the triplet $(\mathbf{x}_1 = (x_1, y_1), \theta_1)$.

[(†)] `http://bocop.org`

- the relative orientations between successive links. For $i \in [2, \cdots, N]$, we note $\alpha_i = \theta_i - \theta_{i-1}$ the angle between the $(i-1)$-th and i-th links. The vector $(\alpha_2, \ldots, \alpha_N)$ defines the shape of the swimmer.

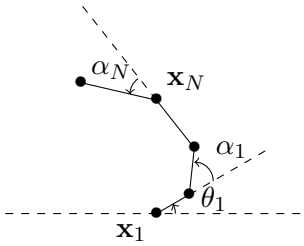


Figure 1. Coordinates for the N-link swimmer.

2.2 Dynamics

We recall in this section the main steps to obtain the equations of motion using the Resistive Force Theory, as in [1]. The dynamics of the swimmer stems from Newton laws, neglecting the inertia:

$$\left\{ \begin{array}{l} \mathbf{F} = 0\,, \\ \mathbf{e}_z \cdot \mathbf{T}_{\mathbf{x}_1} = 0\,, \end{array} \right. \tag{2}$$

where $\mathbf{F}$ is the total force exerted on the swimmer by the fluid and $\mathbf{T}_{\mathbf{x}_1}$ is the total torque about the point $\mathbf{x}_1$.

The Resistive Force Theory uses the local drag approximation for the coupling between fluid and swimmer. We denote by s the arc length coordinate on the i-th link ($0 \le s \le L_i$) and by $\mathbf{v}_i(s)$ the velocity of the corresponding point. We also introduce the local frame $(\mathbf{e}_i, \mathbf{e}_i^\perp)$ defined by

$$\mathbf{e}_i = \begin{pmatrix} \cos(\theta_i) \\ \sin(\theta_i) \end{pmatrix} \quad \mathbf{e}_i^\perp = \begin{pmatrix} -\sin(\theta_i) \\ \cos(\theta_i) \end{pmatrix}$$

and write $\mathbf{x}_i(s) = \mathbf{x}_i + s\mathbf{e}_i$. By differentiation, we obtain,

$$\mathbf{v}_i(s) = \dot{\mathbf{x}}_i + s\ \dot{\theta}_i\ \mathbf{e}_i^\perp\,. \tag{3}$$

The force $\mathbf{f}_i$ acting on the i-th segment is assumed to depend linearly on the velocity. It is defined by

$$\mathbf{f}_i(s) := -\xi\,(\mathbf{v}_i(s)\cdot\mathbf{e}_i)\,\mathbf{e}_i - \eta\left(\mathbf{v}_i(s)\cdot\mathbf{e}_i^\perp\right)\mathbf{e}_i^\perp, \tag{4}$$

where ξ and η are the drag coefficients along the directions of $\mathbf{e}_i$ and $\mathbf{e}_i^\perp$. We thus obtain

$$\left\{ \begin{array}{l} \mathbf{F} = \displaystyle\sum_{i=1}^{N} \int_0^{L_i} \mathbf{f}_i(s)\,ds\,, \\ \mathbf{e}_z \cdot \mathbf{T}_{\mathbf{x}_1} = \mathbf{e}_z \cdot \displaystyle\sum_{i=1}^{N} \int_0^{L_i} (\mathbf{x}_i(s) - \mathbf{x}_1) \times \mathbf{f}_i(s)\,ds\,. \end{array} \right. \tag{5}$$

Then the dynamics of the swimmer is finally expressed as

$$\begin{pmatrix} \dot\alpha_2 \\ \vdots \\ \dot\alpha_N \\ \dot{\mathbf{x}}_1 \\ \dot\theta_1 \end{pmatrix} = \sum_{i=1}^{N-1} \begin{pmatrix} \mathbf{b}_i \\ \tilde{\mathbf{g}}_i\,(\theta_1, \alpha_2, \cdots, \alpha_N) \end{pmatrix} \dot\alpha_{i+1}\,. \tag{6}$$

where $\mathbf{b}_i$ is the i–th vector of the canonical basis of $\mathbf{R}^{N-1}$.

3 Controllability and optimal control problems

3.1 Controllability

This Section is devoted to the controllability of the N-link swimmer. We prove that there exist control functions which allow the swimmer to move everywhere in the plane.

Theorem 3.1 *Consider the N-link swimmer described in Section 2 evolving in the space $\mathbf{R}^2$. Then for almost every lengths of the sticks $(L_i)_{i=1,\cdots,N}$ and for any initial configuration $(\mathbf{x}_1^i, \theta_1^i, \alpha_2^i, \cdots, \alpha_N^i) \in \mathbf{R}^2 \times [0, 2\pi]^N$, any final configuration $(\mathbf{x}_1^f, \theta_1^f, \alpha_2^f, \cdots, \alpha_N^f)$ and any final time $T > 0$, there exists a shape function $(\alpha_2, \cdots, \alpha_N) \in \mathcal{W}^{1,\infty}([0,T])$, satisfying $(\alpha_2, \cdots, \alpha_N)(0) = (\alpha_2^i, \cdots, \alpha_N^i)$ and $(\alpha_2, \cdots, \alpha_N)(T) = (\alpha_2^f, \cdots, \alpha_N^f)$ and such that if the self-propelled swimmer starts in position $(\mathbf{x}_1^i, \theta_1^i)$ with the shape $(\alpha_2^i, \cdots, \alpha_N^i)$ at time $t = 0$, it ends at position $(\mathbf{x}_1^f, \theta_1^f)$ and shape $(\alpha_2^f, \cdots, \alpha_N^f)$ at time $t = T$ by changing its shape along $(\alpha_2, \cdots, \alpha_N)(t)$.*

3.2 Minimum time Optimal Control Problem

Here we present the minimum time optimal control problem for the N-link swimmer, which is well defined from the controllability result proven in Section 3. Then we present the numerical method used to solve this problem.

For any time $t > 0$, we denote the state of the swimmer by $\mathbf{z}(t) := (\alpha_2, \cdots, \alpha_N, \mathbf{x}_1, \theta_1)(t)$, the control function by $\mathbf{u}(t) := (\dot\alpha_2, \cdots, \dot\alpha_N)(t)$ and the dynamics by $\mathbf{f}(\mathbf{z}(t), \mathbf{u}(t)) = \sum_{i=1}^{N-1} \mathbf{g}_i(\mathbf{z}(t))\,\dot\alpha_{i+1}(t)$.

We now assume that the swimmer starts at the initial configuration $\mathbf{z}^i$, and we set a final state $\mathbf{z}^f$. We want to find a swimming strategy that minimizes the time to reach the final configuration, i.e.,

$$(OCP) \begin{cases} \inf\ t_f\,, \\ \dot{\mathbf{z}}(t) = f(\mathbf{z}(t), \mathbf{u}(t))\ ,\ \ \forall t \in [0, t_f]\,, \\ \mathbf{u}(t) \in \mathbf{U} := [-1, 1]^N\ ,\ \ \forall t \in [0, t_f]\,, \\ \mathbf{z}(0) = \mathbf{z}^i \quad , \quad \mathbf{z}(t_f) = \mathbf{z}^f\,. \end{cases}$$

By applying Filippov-Cesary Theorem ([23]), there exists a minimal time such that the constraints are satisfied i.e., the infimum can be written as a minimum.

3.2.1 Numerical simulations for the Purcell swimmer (N= 3)

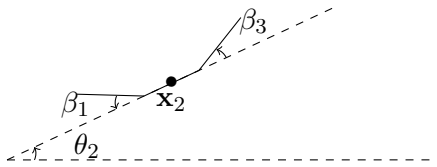


Figure 2. Matching the notations used in [6], we use the variables above to represent the Purcell swimmer.

We present here the numerical simulations for the Purcell swimmer (3 sticks). Without making any assumptions on the structure of the optimal trajectory, we obtain a solution with periodic strokes. We compare this stroke to the one of Purcell ([20], [6]), and observe that it gives a better displacement speed.
In order to solve this optimal control problem, we use a so-called direct approach. The direct approach transforms the infinite dimensional optimal control problem (OCP) into a finite dimensional optimization problem (NLP). This is done by a discretization in time applied to the state and control variables, as well as the dynamics equation. All tests were run using the software Bocop ([9]). The discretized nonlinear optimization problem is solved by the well-known solver Ipopt [24] with Mumps [5], while the derivatives are computed by sparse automatic differentiation with Adol-C [25] and ColPack [14].

3.2.2 The classical Purcell stroke

We recall the stroke presented by Purcell in [20] in order to compare it to the optimal strategy given by our numerical results. Let us denote by $\Delta\theta$ the angular excursion, meaning that β_1 and β_3 belong to $[-\frac{\Delta\theta}{2}, \frac{\Delta\theta}{2}]$. The Purcell stroke is defined by the periodic cycle of deformation over $[0, T]$:

$$\big(\,\beta_1(t), \beta_3(t)\,\big) = \begin{cases} \big(\,\frac{4\Delta\theta}{T}t - \frac{\Delta\theta}{2}, \frac{\Delta\theta}{2}\,\big) & \text{if } 0 \le t \le \frac{T}{4} \\ \big(\,\frac{\Delta\theta}{2}, -\frac{4\Delta\theta}{T}t + \frac{3\Delta\theta}{2}\,\big) & \text{if } \frac{T}{4} \le t \le \frac{T}{2} \\ \big(\,-\frac{4\Delta\theta}{T}t + \frac{5\Delta\theta}{2}, -\frac{\Delta\theta}{2}\,\big) & \text{if } \frac{T}{2} \le t \le \frac{3T}{4} \\ \big(\,-\frac{\Delta\theta}{2}, \frac{4\Delta\theta}{T}t - \frac{7\Delta\theta}{2}\,\big) & \text{if } \frac{3T}{4} \le t \le T \end{cases} .$$

In the following, we call the "classical" Purcell stroke the one corresponding to $\Delta\theta = \frac{\pi}{3}$, with $T = 4\Delta\theta$ chosen to satisfy the constraints on the speed of deformation stated in (OCP), i.e., $u_i(t) := \dot{\beta}_i(t) \in [-1, 1]$.

3.2.3 Comparison of the optimal stroke and Purcell stroke

We set the initial position $\mathbf{x}_2, \theta_2 = (0, 0, 0)$ and the final position $\mathbf{x}_2, \theta_2 = (-0.25, 0, 0)$. We also constrain the angles β_1 and β_3 in $[-\frac{\pi}{6}, \frac{\pi}{6}]$ for all time. Solving the minimum time problem with the direct method gives us a solution that is actually periodic.

Extracting one stroke from this solution, and comparing it with the Purcell stroke we have

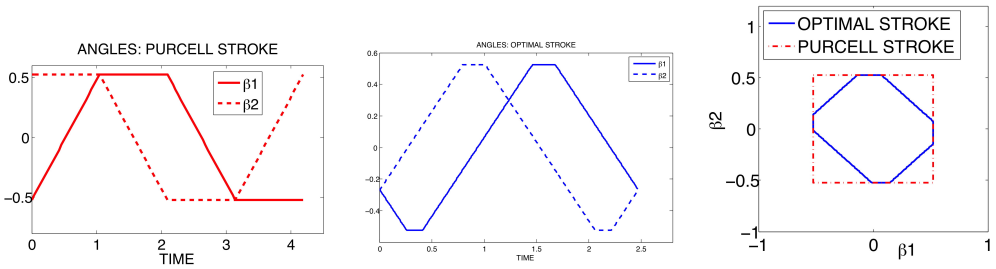


Figure 3. Angles and phase portrait - Purcell stroke and optimal stroke.

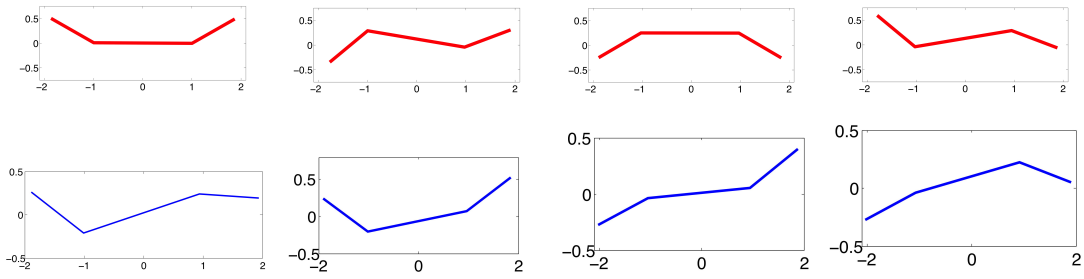


Figure 4. Purcell stroke (above) and optimal stroke (below).

We observe that using Purcell strokes, the swimmer only reaches $(\approx -0.18, 0)$, which confirms that our optimal stroke allows a greater x-displacement.

3.3 Optimal design for the three link Purcell swimmer

Optimal control problem. We are interested in finding a periodic sequence of deformations which maximizes the displacement of the swimmer along the x-axis. More precisely, we optimize both the link length ratio L_2/L and the deformation of the swimmer over time. Taking the deformation speed $\dot{\beta}_{1|3}$ as control functions, we obtain the optimal control problem

$$(OCP2)\begin{cases} \max\ x_2(T)\ \text{ s.t.} \\ \dot{\mathbf{z}}(t) = f(\mathbf{z}(t), \dot{\beta}_1, \dot{\beta}_3) \quad \forall t \in [0,T]\,, \\ \dot{\beta}_{1|3} \in \mathbf{U} = [-b, b] \quad \forall t \in [0,T]\,, \\ \beta_{1|3}(t) \in [-a, a] \quad \forall t \in [0,T]\,, \\ x_2(0) = y_2(0) = \theta_2(0) = 0, y_2(T) = \theta_2(T) = 0\,, \\ \beta_{1|3}(0) = \beta_{1|3}(T), \\ 2L + L_2 = c. \end{cases}$$

We set the constraints a and b over the amplitude and deformation speed, as well as the total length c of the swimmer. The final time T is fixed, and the constraint $\beta_{1|3}(0) = \beta_{1|3}(T)$ ensures that the swimmer is in the same configuration at the initial and final time.

Pontryagin's Maximum Principle (PMP). We recall here the PMP as it gives some insight on the shape of optimal strokes. This theorem in optimal control introduced by Pontryagin et al. in [7] gives necessary conditions for local optimality. The PMP is characterized by an Hamiltonian function H that formally depends on the state variables $\mathbf{z}$, the control functions $\dot{\beta}_{1|3}$, and so-called *co-state* variables noted $\mathbf{p}$. Let the Hamiltonian be

$$H(\mathbf{z}, \mathbf{p}, \dot{\beta}_1, \dot{\beta}_3) = \langle \mathbf{p}, \mathbf{g}_1(\mathbf{z}) \rangle \dot{\beta}_1 + \langle \mathbf{p}, \mathbf{g}_2(\mathbf{z}) \rangle \dot{\beta}_3. \tag{7}$$

Under the assumption that $\mathbf{g}_{1|2}$ are continuous and C^1 with respect to $\mathbf{z}$, the PMP states that:
if $(\mathbf{z}^*, \dot{\beta}_1^*, \dot{\beta}_3^*)$ is a solution of (OCP) then there exists $\mathbf{p}^* \neq 0$ absolutely continuous such that $\dot{\mathbf{z}}^* = H_p(\mathbf{z}^*, \mathbf{p}^*, \dot{\beta}_1^*, \dot{\beta}_3^*)$, $\dot{\mathbf{p}}^* = -H_z(\mathbf{z}^*, \mathbf{p}^*, \dot{\beta}_1^*, \dot{\beta}_3^*)$, $\mathbf{p}^*(T)$ is orthogonal to the cotangent cone of the final conditions at $\mathbf{z}^*(T)$ and $(\dot{\beta}_1^*, \dot{\beta}_3^*)$ maximizes the Hamiltonian for almost every time $t \in [0, T]$.

- **Bang arcs:** If $\langle \mathbf{p}, \mathbf{g}_i(\mathbf{z}) \rangle \neq 0$ for $i = 1, 2$ over a time interval, then the optimal control $\dot{\beta}_{1|3^*}$ must be on the boundary of $U = \{(-b, -b), (-b, b), (b, -b), (b, b)\}$.
- **Constrained arcs:** When $|\beta_i| = a$, the corresponding control $\dot{\beta}_i = 0$.
- **Symmetries:** optimal strokes should be symmetric with respect to the diagonal axes $\beta_1 = \beta_3$ and $\beta_1 = -\beta_3$ (linearity and time independence).

3.3.1 Optimal swimmer design

In this section, we express the leader term of the swimmer's displacement for a stroke of small perimeter which satisfies all properties stated in the previous section. We represent the stroke by a closed octagonal curve γ in the phase portrait (β_1, β_3), see Fig. 5.

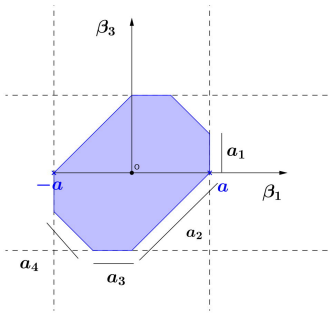


Figure 5. (Color online) Phase portrait (β_1, β_3) of the octagonal stroke considered for the expansion of the displacement.

The expansion of the displacement over the complete stroke is

$$\mathbf{z}(T) - \mathbf{z}(0) = C\ [\mathbf{g}_1, \mathbf{g}_2](\mathbf{z}(0)) + o(a_i^3)_{i=1-4}\,, \tag{8}$$

where

$$C = \frac{a_1 a_2\sqrt{2}}{2} + a_1 a_3 + \frac{a_2 a_3\sqrt{2}}{2} + \frac{a_1 a_4\sqrt{2}}{2} + a_2 a_4 + \frac{a_3 a_4\sqrt{2}}{2}$$

Using the explicit expression of the Lie brackets around $\mathbf{z}(0) = (0, 0, x, y, 0)$

$$x(T) - x(0) = C\left(\frac{\eta - \xi}{\xi}\right)\left(\frac{L^3 L_2(3L + 2L_2)}{(2L + L_2)^4}\right) + o(a_i^3)_{i=1-4} \tag{9}$$

Setting the total length of the swimmer by a constant equal to c, i.e., $2L + L_2 = c$, we find that (9) has a unique maximum at

$$L^* = c\Big(1 - \sqrt{\frac{2}{5}}\Big), \quad L_2^* = c\Big(2\sqrt{\frac{2}{5}} - 1\Big), \tag{10}$$

which gives an optimal ratio of

$$\left(\frac{L_2}{L}\right)^* = \frac{\sqrt{10} - 1}{3} \sim 0.721\,. \tag{11}$$

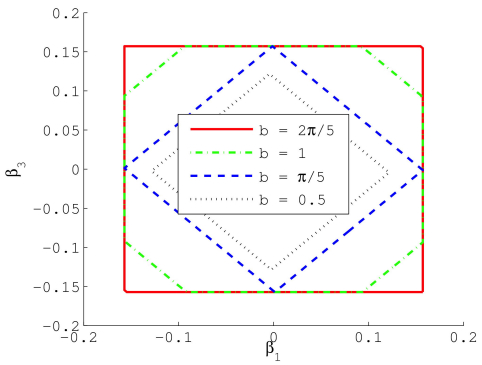
3.4 Numerical simulations

We solve now the optimal control problem $(OCP2)$ numerically, in order to determine the optimal swimming strategy and link ratio. Simulations are performed with the toolbox Bocop ([8]) that implements a direct transcription method.

We explore different values for the bounds a, b on the shape angles and deformation speed and see their influence on the optimal stroke and link ratio.

3.4.1 Small amplitudes, influence of speed limits

We start with small amplitudes by setting $a = \pi/20$ and solve $(OCP2)$ for different values of the speed limit b. Here we set $T = 1$ and use 250 time steps for the discretization. Results are given in Table below, with the phase portraits for the shape angles β_1, β_3.



b	$x(T)$	L_2/L	stroke
0.5	2.68E-3	0.719	diamond
$\pi/5$	4.23E-3	0.719	diamond
0.75	5.70E-3	0.719	octagon
1	7.73E-3	0.719	octagon
$2\pi/5$	8.42E-3	0.717	square
1.5	1.14E-2	0.719	octagon (x2)
2	1.55E-2	0.719	octagon (x2)

First, we observe that the optimal ratio L_2/L is very close to its theoretical value of 0.721 from (11), regardless of b. The three strokes observed (diamond, octagon, square) match the discussion from previous Section. They include only diagonal lines (bang arcs saturating the speed limit b) and horizontal/vertical lines (constrained arcs for the amplitude limit a).

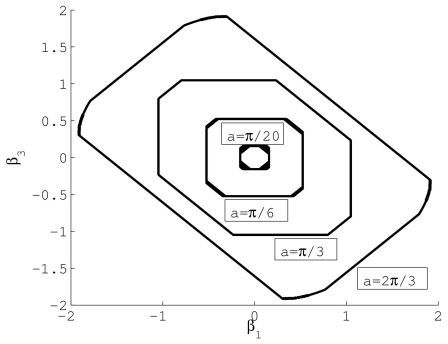
3.5 Large amplitudes, influence of angle limits

Now we study the influence of the maximal amplitude of the stroke, set by the bound a. In this last part we set the deformation speed limit $b = 1$ to focus on the amplitude.

The results are illustrated in the table and figure below. The shape of the optimal stroke is always octagonal until it becomes unconstrained for very large values of a. We observe that the central symmetry observed for small amplitudes is lost for larger a, however symmetry w.r.t both diagonal axes still holds as expected.

In the unconstrained case, we see arcs that are neither bang arcs (diagonal) or constrained arcs (horizontal/vertical), but rather appear as smooth curves (see Fig.below). These are characteristic of so-called *singular arcs*, namely the case where $\langle p, g_i(z)\rangle = 0$ in the PMP.

The optimal ratio L_2/L shows a steady decrease with a, starting quite close to the value 0.721 computed for small amplitudes, the seemingly reaching a limit value of 2/3 in the unconstrained case (i.e. $L = 1.5, L_2 = 1$). We recall that the classical Purcell swimmer has a link ratio of 2 ($L = 1, L_2 = 2$).



a	x(T)	L_2/L	stroke
$\pi/20$	0.192	0.719	octagon x26
$\pi/10$	0.384	0.712	octagon x13
$\pi/6$	0.593	0.697	octagon x7
0.75	0.811	0.676	octagon x5
$\pi/3$	1.088	0.660	octagon x4
1.25	1.266	0.660	octagon x4
1.5	1.263	0.660	octagon x3
1.75	1.329	0.667	octagon x3
$2\pi/3$	1.335	0.667	unconstrained x3
2.5	1.335	0.667	unconstrained x3

4 Conclusions

In this work we have presented a discrete model of a slender swimmer inspired by the one of Purcell [20] which swims changing its shape, discretizing its body with a chain of N links . We prove that for N greater than 3 and for almost any N-uplet of sticks lengths, the swimmer is globally controllable in the whole plane. Then, we focus on finding a swimming strategy that leads the N-link swimmer from an fixed initial position to a given final position, in minimum time. As a consequence of the controllability result, we show

that there exists a shape change function which allows to reach the final state in a minimal time. Instead of using the approach of the minimum time function [11, 12], we formulate this optimal control problem and solve it with a direct approach (BOCOP) for the case $N = 3$ (Purcell swimmer). Without any assumption on the structure of the trajectory, we obtain a periodic solution, from which we identify an optimal stroke. Comparing this optimal stroke with the Purcell one confirms that it is better and gives a speed greater by about 20%.
Current work includes solving the optimal control problem for more complex displacements (along the y axis, rotations) and for the optimal design, i.e. the optimization of the link ratio of the three-link swimmer for maximal displacement. We provide an estimate of the displacement based on an expansion for small deformations, which gives a theoretical optimal link ratio. Numerical simulations are consistent with this theoretical ratio for small amplitudes of deformation. We also observe that the optimal ratio changes for large amplitudes of deformation, with a limit value of 0.667 in the unconstrained case versus a theoretical ratio of 0.721 obtained for small amplitudes of deformation. For an amplitude of $\pi/3$, the displacement gain is about 60% compared with the classical Purcell swimmer design. A possible continuation of this work is the comparison of different objective functions, such as speed or efficiency.

Also, noticing that the N-link swimmer was introduced in [1] in the perspective of approximating the motion of several living micro organisms, an interesting extension of this model is to generalize the simulations to greater values of N. Of course, comparing the candidate for the optimal motion strategy with the one used by real micro organism could be a more tricky issue. On the other hand, another interesting direction is to study formally the existence of the periodic solution for the optimal problem.

References

[1] F. Alouges, A. De Simone, L. Giraldi, and M. Zoppello, *Self-propulsion of slender micro-swimmers by curvature control: N-link swimmers.* J. Nonlinear Science, 56 (2013), 132–141.

[2] F. Alouges, A. De Simone, L. Heltai, A. Lefebvre, and B. Merlet, *Optimally swimming stokesian robots.* Discrete and Continuous Dynamical Systems - Series B, 18(5) (2010),1189–1215.

[3] F . Alouges, A. De Simone, and A. Lefebvre, *Optimal strokes for low Reynolds number swimmers: an example.* Journal of Nonlinear Science, 18(3) (2008), 277–302.

[4] F. Alouges and L. Giraldi, *Enhanced controllability of low reynolds number swimmers in the presence of a wall.* Acta Applicandae Mathematicae, 128(1) (2012), 153–179.

[5] P. R. Amestoy, I. S. Duff, J. Koster, and J.-Y. L. Excellent, *A fully asynchronous multifrontal solver using distributed dynamic scheduling.* SIAM Journal of Matrix Analysis and Applications, 23(1) (2001), 15–41.

[6] L. E. Becker, S. A. Koehler, and H. A. Stone, *On self-propulsion of micro-machines at low reynolds number: Purcell?s three-link swimmer.* J. Fluid Mech., 490 (2003), 15–35.

[7] V. G. Boltyanskii, R. V. Gamkrelidze, and L. S. Pontryagin, *Towards a theory of optimal processes.* Dokl. Akad. Nauk SSSR, 110(1) (1956), 7–10.

[8] F. Bonnans, D. Giorgi, S. Maindrault, P. Martinon, and V. Grélard, *Bocop - a collection of examples.* Technical report, INRIA, `http://www.bocop.org`, 2014.

[9] J. Bonnans, Frédéric, Pierre Martinon, and Vincent Grélard, *Bocop - A collection of examples.* Technical report, INRIA, 2012. RR-8053.

[10] T. Chambrion and A. Munnier, *Generic controllability of 3d swimmers in a perfect fluid.* SIAM J. Control Optim., 50(5) (2011), 2814–2835.

[11] G. Colombo and Khai T. Nguyen, *On the structure of the minimum time function.* SIAM J. Contr. Optim., 48(7) (2010), 4776–4814.

[12] G. Colombo and Khai T. Nguyen, *On the minimum time function around the origin.* Mathermatics od Control and Related Fields, 3(1) (2013), 51–82.

[13] B. M. Friedrich, I. H. Riedel-Kruse, J. Howard, and F. Jülicher, *High-precision tracking of sperm swimming fine structure provides strong test of resistive force theory.* The Journal of Experiment Biology, 213 (2010), 1226–1234.

[14] A. Gebremedhin, A. Pothen, and A. Walther, *Exploiting sparsity in jacobian computation via coloring and automatic differentiation: a case study in a simulated moving bed process.* In C. Bischof et al, editor, Lecture Notes in Computational Science and Engineering 64, pp. 339–349. Springer, 2008. Proceedings of the Fifth International Conference on Automatic Differentiation (AD2008).

[15] J. Gray and J. Hancock, *The propulsion of sea-urchin spermatozoa.* Journal of Experimental Biology (1955).

[16] J. Lohéac, J. F. Scheid, and M. Tucsnak, *Controllability and time optimal control for low Reynolds numbers swimmers.* hal-00635981 (2011).

[17] R. Montgomery, "A tour of subriemannian geometries, theirs geodesics and applications". American Mathematical Society, 2002.

[18] E. Passov and Y. Or, *Dynamics of Purcell?s three-link microswimmer with a passive elastic tail.* Eur Phys J E, 78(35) (2012), 1–9.

[19] T. Powers and E. Lauga, *The hydrodynamics of swimming microorganisms.* Rep. Prog. Phys., 72(096601), 2009.

[20] E. M. Purcell, *Life at low Reynolds number.* American Journel of Physics, 45 (1977), 3–11.

[21] D. Tam and A. E. Hosoi, *Optimal strokes patterns for Purcell?s three link swimmer.* Physical Review Letters, 2007.

[22] G. Taylor, *Analysis of the swimming of microscopic organisms.* Proc. R. Soc. Lond. A, 209 (1951), 447–461.

[23] E. Trelat, "Contrôle optimal : théorie and applications". Vuibert, Collection Mathématiques Concrtes, 2005.

[24] A. Wächter and L. T. Biegler, *On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming.* Mathematical Programming, 106(1) (2006), 25–57.

[25] A. Walther and A. Griewank, "Getting started with adol-c". Naumann and O. Schenk, editors, Combinatorial Scientific Computing. Chapman-Hall CRC Computational Science, 2012.

Computational social choice: between AI and Economics

Andrea Loreggia (*)

Abstract. During the last decades, the trend has been for disciplines to converge on common techniques to be used in similar problems, besides focusing on specific techniques to be used in narrow domains. AI is one of the best examples: the cross-fertilisation process leads to a very fascinating solutions. Consider for example genetic algorithms, which mimic evolutionary mechanisms to solve search and optimization problems. The individualistic approach of problem solving becomes insufficient: concepts, techniques and experts need to collaborate to get a better understanding of the problems they would like to solve. The techniques that AI makes available are being used by many other disciplines. AI nowadays inundates our everyday life with tools and methods that are hidden in our household electrical devices, smartphones and much more. Starting from the field of multi-agent systems, researchers in AI recently considered the use of models and problems from economics. Notable examples are voting systems used to aggregate the results of several search engines, game theoretic methods that analyse the complex interaction of autonomous agents, and matching procedures implemented on large-scale problems such as the coordination of kidneys transplants and the assignment of students to schools. In this scenario, a number of research lines federated under the name of computational social choice. The need for a computational study of collective decision procedures is clear. On the one hand, from crowdsourcing to university admission ranking, many real-life applications apply existing social choice methods to large scale problems. On the other hand, collective decision-making is not a prerogative of human societies, and multi-agent systems can use these methods to coordinate their actions when facing complex situations. In this artcile, we would like to focus on two examples that highlight the impact of a computational approach to classical problems of collective choice. First, by studying repeated decisions (think of opinion polls that precede an election) to evaluate the quality of the result, and, second, by devising innovative procedures to predict the preferences of a collection of individuals.

1 Introduction

During the last few decades, the trend has been for disciplines to converge on common techniques to be used in similar problems, besides focusing on specific techniques to be used in narrow domains. AI is one of the best examples: the cross-fertilisation process has

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: andrea.loreggia@gmail.com . Seminar held on December 17th, 2015.

led to very fascinating solutions. Consider for example genetic algorithms, which mimic evolutionary mechanisms to solve search and optimization problems [Gol89]. Or think of bird flocking or fish schooling, which are reproduced in particle swarm optimization [EK95] and used in coordinating autonomous driverless cars [GGLV12].

The individualistic approach of problem solving becomes insufficient: concepts, techniques and experts need to collaborate to get a better understanding of the problems they would like to solve. The techniques that AI makes available are being used by many other disciplines. Just think of the variety of machine learning techniques used in medicine, physics or astronomy, or the constraint programing algorithms that AI researchers use to solve planning problems. AI nowadays inundates our everyday life with tools and methods that are hidden in our household electrical devices, smartphones and much more.

Starting from the field of multi-agent systems, researchers in AI recently considered the use of models and problems from economics. Notable examples are voting systems used to aggregate the results of several search engines [DKNS01], game theoretic methods that analyse the complex interaction of autonomous agents [SLB09], and matching procedures implemented on large-scale problems such as the coordination of kidneys transplants [ABS07] and the assignment of students to schools [GC10].

In this scenario, a number of research lines federated under the name of computational social choice [RVW11]. The need for a computational study of collective decision procedures is clear. On the one hand, from crowdsourcing to university admission ranking, many real-life applications apply existing social choice methods to large scale problems. On the other hand, collective decision-making is not a prerogative of human societies, and multi-agent systems can use these methods to coordinate their actions when facing complex situations.

A prime example is the Sydney Coordinated Adaptive Traffic System (SCATS) is a real-life multi-agent system implementation used in different cities of 27 countries around the world to manage city traffic. The system uses an adaptive approach [Sys14] which permits to adjust the management plan to the different daily traffic situations. Each intersection has a computer that manages the traffic based on an assigned plan. There are also sensors to analyse the traffic flow, this analysis allows to adjust the management of the traffic by extending or reducing the green phase. But the adjustment cannot be computed using only what a single traffic light can capture. Data from the different traffic lights of the city is sent to a central computer which produces different plausible plans. The plan is then chosen by the intersections using a voting system: each intersection votes for its preferred plan basing its preferences on what have been captured by the sensors. The plan with more preferences is chosen to manage the traffic for a specified period of time.

In the paper, we focus on two additional examples that highlight the impact of a computational approach to classical problems of collective choice. First, by studying repeated collective decisions (that models opinion polls that precede an election) to evaluate the quality of the result, and, second, by devising innovative procedures to predict the preferences of a collection of individuals.

1.1 Voting Systems

Let C be a finite set of m candidates and V be a finite set of n individuals. We assume individuals have preferences p_i over candidates in C in the form of *strict linear orders*, i.e., transitive, anti-symmetric and complete binary relations. Individuals express their preferences in form of a ballot P_i (e.g., the top candidate, a set of approved candidates, or the full linear order) and we call the choice of a ballot for each individual a profile $P = (P_1, \dots, P_n)$. Observe that we do not allow agents to express ties among candidates, i.e., it is not possible for an agent to state that two candidates in C are equally preferred. We write $a \; P_i \; b$ to denote that agent i prefers candidate a to candidate b in profile P, on the same way we write $a \; P_i \; C \setminus \{a\}$ to denote that agent i prefers candidate a to all the candidate in the set C. In this work, we assume that individuals submit as a ballot for the election their full linear order, and we thus use the two notions of ballot and preference interchangeably. An election E is then a pair (C, V) where C is a set of m candidates and V is a collection of n votes (linear orders over C), as already said here we assume that each voter gives a complete preference order over the set of candidates. For example given $C = \{a, b, c\}$, suppose voter v_1 prefers candidate b to a and c is her less preferred candidate, then her ballot can be represented as $v_1 : b \succ a \succ c$. As usual in the literature given an arbitrary order over a set of candidates $C = \{c_1, \dots, c_m\}$ a preference like $v_i : C$ means that the voter i preferences respects that arbitrary order and so the preference corresponds to $v_i : a \succ b \succ c$. On the same way a preference likes $v_i : \overleftarrow{C}$ means that in the voter i preferences that arbitrary order is inverted and so we could rewrite it in the following way $v_i : c \succ b \succ a$.

1.2 Voting Rules

A (non-resolute) *voting rule* F associates with every profile $P = (P_1, \dots, P_n)$ a non-empty subset of winning candidates $F(P) \in 2^C \setminus \{\varnothing\}$. Let us borrow from the literature some notations useful to define voting rules and later some properties [EFS11]. In particular given two candidates c, a we set $W(c, a) = |\{i : c \; P_i \; a\}|$. There is a wide collection of voting rules that have been defined in the literature, and here we focus on the following ones:

Positional scoring rules (PSR): Let $(s_1, \dots, s_m)$ be a scoring vector such that $s_1 \geq \cdots \geq s_m$ and $s_1 > s_m$. If a voter ranks candidate c at j-th position in her ballot, this gives s_j points to the candidate. The candidates with the highest score win. We focus on four particular PSR: *Plurality* with scoring vector $(1, 0, \dots, 0)$, *veto* with vector $(1, \dots, 1, 0)$, *k-approval* with vector $(1, 1, \dots, 1, 0, \dots, 0)$, where the scoring rule rewards with 1 point k candidates, and *Borda* with vector $(m-1, m-2, \dots, 0)$.

Approval: Given a subset of approved alternatives $c_i \subseteq C$ for each $i \in V$, the winners of *approval voting* are the candidates that receive the highest number of approvals.

Copeland: Any candidate c gets 1 point for each won pairwise comparison, she gets 0 point for each tie and she gets -1 point for each lost pairwise comparison. The score of c is $score(c)$=$|\{a : W(c,a) > W(a,c)\}|$-$|\{a : W(a,c) > W(c,a)\}|$.

Maximin: The score of a candidate c is the smallest number of voters preferring it in any pairwise comparison, i.e. $score(c)=\min_{a\in C} W(c,a)$.

Single Transferable Vote (STV): If there exists a candidate that is ranked first by the majority of the voters than this is the winner, otherwise the candidate that is ranked first by the fewest number of voters gets eliminated (ties are broken following a predetermined order of candidates). Votes initially given to the eliminated candidate are then transferred to the candidate that comes immediately after in the individual preferences. This process is iterated until one alternative is ranked first by a majority of voters.

Despite its simple definition, approval voting has been the subject of an extensive literature since its first appearance (see, e.g., [LS10]).

All the previous voting rules can be adapted to output a ranking of the candidates (from higher to lower score) transforming the voting rules into *social welfare functions* [RVW11], i.e., functions which associate with every profile of preferences a ranking of the alternatives.

1.3 Iterative Voting

In a voting system, a voting rule is used to decide which decision to take, mapping the agents' preferences over the possible candidate decisions into a winning decision for the collection of agents. In these kind of scenarios, it may be desirable that agents do not have any incentive to act strategically, that is, to misreport their preferences in order to influence the result of the voting rule in their favor. Indeed, manipulation and control are usually seen as a bad behavior from an agent, to be avoided or at least to be made computationally difficult to accomplish. We know that every reasonable voting rule is manipulable when no domain restriction is imposed on the agents' preferences [Gib73, Sat75]. Following this finding, a considerable amount of work has been spent on devising conditions to avoid manipulation from the perspective of the designer of an election.

For instance, one can devise restrictive conditions on the preference profiles that can be expressed, or study computational barriers that make the calculation of manipulation strategies too hard for the agents to be performed Iterative voting models an electoral process during which voters are allowed to change their mind when the outcome of the election does not satisfy them. Voters can change their preferences in order to make another more preferred candidate win the election. The process can reproduce a multi-agent system where agents cannot share their complete knowledge (in this case their preferences), either because of media limitations which do not allow to send enough information or simply because they do not trust one another.

In this scenario the iterative process helps the system to reach an equilibrium where all the agents are satisfied. During the talk we show some theoretical results describing under which assumptions this systems converges to a stable state where no voter has incentive to cheat, either because she is satisfied, or because she cannot affect the outcome. We will also show the results of our simulations, showing that the quality of the winner after iteration is often higher than that of the winner of the initial state [GLR+13].

1.4 Sentiment Analysis

We live in a world where we communicate more and more on social media, writing text that reflects our opinions and feelings. Being able to formalize such opinions and reason with them can be very useful for a number of practical applications. First, service providers may personalize their offer based on customers opinions. Second, companies may test what products would be better received by potential consumers, and adjust their strategy accordingly. Third, community councils and candidates in political elections may evaluate the reception of their proposals, and focus their attention on the most preferred ones. It comes therefore as no surprise that the extraction of individual opinions from textual expressions, such as tweets, blog posts, or product reviews, has been the subject of a very active area of research in recent years.

Researchers in sentiment analysis and opinion mining [Liu12, PL08] developed a collection of tools in natural language processing (NLP) for the extraction of opinions, sentiments, or attitudes of individuals from their textual expressions. In order to summarize the opinion of all the individuals in a unique indicator, the opinions extracted are then used to define a notion of collective sentiment about the entities under consideration, be they commercial products, policies or candidates. Sentiment analysis is used to classify the collective opinion about a given item [Liu12]. This is done by extracting the individual opinions from text that individuals write, such as Twitter or blog posts, via natural language processing techniques. Sentiment analysis is then used to predict the opinion of the collectivity. More often it is used to predict the outcome of political elections or guessing the trend of the stock market. While sentiment analysis works quite well when we have just one item for which we would like to know what the community thinks, things change when we would like to compare multiple entities. We present our proposal to cope with the challenges of sentiment analysis over multiple items [GLRS14]. Nevertheless, the problem of generalising existing sentiment analysis techniques to account for more complex individual expressions remains mostly an open and interesting area of research.

References

[ABS07] David J. Abraham, Avrim Blum, and Tuomas Sandholm, *Clearing algorithms for barter exchange markets: enabling nationwide kidney exchanges.* In Jeffrey K. MacKie-Mason, David C. Parkes, and Paul Resnick, editors, ACM Conference on Electronic Commerce, pp. 295–304. ACM, 2007.

[DKNS01] Cynthia Dwork, Ravi Kumar, Moni Naor, and D Sivakumar, *Rank aggregation methods for the Web.* In Proceedings of the 10th international conference on World Wide Web, pp. 613–622, New York, NY, USA, 2001.

[EFS11] Edith Elkind, Piotr Faliszewski, and Arkadii M. Slinko, *Cloning in elections: Finding the possible winners.* J. Artif. Intell. Res. (JAIR), 42, pp. 529–573, 2011.

[EK95] R C Eberhart and J Kennedy, *Particle swarm optimization.* IEEE International Conference on Neural Networks, 4, pp. 1942–1948, 1995.

[GC10] Mingyu Guo and Vincent Conitzer, "Computationally feasible automated mechanism design: General approach and case studies". In Maria Fox and David Poole, editors, AAAI. AAAI Press, 2010.

[GGLV12] Jorge Godoy, Dominique Gruyer, Alain Lambert, and Jorge Villagra, *Development of an particle swarm algorithm for vehicle localization.* In Intelligent Vehicles Symposium, pp. 1114–1119. IEEE, 2012.

[Gib73] A. Gibbard, *Manipulation of voting schemes: A general result.* Econometrica, 41, pp. 587?601, 1973.

[GLR+13] Umberto Grandi, Andrea Loreggia, Francesca Rossi, Kristen Brent Venable, and Toby Walsh, *Restricted manipulation in iterative voting: Condorcet efficiency and borda score.* In Patrice Perny, Marc Pirlot, and Alexis Tsoukis, editors, ADT, volume 8176 of Lecture Notes in Computer Science, pp. 181–192. Springer, 2013.

[GLRS14] Umberto Grandi, Andrea Loreggia, Francesca Rossi, and Vijay Saraswat, *From sentiment analysis to preference aggregation.* In Proceedings of the International Symposium on Artificial Intelligence and Mathematics (ISAIM-2014), 2014.

[Gol89] David E. Goldberg, "Genetic Algorithms in Search, Optimization and Machine Learning". Addison-Wesley, Reading, MA, 1989.

[Liu12] Bing Liu, "Sentiment Analysis and Opinion Mining". Synthesis Lectures on Human Language Technologies. Morgan and Claypool Publishers, 2012.

[LS10] Jean-François Laslier and M. Remzi Sanver, editors, "Handbook of Approval Voting". Springer, 2010.

[PL08] Bo Pang and Lillian Lee, *Opinion mining and sentiment analysis.* Foundations and Trends in Information Retrieval, 2(1-2), pp. 1?135, 2008.

[RVW11] Francesca Rossi, Kristen Brent Venable, and Toby Walsh, "A Short Introduction to Preferences: Between Artificial Intelligence and Social Choice". Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan and Claypool Publishers, 2011.

[Sat75] Mark. A. Satterthwaite, *Strategy-proofness and arrow?s conditions: Existence and correspondence theorems for voting procedures and social welfare functions.* Journal of Economic Theory, 10, pp. 187?217, 1975.

[SLB09] Yoav Shoham and Kevin Leyton-Brown, "Multiagent Systems - Algorithmic, Game-Theoretic, and Logical Foundations". Cambridge University Press, 2009.

[Sys14] Sydney Coordinated Adaptive Traffic System, *SCATS: How it works — Adaptive control.* `http://www.scats.com.au/how-scats-worksadaptive.html`, 2014.

Quivers, representations of algebras and beyond

Gabriella D'Este (*)

Abstract. I will illustrate some results obtained by using techniques and general ideas coming from representation theory of finite dimensional algebras. These algebras will almost always be "path algebras" given by quivers, that is oriented graphs, with finitely many vertices and arrows. In less technical words, I will describe some results of applied linear algebra.

Introduction

In this written version of my talk I will present some more or less new results that I may describe with few definitions and many pictures. This note is divided in four sections. Section 1 describes the multiplicity of simple modules in the socle of certain extremely large modules. Section 2 illustrates the socalled tilting equivalence (resp. cotilting duality) induced by a rather small and concrete tilting object, that is by a 6–dimensional vector space. Section 3 describes some infinite dimensional modules. Finally, Section 4 describes the Auslander–Reiten quivers ([AuReS] or [R3]) of certain finite dimensional algebras of finite representation type, that is admitting only finitely many indecomposable modules up to isomorphism. I refer to the notes of Alice Pavarin [P] and Jorge Vitoria [Vi] in the PDF Graduate Seminar for both the first definitions of modules, quivers and representations of a quiver and for the main properties of tilting modules defined over any ring and of any finite projective dimension. For a visual presentation of quivers and tilting modules related to my work, I refer to [D6] and [D7].

I wish to thank Claus Michael Ringel who suggested me the question illustrated in Section 4, and gave me useful hints to solve it.

1 Multiplicities of simple modules

We begin with some definitions needed to state a result on multiplicities. First of all, a module P is projective (resp. I is injective) if P (resp. I) is a direct summand –up to

(*)Università di Milano, Dip. Matematica "Federigo Enriques", Via Cesare Saldini 50, Milano, Italy; E-mail: gabriella.deste@unimi.it . Seminar held on January 20th, 2016.

isomorphism– of every module X such that P is a quotient of X (resp. I is a submodule of X). Given a ring R, a module C is a cogenerator for the category of all R–modules if every module may be embedded in a direct product of copies of C. Next, a projective R–module is isomorphic to a direct summand of a free module, that is of a direct sum of copies of the regular module R . On the other hand, an injective module has a less visible structure. For instance, only over well behaved rings [J, Proposition 3.16] the injective modules are exactly the divisible modules. Over an arbitrary ring R the well - known injective modules are of the form $H = \mathrm{Hom}_Z(R, G)$, where G is any injective abelian group [J, Lemma 2 , page 159]. Moreover, if G contains Q/Z, that is if G is also a cogenerator, then H is an injective cogenerator for the category of all R–modules [Pi, Exercice 5 , page 90]. Finally, we recall that the abelian group of all morphisms between suitable bimodules is a bimodule with the structure defined in [J, Propositions 3.4 and 3.5, page 134]. In particular we know from [Pi, Exercice 5, page 90] that, for any K–algebra A, the left A–module $D(A) = Hom_K(A_A, K)$ is an injective cogenerator with its usual structure [J, Proposition 3.5]. This means that, for any $f \in D(A)$ and $a \in A$, the element af is defined by the formula $(af)(b) = f(ba)$ for all $b \in A$. Following the terminology of [J, page 121], for any simple module S, the homogeneous component determined by S in a semisimple module M is the sum, say L, of all submodules of M isomorphic to S. Moreover, by [J, Corollary 1 , page 119], L is the direct sum of m copies of S, where the cardinal m is called the multiplicity of S in M.

The following statement describes precisely the socle, that is the sum of all its simple submodules, of the dual $D(A)$ of a K–algebra A.

Theorem 1 [D1, Theorem 3] *Let A be an algebra over the field K and let S be a simple left A–module of dimension d over K. Let $D(A) = \mathrm{Hom}_K(A_A, K)$ be the dual of A, and let m be the multiplicity of S in the socle of $D(A)$. Then the following facts hold:*

(i) *If d is finite, then we have $m = d/\dim_K End_A(S)$.*

(ii) *If d is infinite, then we have $m = |K|^d$.*

A similar result holds for the multiplicity of simple modules over any ring. To state the precise result, we need some notation. First of all, for any ring A, let $A^\#$ (resp. $A^\star$) denote the socalled algebraic or topological character module in the sense of Rowen [Ro], that is the following injective cogenerator for the category of all left A–modules: $A^\# = \mathrm{Hom}_\mathbb{Z}(A_A, \mathbb{Q}/\mathbb{Z})$ (resp. $A^* = \mathrm{Hom}_\mathbb{Z}(A_A, \mathbb{R}/\mathbb{Z})$). As we shall see, also in this case, the multiplicity of a simple module S is either rather small or extremely large. However, in this case, the multiplicity depends on the cardinality of S and on the cardinality of $\mathrm{End}_A(S)$.

Theorem 2 [D1, Theorem 6] *Let A be a ring, let H denote one of the injective cogenerators $A^\#$ and $A^\star$. Let S be a simple left A–module, let F denote the smallest subfield of the division ring $\mathrm{End}_A(S)$, and let m denote the multiplicity of S in the socle of H. Then the following facts hold:*

(i) *If S is finite, then $m = \dim_F S/\dim_F \mathrm{End}_A(S)$.*

(ii) *If S is infinite, then $m = 2^{|S|}$.*

With the previous notation we clearly have $A^{\#} \subseteq A^{\star}$ and the following result holds.

Corollary 3 [D2, Theorem 1] *Let A be a K–algebra and let L be a subfield of K. Then all the injective cogenerators $\mathrm{Hom}_L(A_A, L)$, $A^{\#}$ and $A^{\star}$ are isomorphic to a direct products of copies of $D(A)$ and there is an embedding of the form $\mathrm{Hom}_L(A_A, L) \longrightarrow A^{\#}$.*

2 A toy example of a finite dimensional bimodule

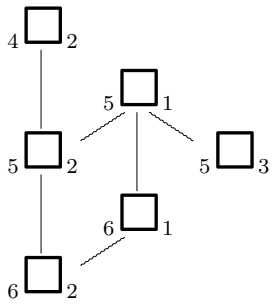
Example 4 Throughout this section A and B denote the K–algebras given by the quivers

respectively. In other words, A is isomorphic to the algebra of all 3×3 lower triangular matrices with entries in K, while B is isomorphic to the algebra of all 3 by 3 upper triangular matrices with entries in K of the form

$$\begin{pmatrix} \star & \star & \star \\ 0 & \star & 0 \\ 0 & 0 & \star \end{pmatrix}$$

The next picture visualizes the structure of an $A - B$ bimodule U of dimension 6 over K . As a left (resp. right) module, U is the direct sum of 3 indecomposable summands denoted by $\begin{smallmatrix} 4 \\ 5 \\ 6 \end{smallmatrix}$, $\begin{smallmatrix} 5 \\ 6 \end{smallmatrix}$, 5 , (resp. 2 , $\begin{smallmatrix} 1 \\ 2\,3 \end{smallmatrix}$, $\begin{smallmatrix} 1 \\ 2 \end{smallmatrix}$).

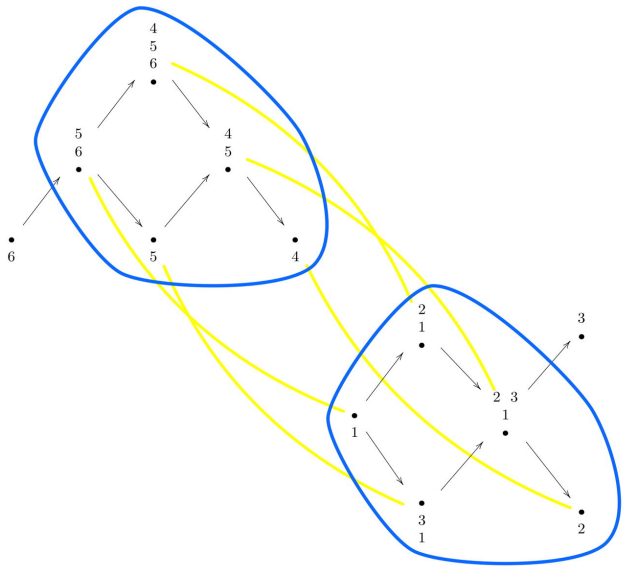
Figure 1



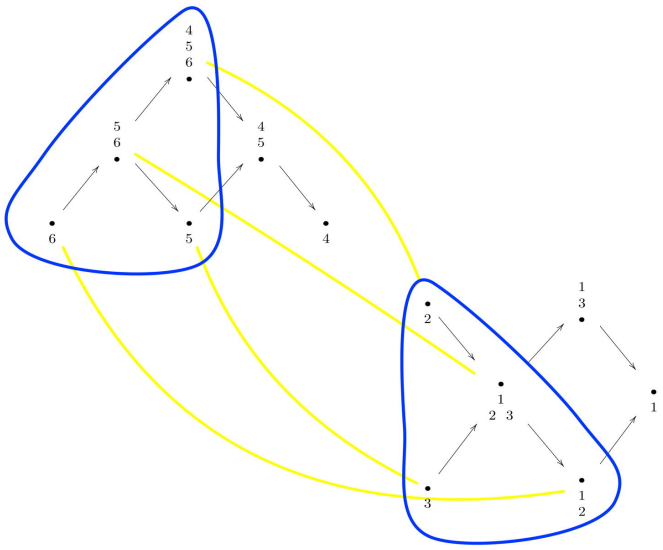
In both cases, the first two summands are projective, while the third is the quotient of two projective modules. In other words, U has projective dimension one on either side. Dually, $\begin{smallmatrix} 4 \\ 5 \\ 6 \end{smallmatrix}$ and $\begin{smallmatrix} 1 \\ 2 \end{smallmatrix}$ are injective, while any of the remaining four summands X is a submodule of an injective module E and the factor module E/X is injective. In other words, U has

injective dimension one on either side. By a direct check, or by the structure of the Auslander–Reiten quivers of A and B and by a well-known formula about extensions and morphisms ([R3, (6) page 76] or [AuReS, Corollary 4.7 , page 132]), we conclude that $\mathrm{Ext}^1_?(U,U) = 0$ where $? = A, B$. Since three is the number of the pairwise non isomorphic direct summands of U, it follows [R3, page 167] that U is a tilting and cotilting module in the sense of Brenner and Butler [BB]. We refer to [D5, Example D] or [D7, Section 4.3] for other properties of U with respect to left and right injective envelopes [J, page 163].

The next picture describes the tilting equivalence, that is one of the equivalences described by the Tilting Theorem [CF], represented by the bimodule U , between the class of modules generated by U as a left A-module and the class of modules cogenerated by the left B-module $\mathrm{Hom}_K(U_B, K)$, the dual of the right B-module U. In this case, by graph theoretical reasons, this tilting equivalence is the unique equivalence between the above classes.



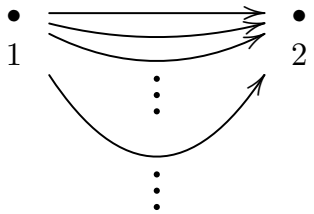
The next picture describes the cotilting duality, that is the duality described by the Cotilting Theorem [CF], induced by U between the class of modules cogenerated by U as a left A-module and the class of modules cogenerated by U as a right B-module. Also in this case, by graph theoretical reasons, this cotilting duality is the unique duality between the above classes.



3 Some examples of infinite dimensional modules.

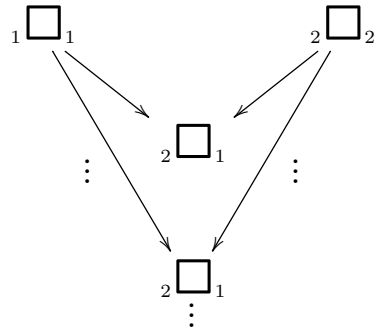
Concerning the importance of direct sums, we recall the following remark of Vamos [V, page 476] on direct sums: "The principle behind this framework is our belief that the only really well – understood construction is the direct sum decomposition." I believe this is true for many reasons. Indeed, the "dual" construction (that of making direct products) may be much more complicated, even in case of countably many factors. For instance, even the direct product of infinitely many factors, all isomorphic to a fixed indecomposable module M of countable dimension over some field K, may have indecomposable direct summands not isomorphic to M, that is M fails to be product complete in the sense of Krause and Saorin [KS]. To see this, it suffices to consider the following example.

Example 5 Let $P(1) = Ae(1)$, where A is the algebra $\left(\begin{smallmatrix} K & 0 \\ V & K \end{smallmatrix}\right)$, $e(1) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ and V is a K-vector space of infinite and countable dimension. In other words A is the algebra given by a quiver of the form



with infinitely may arrows from 1 to 2 . Then any direct product of infinitely many copies of $P(1)$ has a direct summand isomorphic to the simple projective module $P(2) = Ae(2)$ with $e(2) = \left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ [D3, Lemma 2.1, first part of the proof of (ii)]. The next picture describes the structure of the regular $A - A$ bimodule A, which induces a cotilting duality

(between finitely generated projective left and right modules), sending simple modules to indecomposable non simple ones and, conversely, indecomposable non simple modules to simple ones.



Since the unique maximal submodule of any non simple indecomposable projective module is not reflexive with respect to A [D4, Theorem 2.5], it follows that reflexive modules are not closed under submodules. We add a remark on the dual construction (that of making direct products). Only in very special cases direct products with a component - wise structure describe indecomposable injective modules, as the next example shows.

Example 6 Let B be algebra given by the quiver $a \circlearrowright \bullet_1 \xrightarrow{b} \bullet_2$, and let $I(2)$ denote the injective envelope of the simple module $S(2)$ of the form $\circlearrowright 0 \longrightarrow K$. Then a direct check shows that $I(2)$ is of the form $a \circlearrowright K^{\mathbb{N}} \xrightarrow{b} K$, where the action of a and b on $K^{\mathbb{N}}$ is defined as follows:

$$a(k(0), k(1), k(2), \dots) = (k(1), k(2), k(3), \dots) \qquad \text{and}$$
$$b(k(0), k(1), k(2), \dots) = k(0).$$

Sometimes, direct products of one dimensional vector spaces with multiplications component - wise defined, describe big proper submodules (of uncountable dimension) of indecomposable injective modules. We give an example suggested by Example 6.

Example 7 Let A denote the free K-algebra $K\langle X, Y\rangle$ in two non commutative variables X and Y, and let S denote the simple left A-module $A/\langle X, Y\rangle$. Next let L denote the left A-module $K^{\mathbb{N}}$ such that X and Y act component-wise as follows:

$$X(k(0), k(1), k(2), \dots) = (0, k(2), k(3), \dots) \quad \text{and}$$
$$Y(k(0), k(1), k(2), \dots) = (k(1), 0, 0, 0, \dots).$$

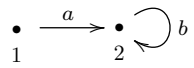
Consequently, for any non–zero element v of $K^{\mathbb{N}}$, we have $(1, 0, 0, 0, \dots) \in Av$. Hence S is the simple essential socle of L, and so L may be embedded in the injective envelope $E(S)$ of S. It suffices to interchange X and Y in the construction of L to obtain a module M, not isomorphic to a submodule of L, admitting an essential socle isomorphic to S. Therefore L fails to be an injective module.

We end this section with the following open (graph theoretical?) problem.

Problem 8 *What is the structure of $E(S)$?*

4 Some examples of finite Auslander-Reiten quivers

Throughout this section, let Q be a quiver of the form



with relations $b^m = 0$ and $b^2a = 0$, with $m > 1$.

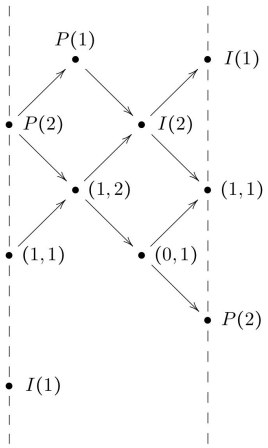
Let A denote the K-algebra given by Q, and let $e(x)$ denote the path of length zero around the vertex x. Then a direct calculation of the whole Auslander-Reiten quiver, by means of the dual of the transpose [AuReS] $\tau(M)$ of any indecomposable non projective module M, shows the following:

- If $m = 2, 3, 4, 5$, then the number of isomorphism classes of left A-modules is equal to 7, 14, 28, 66 respectively.

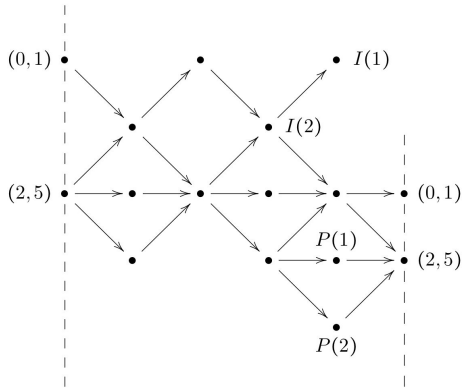
On the other hand, if $m = 6$, then there exist infinitely many indecomposable modules [R1], because almost all, but finitely many, indecomposable representations of the Euclidean diagram $\widetilde{E}_8$ give rise to indecomposable representations of Q.

We end with the picture of the Auslander-Reiten quivers with 7, 14, 28 and 66 vertices mentioned above.

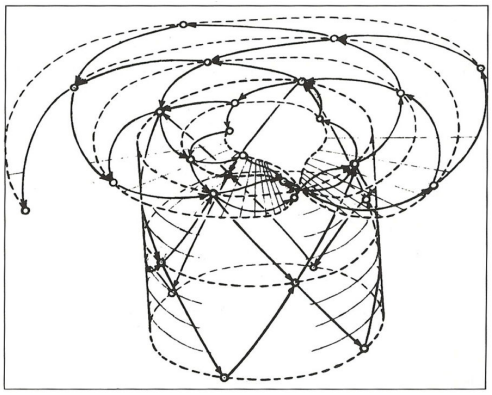
If $m = 2$, then there are 3 stable indecomposable modules, that is belonging to a closed τ-orbit. On the other hand, the 4 indecomposable modules which are either projective or injective are exactly the unstable modules, as indicated in the following Auslander-Reiten quiver. In the picture we have to identify the dotted lines and we replace the 3 stable indecomposable modules M by their dimension vectors [R3], that is by the pairs $(\dim_K e(1)M, \dim_K e(2)M)$.



If $m = 3$, then all the 14 indecomposable modules are unstable and the two τ-orbits have 6 and 8 elements respectively. In this case the Auslander-Reiten quiver has the following shape, where again we have to identify the dotted lines.



If $m = 4$, then 20 indecomposable modules are stable and 8 are unstable. For an elegant and topological form of the Auslander-Reiten quiver, we refer to Ringel's paper [R2, page 93]. We take from Ringel's home page [R4, Abbildung 3] the following picture of this Auslander-Reiten quiver.



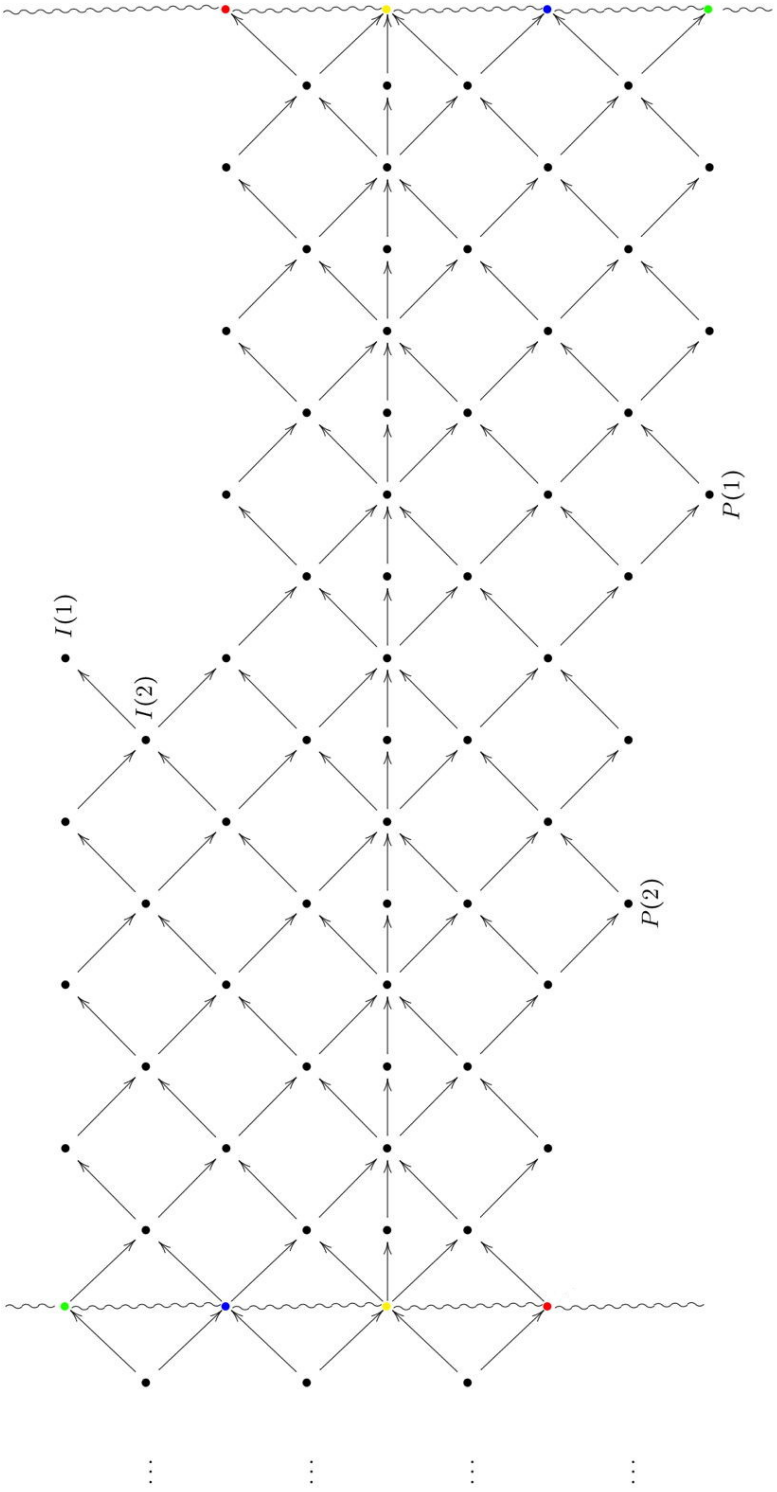
We refer to Section 3.1 of [D6] for a naïve description of the same Auslander-Reiten quiver, admitting

- four stable τ-orbits with 4, 4, 4 and 8 elements;
- two unstable τ-orbits with 3 and 5 elements.

If $m = 5$, then there exist

- 48 stable modules, belonging to four τ-orbits with 8, 8, 16 and 16 elements;
- 18 unstable modules, belonging to two τ-orbits with 8 and 10 elements.

The following picture illustrates the shape of the Auslander-Reiten quiver.



References

[AuReS] M. Auslander, I. Reiten and S. O. Smalø, "Representation theory of Artin algebras". Cambridge University Press, 1995.

[BB] S. Brenner and M. C. R. Butler, *Generalizations of the Bernstein-Gelfand-Ponomarev reflection functors.* Representation Theory II (Ottawa, 1979), V. Dlab and P. Gabriel (eds.), Lec. Notes in Math., 832, Springer, Berlin (1980), 103–169.

[CF] R. R. Colby and K. R. Fuller, "Equivalence and Duality for Module Categories". Cambridge University Press, 2004.

[D1] G. D'Este, *Simple submodules and multiplicities.* Forum Math. 4 (1992), 1–11.

[D2] G. D'Este, *Standard duals of algebras.* Comm. Algebra 20 (12) (1992), 3503–3513.

[D3] G. D'Este, *Free modules obtained by means of infinite direct products.* International Conference Algebra and Its Applications, AMS 259 (2000), 161–174.

[D4] G. D'Este, *Reflexive modules are not closed under submodules.* Representations of algebras (Sao Paulo), LNPAM 224, M. Dekker (2002), 53–64.

[D5] G. D'Este, *Symmetries and asymmetries for cotilting bimodules.* Proceedings of the Venice Algebra Conference, M. Dekker 236 (2004), 103–118.

[D6] G. D'Este, *Unofficial history of a joint work with Dieter Happel and of two unexpected quotations.* Rend. di Matematica, Serie VII, 5 (2014), 115–130, `http://arxiv.org/abs/1401.2085`.

[D7] G. D'Este, *Recent and less recent results on Tilting Theory.* `http://arxiv.org/abs/1411.4418`.

[J] N. Jacobson, "Basic Algebra II". Freeman and C., San Francisco 1980.

[KS] H. Krause and M. Saorìn, *On minimal approximations of modules.* Contemp. Math. 229 (1998), 227–236.

[P] A. Pavarin, *Recollements of infinitely generated tilting modules.* Seminario Dottorato 2012-13, 9–17.

[Pi] R. S. Pierce, "Associative algebras". Springer GTM 8, 1982.

[R1] C. M. Ringel. Private communication to the author.

[R2] C. M. Ringel, *Unzerlegbare Darstellungen endlich-dimensionaler Algebren.* Jber. d. Dt. Math. Verein. 85 (1983), 86–105.

[R3] C. M. Ringel, "Tame algebras and integral quadratic forms". Springer LMN 1099 (1984).

[R4] C. M. Ringel, *Abbildungen zu Diskrete Methoden in der Darstellungstheorie.* Bielefelder Universitätszeitung (1992), `www.math.uni-bielefeld.de/.../diskret/pictur.....`

[Ro] L. H. Rowen, "Ring Theory". Volume 1, Pure and Applied Math. 127, Academic Press, 1988.

[V] P. Vamos, *The Holy Grail of algebra: seeking complete sets of invariants.* Proceedings of the Padova Algebra Conference. Mathematics and Its Applications 343, Kluwer (1995), 475–490.

[Vi] J. Vitoria, *A visual introduction to Tilting.* Seminario Dottorato 2013-14, 147–155.

On the behavior of membranes and plates upon perturbations of shape and density

Luigi Provenzano (*)

Abstract. In this article we consider eigenvalue problems for second and fourth order partial differential operators. Such problems arise from the study of the transverse vibrations of thin membranes and plates, respectively. We are interested in the behavior of the normal modes of vibration (i.e., the eigenvalues) upon variations of the shape and the density of the membrane/plate. In particular, we shall consider the issue of the optimization of the eigenvalues depending on such parameters, under suitable constraints (of fixed volume or mass, for example). The presentation is of introductory type and is intended for a general reader, no matter the field of expertise.

1 Introduction

An ubiquitous object in Mathematical Analysis is the Laplace operator, which is a differential operator of order 2, acting on functions u on $\mathbb{R}^N$ and which is defined by

$$\Delta u := \sum_{i=1}^{N} \frac{\partial^2 u}{\partial x_i^2}.$$

This object has been widely studied for centuries and has applications in many different fields of Mathematics and applied sciences, e.g., it is extremely important in mechanics, electromagnetics, wave theory, quantum mechanics, statistical mechanics and many more. For an introduction on the Laplace operator, the Laplace equation, and their properties, we refer to [23, 25]. Less known is the biharmonic operator, or bi-Laplacian, defined by

$$\Delta^2 u := \Delta\Delta u = \sum_{i,j=1}^{N} \frac{\partial^4 u}{\partial x_i^2 \partial x_j^2},$$

which is an operator of order 4. Even if at a first glance the Laplace operator seems to have an intimate relation with the biharmonic operator, actually the two operators behave

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: prozhunter@gmail.com . Seminar held on February 17th, 2016.

in very different ways. In particular, the theory of linear elliptic second order equations (the Laplace operator is the prototype of a second order linear elliptic operator) is well developed and nowadays classical (see e.g. [27]). On the contrary, there are much less results for higher order differential equations. Among the various reasons, higher order operators like the biharmonic operator do not enjoy the maximum principle, which is a fundamental property of the Laplacian. We refer to the monography [26] for a comprehensive exposition on the state of the art of the theory of poly-harmonic operators.

In this notes we consider eigenvalue problems for the Laplace and the biharmonic operator, namely the problems

$$-\Delta u = \lambda u \quad \text{and} \quad \Delta^2 u = \lambda u\,, \quad \text{in } \Omega,$$

in the unknowns $u \in C^2(\Omega)$ and $u \in C^4(\Omega)$, respectively (the eigenfunction) and $\lambda \in \mathbb{R}^N$ (the eigenvalue), where Ω is a bounded domain (i.e., an open connected subset, see Figure 1 of $\mathbb{R}^N$, subject to suitable boundary conditions.

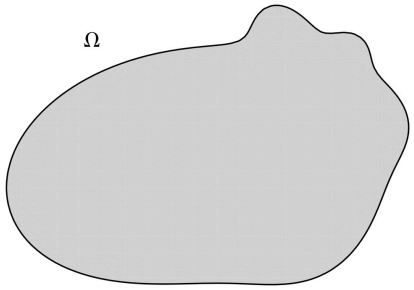


Figure 1. A domain Ω.

When $N = 2$, equations $-\Delta u = \lambda u$ and $\Delta^2 u = \lambda u$ arise in the study of a *thin vibrating membrane* and a *thin vibrating plate*, respectively, whose position at rest is described by the domain Ω. Given a solution $(u, \lambda) \in C^2(\Omega) \times \mathbb{R}$ and $(u, \lambda) \in C^4(\Omega) \times \mathbb{R}$ respectively, the eigenfunction u represents a normal mode of vibration of the membrane/plate, while the eigenvalue λ represents the square of the corresponding vibrational frequency (see Figure 2).

In order to find a solution u to the eigenvalue equations, we need to impose that u satisfies suitable homogeneous boundary conditions. We consider first the case of the Laplacian. We have the eigenvalue problem for the Laplace operator subject to Dirichlet boundary conditions, namely the problem

$$\begin{cases} -\Delta u = \lambda \rho u, & \text{in } \Omega, \\ u = 0, & \text{on } \partial\Omega, \end{cases} \tag{1.1}$$

in the unknowns $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$, $\lambda \in \mathbb{R}$. We also have the eigenvalue problem for the Laplace operator subject to Neumann boundary conditions, namely the problem

$$\begin{cases} -\Delta u = \lambda \rho u, & \text{in } \Omega, \\ \frac{\partial u}{\partial \nu} = 0, & \text{on } \partial\Omega, \end{cases} \tag{1.2}$$

in the unknown $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$, $\lambda \in \mathbb{R}$. Here $\partial\Omega$ denotes the boundary of Ω and ν denotes the outer unit normal to Ω. Hence, $\partial u/\partial\nu$ is the derivative of u along the unit outer normal direction (usually called the normal derivative). The function ρ is a measurable, positive and bounded function and represents the density of the membrane, i.e., how the mass is displaced on it (if $\rho \equiv 1$ we have a homogeneous membrane with density constantly equals to 1). We refer to the quantity $M := \int_\Omega \rho dx$ as the total mass of the membrane.

Problem (1.1) models a thin vibrating membrane with a *fixed frame*, while problem (1.2) models a thin vibrating membrane with a *free frame*. It is standard to prove that problems (1.1) and (1.2) admit an increasing sequence of non-negative eigenvalues of finite multiplicity

$$0 \le \lambda_1 < \lambda_2 \le \cdots \le \lambda_j \le \cdots \nearrow +\infty.$$

We remark that in the case of problem (1.2) in order to have the discreteness of the spectrum we need that Ω has at least a Lipschitz boundary. We note that in the case of problem (1.1) $\lambda_1 > 0$, while in the case of problem (1.2) $\lambda_1 = 0$, $\lambda_2 > 0$. We refer to [19] for the characterization of the spectrum of problems (1.1) and (1.2).

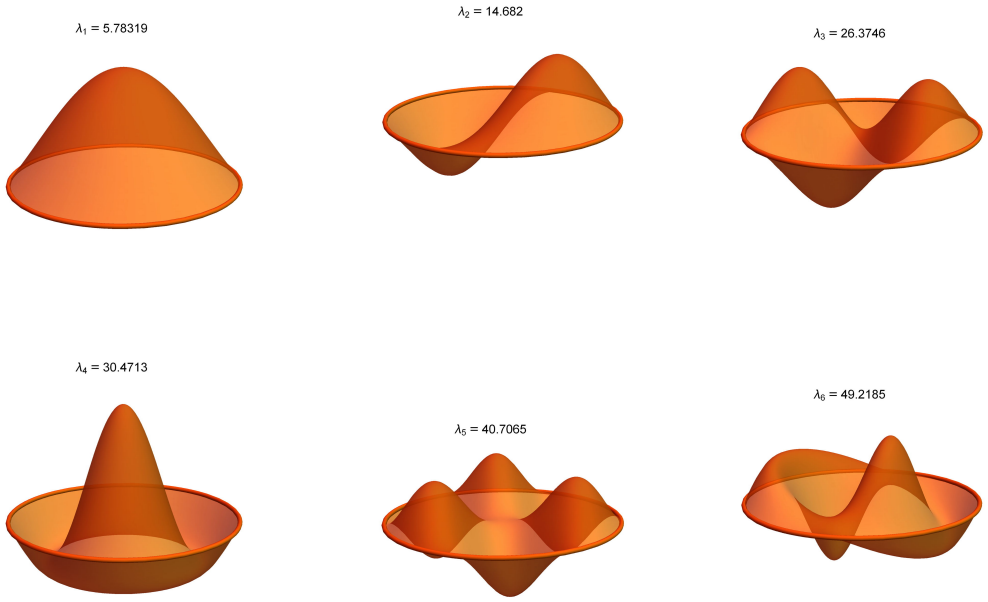


Figure 2. First six eigenfunctions of the Dirichlet Laplacian on the unit disk in $\mathbb{R}^2$.

We consider now the biharmonic operator. We have the eigenvalue problem for the biharmonic operator subject to Dirichlet boundary conditions, namely the problem

$$\begin{cases} \Delta^2 u = \lambda \rho u, & \text{in } \Omega, \\ u = \frac{\partial u}{\partial \nu} = 0, & \text{on } \partial\Omega, \end{cases} \tag{1.3}$$

in the unknowns $u \in C^4(\Omega) \times C^1(\overline{\Omega})$, $\lambda \in \mathbb{R}$. We also have the eigenvalue problem for the biharmonic operator subject to Neumann boundary conditions, namely the problem

$$\begin{cases} \Delta^2 u = \lambda \rho u, & \text{in } \Omega, \\ \frac{\partial^2 u}{\partial \nu^2} = \operatorname{div}_{\partial\Omega}\left(D^2 u \cdot \nu\right) + \frac{\partial \Delta u}{\partial \nu} = 0, & \text{on } \partial\Omega, \end{cases} \tag{1.4}$$

in the unknowns $u \in C^4(\Omega) \cap C^3(\overline{\Omega})$, $\lambda \in \mathbb{R}$. In the case of the biharmonic operator we have also intermediate boundary conditions, namely

$$\begin{cases} \Delta^2 u = \lambda \rho u, & \text{in } \Omega, \\ u = \frac{\partial^2 u}{\partial \nu^2} = 0, & \text{on } \partial\Omega, \end{cases} \tag{1.5}$$

in the unknowns $u \in C^4(\Omega) \cap C^2(\overline{\Omega})$, $\lambda \in \mathbb{R}$.

Here $\operatorname{div}_{\partial\Omega} F$ denotes the tangential divergence of a vector field F and is defined by $\operatorname{div}_{\partial\Omega} F := \operatorname{div} F|_{\partial\Omega} - (DF \cdot \nu) \cdot \nu$, and $D^2 u$ denotes the Hessian matrix of u. Again, the function ρ is a measurable, positive and bounded function which represents the density of the plate. The quantity $M := \int_\Omega \rho dx$ is therefore total mass. We refer to [19, 26] for the derivation of the boundary conditions of problems (1.3) and (1.5). We refer to [18] for the derivation of the boundary conditions in (1.4).
Problems (1.3), (1.4) and (1.5) model respectively clamped, free and hinged vibrating plates. In Figure 3 we have two sections of a vibrating plate. The left ends of both sections are clamped. The right end of the first section if hinged. The right end of the second section is free.

Figure 3. Sections of vibrating plates.

Also in this case, problems (1.3), (1.4) and (1.5) admit an increasing sequence of non-negative eigenvalues of finite multiplicity

$$0 \le \lambda_1 \le \lambda_2 \le \cdots \le \lambda_j \le \cdots \nearrow +\infty$$

We remark that in the case of problem (1.4) in order to have the discreteness of the spectrum we need that Ω has at least a Lipschitz boundary. We note that for problems (1.3) and (1.5) $\lambda_1 > 0$, while for problem (1.4) $\lambda_1 = \lambda_2 = \cdots = \lambda_{N+1} = 0$, $\lambda_{N+2} > 0$.

Eigenvalues and eigenfunctions are important in *linear elasticity*. We have seen that the eigenfunctions represent the natural modes of vibration of a membrane or a plate, while the eigenvalues are the squares of the corresponding vibrational frequencies. Then every possible vibration of the membrane/plate can be described in terms of the natural modes of vibration (see [19] for an exhaustive discussion).

Remark 1.6 Problems (1.1), (1.2), (1.3), (1.4) and (1.5) make sense for any $N \ge 2$ (and also for $N = 1$). Thus in what follows we shall not make any restriction on the space dimension.

It is important, in several fields of engineering (e.g., in the construction of suspeded bridges) to study the dependence of the eigenvalues and eigenfunctions upon some parameters which enter the equations, in particular, the shape Ω and the density ρ. Hence, we are interested in the maps

$$\rho \mapsto \lambda_j[\rho]$$

and

$$\Omega \mapsto \lambda_j[\Omega]$$

We ask the following questions: are these maps continuous? Differentiable? Analytic? What it is possible to say about

$$\max_{|\Omega|=\text{const.}} / \min_{|\Omega|=\text{const.}} \lambda_j[\Omega]$$

$$\max_{\int_\Omega \rho=\text{const.}} / \min_{\int_\Omega \rho=\text{const.}} \lambda_j[\rho]$$

or other critical points? (here $|\Omega|$ denotes the Lebesgue measure of Ω).

There is a huge literature on the dependence of the eigenvalues upon shape or density. We refer e.g., to the book [28] for a quite exhaustive discussion on the state of the art of shape and density optimization problems.

2 Classical results in shape and density optimization

In this section we recall some classical results in shape and density optimization. We start by recalling some results of shape optimization for the eigenvalues. In this case the density is fixed and equals 1. The fundamental question in shape optimization is the following: "*are there optimal sets, (maximizers or minimizers) for the eigenvalues under the constraint that the measure of* Ω *is fixed?*".

The most famous result is perhaps the Faber-Krahn inequality for the first eigenvalue of problem (1.1) (see [24, 31]).

Theorem 2.1 (Faber-Krahn) *Let* Ω *be an open set in* $\mathbb{R}^N$ *of finite measure. Let* $\lambda_1[\Omega]$ *be the first eigenvalue of problem (1.1). Then*

$$\lambda_1[\Omega^*] \leq \lambda_1[\Omega],$$

where Ω^* *is a ball with the same measure as* Ω*. The equality holds only if* $\Omega = \Omega^*$ *is a ball.*

The Faber-Krahn inequality states that among all open sets with a fixed measure, the ball is the unique minimizer of the first Dirichlet eigenvalue of the Laplace operator. We have an analogous result for problem (1.2), due to Szegö and Weinberger (see [42, 43]).

Theorem 2.2 (Szegö-Weinberger) *Let* Ω *be bounded domain of class* C^1 *in* $\mathbb{R}^N$*. Let* $\lambda_2[\Omega]$ *be the first positive eigenvalue of problem (1.2). Then*

$$\lambda_2[\Omega^*] \geq \lambda_2[\Omega],$$

where Ω^* is a ball with the same measure as Ω. The equality holds only if $\Omega = \Omega^*$ is a ball.

The Szegö-Weinberger inequality states that among all bounded domains of class C^1 with a fixed measure, the ball is the unique maximizer of the first positive Neumann eigenvalue of the Laplace operator. As for the biharmonic operator, much less is known. We have the following inequality for the first eigenvalue of problem (1.3), which holds only in dimension $N = 2$ and $N = 3$ (the proof for dimension $N = 2$ was obtained by Nadirashvili and soon generalized to dimension $N = 3$ by Ashbaugh and Benguria, see [39] and [4]).

Theorem 2.3 (Nadirashvili, Ashbaugh-Benguria) *Let Ω be an open set in $\mathbb{R}^N$, with $N = 2, 3$, of finite measure. Let $\lambda_1[\Omega]$ be the first eigenvalue of problem (1.3). Then*

$$\lambda_1[\Omega^*] \leq \lambda_1[\Omega],$$

where Ω^* is a ball with the same measure as Ω. The equality holds only if $\Omega = \Omega^*$ is a ball.

The Nadirashvili-Ashbaugh-Benguria inequality states that in dimensions $N = 2$ and $N = 3$, among all open sets with a fixed measure, the ball is the unique minimizer of the first Dirichlet eigenvalue of the biharmonic operator. The problem is still open for $N \geq 4$. The conjecture is that the ball is a minimizer for all dimensions N. This is the famous Rayleigh's Conjecture.

Now we recall some classical results on density optimization. The fundamental question in density optimization is the following: "*are there optimal densities (maximizers or minimizers) for the eigenvalues under the constraint that the mass is fixed?*".

We start with the following one-dimensional result, due to Krein. Optimal densities for all the eigenvalues of the one-dimensional Laplacian which preserve the total mass exist, provided they satisfy the additional assumption that they are bounded from below and above by two positive constants A and B. More precisely we have the following theorem (see [32]).

Theorem 2.4 (Krein) *Let $\Omega =]0, \pi[$. Then there exist maximizers and minimizers in $L^\infty(\Omega)$ for all the eigenvalues of problems (1.1) and (1.2) on Ω under the constraints*

$$\int_0^\pi \rho dx = \text{const.}$$

and

$$A \leq \rho \leq B, \quad \text{a.e.},$$

where $A, B \in \mathbb{R}$, $0 < A < B$.

In this case it is shown that the optimal densities are 'bang-bang' (see Figures 4 and 5, which means that they are of the form

$$\rho = A\chi_\omega + B\chi_{\Omega \setminus \overline{\omega}},$$

for a suitable $\omega \subseteq \Omega$.

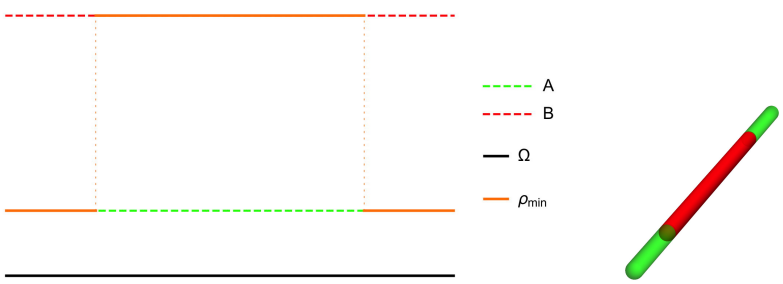


Figure 4. Minimizer of the first Dirichlet eigenvalue on $]0,\pi[$.

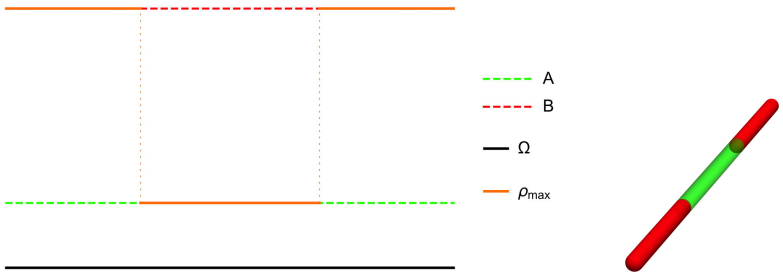


Figure 5. Maximizer of the first Dirichlet eigenvalue on $]0,\pi[$.

For $N \geq 2$, Cox and Mc.Laughlin generalized the result of Krein in the case of Dirichlet boundary conditions. We have the following theorem (see [20, 21]).

Theorem 2.5 (Cox-Mc.Laughlin) *Let Ω be a bounded domain in $\mathbb{R}^N$. Then there exist maximizers and minimizers for all the eigenvalues of problem (1.1) on Ω which preserve the total mass and which satisfy $A \leq \rho \leq B$. Moreover such critical points are of the form*

$$A\chi_\omega + B\chi_{\Omega\setminus\overline{\omega}},$$

for a suitable $\omega \subset \Omega$.

A complete description of the optimal densities as in the case $N \geq 1$ is in general unavailable. We refer to [28] and to the references therein for an updated collection of results in shape and density optimization.

3 The Steklov eigenvalue problem

In this section we introduce the Steklov eigenvalue problem for the Laplace operator. Let Ω be a bounded domain in $\mathbb{R}^N$ with a Lipschitz boundary. We consider the following problem

$$\begin{cases} \Delta u = 0, & \text{in } \Omega, \\ \frac{\partial u}{\partial \nu} = \lambda \rho u, & \text{on } \partial\Omega, \end{cases} \tag{3.1}$$

in the unknowns $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$, $\lambda \in \mathbb{R}$. In this case ρ is a measurable, bounded and positive function defined on the boundary $\partial\Omega$. We note that in problem (3.1) the eigenvalue appears in the boundary conditions. Also in this case it is possible to prove that the spectrum of problem (3.1) is discrete and consists in a sequence of non-negative eigenvalues of finite multiplicity which we denote by

$$0 = \lambda_1 < \lambda_2 \leq \cdots \leq \lambda_j \leq \cdots \nearrow +\infty.$$

For $N = 2$ this problem models the vibrations of a free membrane with mass displaced at the boundary with surface density ρ. For the proof of the discreteness of the spectrum and for more information on the properties of problem (3.1) we refer to the paper [41] where the problem has been introduced for the first time.

We recall now two important results in shape and density optimization for the first positive eigenvalue λ_2 of problem (3.1). We start with the a result in shape optimization (see [9, 44]).

Theorem 3.2 (Brock-Weinstock) *Let Ω be a bounded domain of class C^1 in $\mathbb{R}^N$. Let $\lambda_2[\Omega]$ be the first positive eigenvalue of problem (3.1) in Ω. Then*

$$\lambda_2[\Omega^*] \geq \lambda_2[\Omega],$$

where Ω^* is a ball with the same measure as Ω. Equality holds only if $\Omega = \Omega^*$ is a ball.

The Brock-Weinstock inequality states that among all bounded domains of class C^1 in $\mathbb{R}^N$ with fixed measure, the ball is the unique maximizer of the first positive Steklov eigenvalue. We note that with respect to shape optimization, the Steklov problem shares a similarity with the Neumann problem. We also recall that both problems arise in the study of vibrating membranes with a free frame.

Now we present a result in density optimization (see [29]).

Theorem 3.3 (Hersch-Payne-Schiffer) *Let B be the unit disk in $\mathbb{R}^2$. Let $M > 0$ and let $\rho \in L^\infty(B)$ be such that $\operatorname{ess\,inf}_B \rho > 0$ and that $\int_B \rho dx = M$. Let $\lambda_2[\rho]$ be the first positive eigenvalue of problem (3.1) with density ρ on B. Then*

$$\lambda_2[M/2\pi] \geq \lambda_2[\rho],$$

where $\lambda_2[M/2\pi]$ is the first positive eigenvalue of problem (3.1) with constant density $\rho \equiv M/2\pi$. The equality is attained only if $\rho \equiv M/2\pi$.

We note that the constant surface density is the unique maximizer for the first positive Steklov eigenvalue among all densities which preserve the total mass M. Hence, for problem (3.1), a maximizer exists under the sole contraint that the mass is fixed. The same result is conjectured to be true also for $N \geq 3$ but it is still not proved. We note that problem (3.1) behaves in different way from the problems considered in Section 2 with respect to mass density perturbations.

We refer to [5, 28] for more information and results on the eigenvalues of the Steklov problem.

4 Updates in shape and density perturbations. A new biharmonic Steklov problem

In this section we present some recent results on shape and density perturbation and optimization problems obtained in collaboration with D. Buoso (Politecnico di Torino), L.M. Chasman (University of Minnesota, Morris), M. Dalla Riva (The University of Tulsa, OK) and P.D. Lamberti (Università degli Studi di Padova).

First, we consider mass density perturbation problems, in particular the problem of characterizing the critical mass densities preserving the total mass, without any additional constraint. This lead us state a 'maximum principle' in spectral perturbation problems. Then we investigate some particular mass displacements, i.e., mass densities which concentrate at the boundary of Ω. We consider the eigenvalues of the Neumann problem for the Laplace operator with mass concentrating near the boundary and we prove that at the limit, the Neumann eigenvalues converge to the Steklov eigenvalues. Thus, the Steklov eigenvalues can be considered as limiting Neumann eigenvalues in a mass concentration phenomenon. Moreover, we consider the analogous mass concentration phenomenon for the biharmonic Neumann problem. As a bypass product, we obtain a fourth order Steklov problem which is the analogue for the biharmonic operator of the classical Steklov problem. Then, we consider the dependence of the eigenvalues of this new problem upon the domain Ω. We provide Hadamard-type formulas for the derivatives of the eigenvalues which we use to characterize critical sets. We prove that balls are critical sets for all simple eigenvalues and all symmetric functions of multiple eigenvalues under measure constraint. Moreover, we prove that the ball is actually the unique maximizer for the first positive eigenvalue among all bounded domains with a fixed measure. We provide a quantitative version of such an isoperimetric inequality which turns out to be sharp.

4.1 Mass density perturbations

Let Ω be a bounded domain in $\mathbb{R}^N$. Let $\mathcal{R} := \{\rho \in L^\infty(\Omega) : \operatorname{ess\,inf}_\Omega \rho > 0\}$. We consider the map from $\mathcal{R}$ to $\mathbb{R}$ which maps $\rho \in \mathcal{R}$ to $\lambda_j[\rho]$, where $\lambda_j[\rho]$ denotes the j-th eigenvalue of problems (1.1), (1.3) or (1.5). Then it is possible to prove that $\lambda_j[\rho]$ is a locally Lipschitz-continuous function of ρ (see [33] for more details). Moreover, the eigenvalues are continuous not only with respect to the strong topology of $L^\infty(\Omega)$, but also with respect to the weak* topology, which is more relevant in optimization problems.

Lemma 4.1 *Let $\rho \in \mathcal{R}$ and let $\lambda_j[\rho]$ be an eigenvalue of problems (1.1), (1.3) or (1.5). Let C be a bounded subset of $\mathbb{R}$. Then the map from C to R defined by*

$$\rho \mapsto \lambda_j[\rho]$$

is continuous with respect to the weak topology of $L^\infty(\Omega)$.*

We note that the subsets of densities such that $A \le \rho \le B$ are weakly* compact in $L^\infty(\Omega)$.

Remark 4.2 The sets of densities $\{\rho \in L^\infty(\Omega) : A \le \rho \le B\}$ for fixed $A, B > 0$ are weakly* compact. Then on such sets there exist maximizers and minimizers for all the eigenvalues.

It is now clear why under the additional assumption $A \le \rho \le B$ maximizers and minimizers exist.

Now we consider the issue of the analyticity of the eigenvalues. It is well-known that the eigenvalues of differential operators depending on more than one real parameter are in general not differentiable with respect to the parameters. This is due to the well-known bifurcation phenomena which occur when multiple eigenvalues split into simple eigenvalues, or vice-versa, when multiple eigenvalues collapse into a simple one. Such phenomena prevent the eigenvalues to be even differentiable functions of the parameters involved in the equation. We refer to [30, 40] for an introduction to perturbation theory for elliptic operators.

It has been pointed out in [37, 38] that, in order to prevent such phenomena, the correct quantities to be considered in spectral perturbation problems are the elementary symmetric functions of the eigenvalues, more than the eigenvalues themselves (in the case of multiple eigenvalues). We need some preliminary definitions. Let F be a non-empty finite subset of $\mathbb{N}$. Let

$$\mathcal{R}[F] := \{\rho > 0 : \lambda_j[\rho] \neq \lambda_l[\rho]\,,\ \forall j \in F, l \in \mathbb{N} \setminus F\}\,.$$

Roughly speaking, $\mathcal{R}[F]$ is the set of densities which 'preserve the multiplicity' of the eigenvalues. For example, if $F = \{1\}$, then $\mathcal{R}[F] = \{\rho > 0 : \lambda_1[\rho] \text{ is simple}\}$. Then we consider the elementary symmetric functions of the eigenvalues, defined by

$$\Lambda_{F,h}[\rho] := \sum_{\substack{j_1,\dots,j_h \in F \\ j_1<\dots<j_h}} \lambda_{j_1}[\rho]\cdots\lambda_{j_h}[\rho]\,, \quad h = 1, \dots, |F|$$

We provide now a simple finite dimensional example in order to motivate the use of the symmetric functions of the eigenvalues when considering the analyticity issue. Let $A(\alpha_1, \alpha_2)$ be the 2×2 real matrix depending on two parameters $\alpha_1, \alpha_2 \in \mathbb{R}$.

$$A(\alpha_1, \alpha_2) = \begin{bmatrix} 1+\alpha_1 & \alpha_2 \\ \alpha_2 & 1-\alpha_1 \end{bmatrix}$$

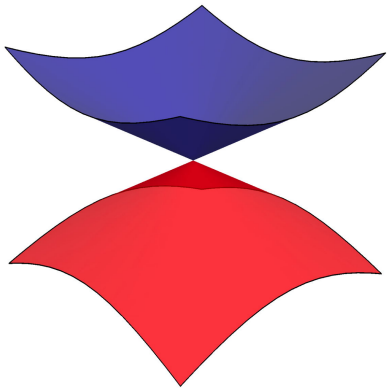


Figure 6. Eigenvalues $\lambda_1[\alpha_1, \alpha_2]$ (red) and $\lambda_2[\alpha_1, \alpha_2]$ (blue) of $A(\alpha_1, \alpha_2)$.

It is standard to compute the eigenvalues of $A(\alpha_1, \alpha_2)$. They are given by the following formulas (see also Figure 6)

$$\lambda_1[\alpha_1, \alpha_1] = 1 - \sqrt{\alpha_1^2 + \alpha_2^2}, \quad \lambda_2[\alpha_1, \alpha_1] = 1 + \sqrt{\alpha_1^2 + \alpha_2^2}$$

At the point $(\alpha_1, \alpha_2) = (0, 0)$ the eigenvalues are clearly not differentiable. If we consider instead the symmetric functions of the eigenvalues

$$\lambda_1[\alpha_1, \alpha_1] + \lambda_1[\alpha_1, \alpha_1] = 2,$$

and

$$\lambda_1[\alpha_1, \alpha_1]\lambda_2[\alpha_1, \alpha_1] = 1 - \alpha_1^2 - \alpha_2^2,$$

they turn out to be analytic.

Exploiting the abstract result of [37, 38] we are able to prove the following theorem.

Theorem 4.3 *Let F be a finite non-empty subset of $\mathbb{N}$. Then*

i) *The set $\mathcal{R}[F]$ is open in $L^\infty(\Omega)$.*

ii) *The function $\Lambda_{F,s}[\rho]$ from $\mathcal{R}[F]$ to $\mathbb{R}$ is real analytic.*

iii) *Let $\rho \in \mathcal{R}[F]$ be s.t. the eigenvalues $\lambda_j[\rho]$ assume the common value $\lambda_F[\rho]$, $\forall j \in F$. Then then the differential of $\Lambda_{F,h}$ at ρ is given by the formula*

$$d\Lambda_{F,h}[\rho][\dot{\rho}] = -C_F \sum_{j \in F} \int_\Omega u_j^2 \dot{\rho}\, dx, \quad \forall \dot{\rho} \in L^\infty(\Omega),$$

where $\{u_j\}_{j \in F}$ is an orthonormal basis of the eigenspace associated with $\lambda_F[\rho]$.

We refer to [33, 36] for the proof of Theorem 4.3. Now we turn our attention to the following extremum problems

$$\min_{\int_\Omega \rho dx = \text{const.}} / \max_{\int_\Omega \rho dx = \text{const.}} \Lambda_{F,s}[\rho].$$

In particular, all ρ's realizing the extremum are critical points under mass constraint. Let $M > 0$ be fixed and let $L_M := \left\{\rho \in \mathcal{R} : \int_\Omega \rho dx = M\right\}$. We want to find critical points for $\Lambda_{F,h}$ restricted on L_M.

We have the following theorem which holds for all the eigenvalues of problems (1.1), (1.3) and (1.5).

Theorem 4.4 *Let F be a nonempty finite subset of $\mathbb{N}$. Then for all $h = 1, ..., |F|$ the function which takes $\rho \in \mathcal{R}[F] \cap L_M$ to $\Lambda_{F,h}[\rho]$ has no critical mass densities.*

We refer to [33, 36] or the proof of Theorem 4.4. We note that Theorem 4.4 does not hold in the case of Neumann boundary conditions. Partial results are obtained for

problem (1.2), see [33] for a more detailed discussion on mass density perturbations and Neumann boundary conditions.

As for the Steklov problem (3.1), the behavior of the eigenvalues upon mass density perturbations is completely different. In fact, we have the following theorem.

Theorem 4.5 *Let B be the unit ball in $\mathbb{R}^N$. Then the constant density is a critical point for all the symmetric functions of the eigenvalues of problem (3.1) under the sole mass constraint.*

We recall that indeed the Hersch-Payne-Schiffer inequality states that for the unit disk in $\mathbb{R}^2$ the constant density is the unique maximizer under mass constraint.

As a consequence of Theorem 4.4 and Lemma 4.1, we have the following 'maximum principle' in density perturbation problems.

Theorem 4.6 *Let $C \subseteq \mathcal{R}[F]$ be a weakly* compact subset of $L^\infty(\Omega)$. Let $M > 0$ such that $C \cap L_M$ is not empty. Then for all $h = 1, ..., |F|$ the function which takes $\rho \in C \cap L_M$ to $\Lambda_{F,h}[\rho]$ has maxima and minima, and such points belong to $\partial C \cap L_M$.*

We refer to [33, 36] for the proof of Theorem 4.6. This generalizes the results of Cox-Mc.Laughlin. In fact the optimal 'bang-bang' densities found by Cox and Mc.Laughlin are boundary points (actually, extremal points) of the weakly* compact sets of densities $A \le \rho \le B$.

4.2 Mass concentration at the boundary. A new biharmonic Steklov problem

In this section we consider Neumann problems with mass densities which concentrate at the boundary of Ω. Let $M > 0$ be fixed. Let $\varepsilon > 0$ small enough and let

$$\omega_\varepsilon := \{x \in \Omega : \operatorname{dist}(x, \partial\Omega) < \varepsilon\}.$$

Let $\rho_\varepsilon : \Omega \to \mathbb{R}^+$ (see Figure 7) be defined by

$$\rho_\varepsilon := \begin{cases} \varepsilon, & \text{in } \Omega \setminus \overline{\omega}_\varepsilon, \\ \frac{M - \varepsilon|\Omega \setminus \overline{\omega}_\varepsilon|}{|\omega_\varepsilon|}, & \text{in } \omega_\varepsilon \end{cases}$$

We note that $\int_\Omega \rho_\varepsilon dx = M$ for all $\varepsilon > 0$.

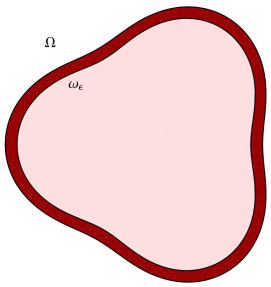


Figure 7. Mass concentration at the boundary.

We consider problem (1.2) on Ω with density $\rho = \rho_\varepsilon$, namely the problem

$$\begin{cases} -\Delta u = \lambda \rho_\varepsilon u, & \text{in } \Omega, \\ \frac{\partial u}{\partial \nu} = 0, & \text{on } \partial\Omega. \end{cases}$$

We have the following theorem, whose proof can be found in [34] (see also [1]).

Theorem 4.7 *For all $j \in \mathbb{N}$, $\lim_{\varepsilon\to 0} \lambda_j[\rho_\varepsilon] = \lambda_j$, where λ_j are the eigenvalues of*

$$\begin{cases} \Delta u = 0, & \text{in } \Omega, \\ \frac{\partial u}{\partial \nu} = \frac{M}{|\partial\Omega|}\lambda u, & \text{on } \partial\Omega. \end{cases}$$

This means that '*the Steklov eigenvalues can be seen as limiting Neumann eigenvalues in a mass concentration phenomenon*'. We refer to [22, 35] for further results on the behavior of the eigenvalues of the Neumann Laplacian upon mass concentration at the boundary.

Then we consider the following biharmonic Neumann problem with density $\rho = \rho_\varepsilon$, which is a generalization of problem (1.4)

$$\begin{cases} \Delta^2 u - \tau\Delta u = \lambda \rho_\varepsilon u, & \text{in } \Omega, \\ \frac{\partial^2 u}{\partial \nu^2} = \tau \frac{\partial u}{\partial \nu} - \operatorname{div}_{\partial\Omega}\left(D^2 u \cdot \nu\right) - \frac{\partial \Delta u}{\partial \nu} = 0, & \text{on } \partial\Omega. \end{cases}$$

Here the coefficient $\tau \geq 0$ is related to the lateral tension (if $\tau = 0$ the plate is not subject to any external tension). If $\tau = 0$ we have problem (1.4). We have the following theorem.

Theorem 4.8 *For all $j \in \mathbb{N}$, $\lim_{\varepsilon\to 0} \lambda_j[\rho_\varepsilon] = \lambda_j$, where λ_j are the eigenvalues of*

$$\begin{cases} \Delta^2 u - \tau\Delta u = 0, & \text{in } \Omega, \\ \frac{\partial^2 u}{\partial \nu^2} = 0, & \text{on } \partial\Omega, \\ \tau \frac{\partial u}{\partial \nu} - \operatorname{div}_{\partial\Omega}\left(D^2 u \cdot \nu\right) - \frac{\partial \Delta u}{\partial \nu} = \frac{M}{|\partial\Omega|}\lambda u, & \text{on } \partial\Omega. \end{cases}$$

The problem defined with this mass concentration argument is the analogue for the biharmonic operator of the classical Steklov problem (3.1). We refer to [14, 15] for the proof of Theorem 4.8. Thus, we have introduced a 'genuine' biharmonic Steklov problem, namely the problem

$$\begin{cases} \Delta^2 u - \tau\Delta u = 0, & \text{in } \Omega, \\ \frac{\partial^2 u}{\partial \nu^2} = 0, & \text{on } \partial\Omega, \\ \tau \frac{\partial u}{\partial \nu} - \operatorname{div}_{\partial\Omega}\left(D^2 u \cdot \nu\right) - \frac{\partial \Delta u}{\partial \nu} = \lambda u, & \text{on } \partial\Omega. \end{cases} \tag{0.1}$$

It is standard to prove that problem (4.9) admits an increasing sequence of non-negative eigenvalues of finite multiplicity, which can be represented by

$$0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_j \leq \cdots \nearrow +\infty.$$

In particular $\lambda_1 = 0$. If $\tau > 0$, the first positive eigenvalue is λ_2. If $\tau = 0$, $\lambda_1 = \lambda_2 = \cdots = \lambda_{N+1} = 0$ and $\lambda_{N+2} > 0$.
This problem has a different nature from other biharmonic Steklov problems already present in the literature (see e.g., [10]). In the next subsection we consider the dependence of the eigenvalues of problem (4.9) upon the domain.

4.3 Shape perturbations and optimization

When dealing with shape perturbation problem, the main issues to be considered are the continuity (or stability), the differentiability and the analyticity of the eigenvalues with respect to the domain. In the present notes we do not consider the issue of the continuity of the eigenvalues with respect to the shape. We refer to [2, 3, 12, 13, 16, 17] for more detailed discussions on spectral stability problems.

We consider the analyticity of the eigenvalues. Again, we shall exploit the abstract perturbation result of [37, 38]. We note that the set of domains has not a linear structure, so it does not make sense to define a directional derivative. We formulate the problem in an alternative way. Let Ω be a fixed domain of class C^1 and let

$$\Phi(\Omega) = \left\{\phi \in \left(C^2(\bar{\Omega})\right)^N : \phi \text{ injective and } \inf_{\Omega} |\det D\phi| > 0\right\}.$$

If Ω is of class C^1 and $\phi \in \Phi(\Omega)$, then $\phi(\Omega)$ is of class C^1 and $\phi^{(-1)} \in \Phi(\phi(\Omega))$. Then we study the Steklov problem on $\phi(\Omega)$. We denote $\lambda_j[\phi] := \lambda_j[\phi(\Omega)]$ and study the map

$$\phi \mapsto \lambda_j[\phi].$$

The space $\Phi(\Omega)$ is a linear space.

Let $F \subset \mathbb{N}$ be fixed. We introduce the following quantity

$$\mathcal{A}_\Omega[F] = \{\phi \in \Phi(\Omega) : \lambda_l[\phi] \neq \lambda_j[\phi] \quad \forall j \in F,\ \forall l \in \mathbb{N} \setminus F\}$$

Then we consider the symmetric functions of the eigenvalues for $s \in \{1, ..., |F|\}$ defined by

$$\Lambda_{F,s}[\phi] = \sum_{j_1 < \cdots < j_s \in F} \lambda_{j_1}[\phi] \cdots \lambda_{j_s}[\phi].$$

We have the following theorem.

Theorem 4.10 *Let Ω be a bounded domain in $\mathbb{R}^N$ of class C^1. Let F be a finite non-empty subset of $\mathbb{N}$. Then*

i) *The set $\mathcal{A}_\Omega[F]$ is open in $\Phi(\Omega)$.*

ii) *The function $\Lambda_{F,s}[\phi]$ from $\mathcal{A}_\Omega[F]$ to $\mathbb{R}$ is real analytic.*

iii) *Let $\tilde{\phi} \in \mathcal{A}_\Omega[F]$ be such that $\lambda_j[\tilde{\phi}] = \lambda_F[\tilde{\phi}]$, $\forall j \in F$ and such that $\tilde{\phi}(\Omega)$ is of class C^4. Let $v_1, ..., v_{|F|}$ be a orthonormal basis of the eigenspace associated with $\lambda_F[\tilde{\phi}]$. Then*

$$d|_{\phi=\tilde{\phi}}\left(\Lambda_{F,s}\right)[\psi] = -\lambda_F^{s-1}[\tilde{\phi}]\binom{|F|-1}{s-1}\sum_{j=1}^{|F|}\int_{\partial\tilde{\phi}(\Omega)}\Big(\lambda_F[\tilde{\phi}]Kv_j^2 + \lambda_F[\tilde{\phi}]\frac{\partial(v_j^2)}{\partial\nu} - \tau|\nabla v_j|^2 - |D^2 v_j|^2\Big)\psi\circ\tilde{\phi}^{(-1)}\cdot\nu d\sigma, \quad \forall\psi\in\left(C^2(\bar{\Omega})\right)^N,$$

where K denotes the mean curvature of $\partial\tilde{\phi}(\Omega)$.

We refer to [14, 15] for the proof of Theorem 4.10. Now we turn our attention to the following extremum problems

$$\min_{|\phi(\Omega)|=\text{const.}} \Big/ \max_{|\phi(\Omega)|=\text{const.}} \Lambda_{F,s}[\phi].$$

In particular, all ϕ's realizing the extremum are critical points under measure constraint. Let $\mathcal{V}_0 > 0$ and let $V(\mathcal{V}_0) = \{\phi\in\Phi(\Omega) : |\phi(\Omega)| = \mathcal{V}_0\}$. We have the following theorem, whose proof can be found in [14, 15].

Theorem 4.11 *Let Ω be a bounded domain of $\mathbb{R}^N$ of class C^1. Let $\tilde{\phi}$ be such that $\tilde{\phi}(\Omega)$ is a ball. Let $\tilde{\lambda}$ be an eigenvalue of problem (4.9) in $\tilde{\phi}(\Omega)$, and let F be the set of $j\in\mathbb{N}$ such that $\lambda_j[\tilde{\phi}] = \tilde{\lambda}$. Then $\Lambda_{F,s}$ has a critical point at $\tilde{\phi}$ on $V(|\tilde{\phi}(\Omega)|)$, for all $s = 1, ..., |F|$.*

Theorem 4.11 states that '*balls are critical domains for all simple eigenvalues and for all the symmetric functions of all multiple eigenvalues under measure constraint*'.

It is natural to ask whether we can say more on the critical nature of balls for the Steklov eigenvalues. By adapting the argument of Brock-Weinstock, we are able to prove the following.

Theorem 4.12 *Let Ω be a bounded domain in $\mathbb{R}^N$ of class C^1. Let $\lambda_2[\Omega]$ be the first positive eigenvalue of problem (4.9) on Ω with $\tau > 0$. Then*

$$\lambda_2[\Omega^*] \geq \lambda_2[\Omega], \tag{4.13}$$

where Ω^* is a ball with the same measure as Ω. The equality holds only if $\Omega = \Omega^*$ is a ball.

Theorem 4.12 states that among all bounded domains of class C^1 with fixed measure, the ball is the unique maximizer of the first non-negative eigenvalue of problem (4.9) with $\tau > 0$. In the case of $\tau = 0$ we have only partial results. We refer to [14, 15] for the proof of Theorem 4.12 and for a discussion on the case $\tau = 0$.

It is natural then to consider the issue of the stability of the inequality (4.13). This means, to answer the following questions: "*if Ω is such that $\lambda_2[\Omega] \sim \lambda_2[\Omega^*]$, then Ω has to resemble a ball? In which way this is quantified?*"
In order to answer this question, we need to introduce a "*distance among shapes*". Let

$$\mathcal{A}(\Omega) := \inf\left\{\frac{|\Omega\triangle B|}{|\Omega|} \,:\, B \text{ ball with } |B| = |\Omega|\right\}$$

be the so-called Fraenkel Asymmetry (see Figure 8). The Fraenkel Asymmetry measures the distance in the L^1 sense of a generic set from the family of balls.

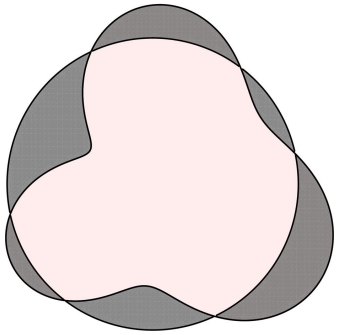


Figure 8. The set $|\Omega \triangle B|$ in dark grey.

We have the following theorem which provides an improved version of inequality (4.13). We refer to [11, 14, 15] for its proof.

Theorem 4.14 *For every domain Ω in $\mathbb{R}^N$ of class C^1 the following estimate holds:*

$$\lambda_2[\Omega] \le \lambda_2[\Omega^*]\left(1 - c_N \mathcal{A}(\Omega)^2\right), \tag{2}$$

where c_N is a suitable constant and Ω^* is a ball with the same measure as Ω.

We refer also to [6, 8] for quantitative isoperimetric inequalities for the eigenvalues of the Neumann and Dirichlet Laplacian. In particular inequality (2) implies (4.13).

Finally we consider the issue of the sharpness of the inequality (2). This means to consider the problem of the optimality of the exponent 2 for the Fraenkel asymmetry in (2). In order to prove the optimality of the exponent 2, we shall exhibit a family $\{\Omega_\varepsilon\}$ of sets approaching the unit ball B such that

$$\mathcal{A}(\Omega_\varepsilon) \simeq \frac{|\Omega_\varepsilon \triangle B|}{|\Omega_\varepsilon|} \simeq \varepsilon \quad \text{and} \quad \lambda_2[B] - \lambda_2[\Omega_\varepsilon] \simeq \varepsilon^2, \quad \varepsilon \ll 1.$$

In many cases, nearly spherical ellipsoids (see Figure 9), the most simple 'deformation' of the ball, realize the sharp exponent in various quantitative inequalities (classical isoperimetric inequality, Faber-Krahn,etc., see [6, 7, 8]).

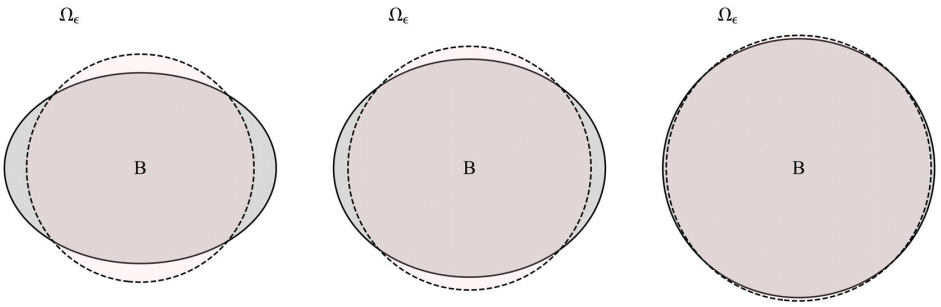


Figure 9. Nearly spherical ellipsoids.

In the case of the Steklov problem (4.9), nearly spherical ellipsoids Ω_ε are such that $\lambda_2[B] - \lambda_2[\Omega_\varepsilon] \simeq \varepsilon$, $\varepsilon \ll 1$ (we refer to [11] for more details). Then the natural question is whether are there shapes realizing the exponent 2 or is 1 (or another number $1 < \alpha < 2$) the right exponent.

We consider a family $\{\Omega_\varepsilon\}$ defined in the following way

$$\Omega_\varepsilon := \left\{ x \in \mathbb{R}^N \,:\, |x| < 1 + \varepsilon\psi(x/|x|) \right\},$$

where $\psi \in C^\infty(\partial B)$ and satisfies

1. $\int_{\partial B} \psi d\sigma = 0$;
2. $\int_{\partial B} (a \cdot x)\psi d\sigma = 0$ for all $a \in \mathbb{R}^N$;
3. $\int_{\partial B} (a \cdot x)^2 \psi d\sigma = 0$ for all $a \in \mathbb{R}^N$,

see Figure 10. This family of sets is such that $\mathcal{A}(\Omega_\varepsilon) \simeq \varepsilon$ and $\lambda_2[B] - \lambda_2[\Omega_\varepsilon] \simeq \varepsilon^2$, proving that the exponent 2 is sharp. We refer to [11] for the proof of this result and for more discussions on the sharpness of quantitative isoperimetric inequalities.

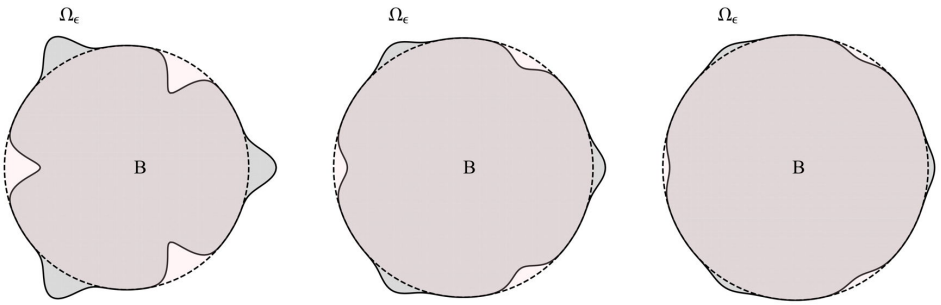


Figure 10. Domains statisfying properties i), ii) and iii).

References

[1] J. M. Arrieta, Á. Jiménez-Casas, and A. Rodríguez-Bernal, *Flux terms and Robin boundary conditions as limit of reactions and potentials concentrating at the boundary.* Rev. Mat. Iberoam., 24(1) (2008), 183–211.

[2] J. M. Arrieta and P. D. Lamberti, *Spectral stability results for higher-order operators under perturbations of the domain.* C. R. Math. Acad. Sci. Paris, 351(19-20) (2013), 725–730.

[3] J. M. Arrieta and P. D. Lamberti, *Higher order elliptic operators on variable domains. Stability results and boundary oscillations for intermediate problems.* arXiv:1502.04373 (2015).

[4] M. S. Ashbaugh and R. D. Benguria, *On Rayleigh?s conjecture for the clamped plate and its generalization to three dimensions*. Duke Math. J., 78(1) (1995), 1–17.

[5] C. Bandle, "Isoperimetric inequalities and applications". Volume 7 of Monographs and Studies in Mathematics. Pitman (Advanced Publishing Program), Boston, Mass.-London, 1980.

[6] L. Brasco, G. De Philippis, and B. Ruffini, *Spectral optimization for the Stekloff-Laplacian: the stability issue*. J. Funct. Anal., 262(11) (2012), 4675–4710.

[7] L. Brasco, G. De Philippis, and B. Velichkov, *Faberkrahn inequalities in sharp quantitative form*. Duke Math. J., 164(9) (2015), 1777–1831.

[8] L. Brasco and A. Pratelli, *Sharp stability of some spectral inequalities*. Geom. Funct. Anal., 22(1) (2012), 107–135.

[9] F. Brock, *An isoperimetric inequality for eigenvalues of the Stekloff problem*. ZAMM Z. Angew. Math. Mech., 81(1) (2001), 69–71.

[10] D. Bucur and F. Gazzola, *The first biharmonic Steklov eigenvalue: Positivity preserving and shape optimization*. Milan Journal of Mathematics, 79(1) (2011), 247–258.

[11] D. Buoso, L. M. Chasman, and L. Provenzano, *On the stability of some isoperimetric inequalities for the fundamental tones of free plates*. In preparation (2016).

[12] D. Buoso and P. D. Lamberti, *Eigenvalues of polyharmonic operators on variable domains*. ESAIM Control Optim. Calc. Var., 19(4) (2013), 1225–1235.

[13] D. Buoso and P. D. Lamberti, *Shape deformation for vibrating hinged plates*. Math. Methods Appl. Sci., 37(2) (2014), 237–244.

[14] D. Buoso and L. Provenzano, *A few shape optimization results for a biharmonic Steklov problem*. J. Differential Equations, 259(5) (2015), 1778–1818.

[15] D. Buoso and L. Provenzano, "On the eigenvalues of a biharmonic steklov problem". In Integral Methods in Science and Engineering (2015).

[16] V. Burenkov and P. D. Lamberti, *Spectral stability of higher order uniformly elliptic operators*. In Sobolev spaces in Mathematics. II, volume 9 of Int. Math. Ser. (N. Y.), Springer, New York, 2009, pp. 69–102.

[17] V. I. Burenkov and P. D. Lamberti, *Sharp spectral stability estimates via the Lebesgue measure of domains for higher order elliptic operators*. Rev. Mat. Complut., 25(2) (2012), 435–457.

[18] L. M. Chasman, *An isoperimetric inequality for fundamental tones of free plates*. Comm. Math. Phys., 303(2) (2011), 421–449.

[19] R. Courant and D. Hilbert, "Methods of mathematical physics". Vol. I. Interscience Publishers, Inc., New York, N.Y., 1953.

[20] S. J. Cox and J. R. McLaughlin, *Extremal eigenvalue problems for composite membranes. I.* Appl. Math. Optim., 22(2) (1990), 153–167.

[21] S. J. Cox and J. R. McLaughlin, *Extremal eigenvalue problems for composite membranes. II.* Appl. Math. Optim., 22(2) (1990), 169–187.

[22] M. Dalla Riva and L. Provenzano, *On vibrating thin membranes with mass concentrated near the boundary*. In preparation (2016).

[23] L. C. Evans, "Partial differential equations". Volume 19 of Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, second edition, 2010.

[24] G. Faber, *Beweis, dass unter allen homogenen Membranen von gleicher Fläche und gleicher Spannung die kreisförmige den tiefsten Grundton gibt*. Münch. Ber. 1923 (1923), 169–172.

[25] G. B. Folland, "Introduction to partial differential equations". Princeton University Press, Princeton, NJ, second edition, 1995.

[26] F. Gazzola, H.-C. Grunau, and G. Sweers, "Polyharmonic boundary value problems". Volume 1991 of Lecture Notes in Mathematics. Springer-Verlag, Berlin, 2010. Positivity preserving and nonlinear higher order elliptic equations in bounded domains.

[27] D. Gilbarg and N. S. Trudinger, "Elliptic partial differential equations of second order". Volume 224 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, second edition, 1983.

[28] A. Henrot, "Extremum problems for eigenvalues of elliptic operators". Frontiers in Mathematics. Birkhäuser Verlag, Basel, 2006.

[29] J. Hersch, L. E. Payne, and M. M. Schiffer, *Some inequalities for Stekloff eigenvalues.* Arch. Rational Mech. Anal., 57 (1975), 99–114.

[30] T. Kato, "Perturbation theory for linear operators". Springer-Verlag, Berlin-New York, second edition, 1976. Grundlehren der Mathematischen Wissenschaften, Band 132.

[31] E. Krahn, *Über eine von Rayleigh formulierte Minimaleigenschaft des Kreises.* Math. Ann., 94(1) (1925), 97–100.

[32] M. G. Krein, *On certain problems on the maximum and minimum of characteristic values and on the Lyapunov zones of stability.* Amer. Math. Soc. Transl. (2), 1 (1955), 163–187.

[33] P. Lamberti and L. Provenzano, *A maximum principle in spectral optimization problems for elliptic operators subject to mass density perturbations.* Eurasian Math. J., 4(3) (2013), 70–83.

[34] P. Lamberti and L. Provenzano, *Viewing the Steklov eigenvalues of the Laplace operator as critical Neumann eigenvalues.* In V. V. Mityushev and M. V. Ruzhansky, editors, Current Trends in Analysis and Its Applications, Trends in Mathematics, pp. 171–178. Springer International Publishing, 2015.

[35] P. Lamberti and L. Provenzano, *Neumann to Steklov eigenvalues: asymptotic and monotonicity results.* Submitted (2016).

[36] P. D. Lamberti, *Absence of critical mass densities for a vibrating membrane.* Appl. Math. Optim., 59(3) (2009), 319–327.

[37] P. D. Lamberti and M. Lanza de Cristoforis, *A real analyticity result for symmetric functions of the eigenvalues of a domain dependent Dirichlet problem for the Laplace operator.* J. Nonlinear Convex Anal., 5(1) (204), 19–42.

[38] P. D. Lamberti and M. Lanza de Cristoforis, *Critical points of the symmetric functions of the eigenvalues of the Laplace operator and overdetermined problems.* J. Math. Soc. Japan, 58(1) (2006), 231–245.

[39] N. S. Nadirashvili, *Rayleigh?s conjecture on the principal frequency of the clamped plate.* Arch. Rational Mech. Anal., 129(1) (1995), 1–10.

[40] F. Rellich, "Perturbation theory of eigenvalue problems". Assisted by J. Berkowitz. With a preface by Jacob T. Schwartz. Gordon and Breach Science Publishers, New York-London-Paris, 1969.

[41] W. Stekloff, *Sur les problèmes fondamentaux de la physique mathématique (suite et fin).* Ann. Sci. École Norm. Sup. (3), 19 (1902), 455–490.

[42] G. Szegö, *Inequalities for certain eigenvalues of a membrane of given area.* J. Rational Mech. Anal., 3 (1954), 343–356.

[43] H. F. Weinberger, *An isoperimetric inequality for the N-dimensional free membrane problem.* J. Rational Mech. Anal., 5 (1956), 633–636.

[44] R. Weinstock, *Inequalities for a classical eigenvalue problem.* J. Rational Mech. Anal., 3 (1954), 745–753.

Introduction to propagation of chaos for mean-field interacting particle systems

Luisa Andreis (*)

Abstract. The purpose of this article is to give an overview on mean-field interacting particle systems. We will focus on the notion of propagation of chaos, which aims to understand the connection between the microscopic and the macroscopic description of phenomena. Usually, an interacting particle system refers to the microscopic level and a corresponding nonlinear process describes the macroscopic one. In a great number of situations, under hypothesis on the symmetry of the system and on the type of interaction, the link between these two levels is precisely given by propagation of chaos. Since the article is intended for a general reader, we start by recalling basic definitions and results of Probability. Then we introduce the basic concepts of the theory, by means of classical examples as well as recent ones.

1 Preliminaries

The aim of this section is to define the basic objects from Probability theory that will be useful in the sequel. In particular, we aim to introduce at an informal and intuitive level diffusion processes with values in $\mathbb{R}^d$ and Markov processes with jumps, that have **càdlàg** paths, from the french expression *continue à droite, limites à gauche*, i.e. right continuous with left limits. Since we want to give a general idea on particle systems and nonlinear Markov processes, even in this preliminary part we will just mention some particular cases and we will not focus on details of stochastic calculus, for a complete introduction see, for instance, Ikeda and Watanabe [4].

1.1 Diffusion processes in $\mathbb{R}^d$

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a **probability space**, where Ω is the sample space, $\mathcal{F}$ is the σ-algebra of events and $\mathbf{P}$ is a probability measure. Informally speaking, a stochastic process is a family of random variables indexed by a time index, that can be discrete, i.e. belonging to a subset of $\mathbb{N}$, or continuous, belonging to a subset of $\mathbb{R}^+$. To describe a stochastic process, we need to define a filtration. It is an increasing sequence of σ-algebras and, in some sense, it carries the "information" available at a certain time.

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: andreis@math.unipd.it . Seminar held on March 2nd, 2016.

Definition 1.1 (Filtration) Given a σ-algebra $\mathcal{F}$, we call **filtration** a family of σ-algebras $(\mathcal{F}_t)_{t\in T}$, with $T \subset \mathbb{R}^+$ such that for all $s \le t \in T$ it holds

$$\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F}.$$

We say that $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in T}, \mathbf{P})$ is a *filtered probability space.*

Definition 1.2 A **stochastic process** with values on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is a family of $\mathbb{R}^d$-valued random variables $X = (X_t)_{t\ge 0}$ that are measurable w.r.t. $\mathcal{F}$. Let $(\mathcal{F}_t)_{t\ge 0}$ be a filtration, we say that the process X is **adapted to the filtration** $(\mathcal{F}_t)_{t\ge 0}$ if X_t is $\mathcal{F}_t$-measurable for all $t \ge 0$. In particular, we say that $X = (X_t)_{t\ge 0}$ is a **Markov process** if it is an adapted stochastic process such that, for all $0 \le s < t$

$$\mathbf{P}\left(X_t \in \cdot | \mathcal{F}_s\right) = \mathbf{P}\left(X_t \in \cdot | X_s\right).$$

Markov processes represent an important class of stochastic processes. Their feature of being "memoryless", in the sense that the future depends only on the current state and not on the whole previous history, makes them really powerful for applications and simulations. Now let us focus on a particular class of Markov processes, that have continuous paths in $\mathbb{R}^d$.

Definition 1.3 A **diffusion process** $X = (X_t)_{t\ge 0}$ with values in $\mathbb{R}^d$ is a Markov process with a.s. continuous paths.

A well-know example of diffusion process is the **Brownian motion**. Just to recall its definition, given a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in T}, \mathbf{P})$, we say that the process

$$B := (B_t)_{t\ge 0},$$

with values in $\mathbb{R}$, is an $\mathcal{F}_t$-Brownian motion if it is an adapted stochastic process with the following properties:

- $B_0 = 0$ a.s.
- $B_t - B_s \sim \mathcal{N}(0, t-s)$ for all $s < t$
- $B_{t_1} - B_{s_1}$ and $B_{t_2} - B_{s_2}$ are independent for all $s_1 < t_1 \le s_2 < t_2$
- B has a.s. continuous paths.

The paths of Brownian motions have some characteristic features, in particular they have infinite variation over every finite time interval. However it is possible to build integrals w.r.t. the Brownian motion, they are called Itô's integrals and they are defined as limits *in probability* of Riemann sums along partitions of the time interval, as the lenght of the partition's element goes to zero. Itô's integrals are the fundamental bricks of *stochastic differential equations*, given the huge literature and theory of SDE we suggest

the reading of [4] for a complete introduction of the topics. Here we aim to describe particular cases, so let us just informally mention that a diffusion process X solves a SDE in the sense that, for all $t \leq 0$, it can be written in the form

$$X_t = X_0 + \int_0^t b(X_s)ds + \int_0^t \sigma(X_s)dB_s.$$

Equivalently, we say that a diffusion process $\{X(t)\}_{t\geq 0}$ solves the following SDE

$$\begin{cases} dX_t = b(X_t)dt + \sigma(X_t)dB_t \\ X_0 \ \text{ initial condition} \end{cases}$$

where

- X_0 is a $\mathcal{F}_0$-measurable r.v.;
- b is the **drift** coefficient;
- $\int_0^t \sigma(X_t)dB_t$ is a stochastic integral w.r.t. d_1-dimensional Brownian motion $B = (B_t)_{t\geq 0}$;
- σ is a $d \times d_1$ matrix, called the **diffusion** coefficient.

Diffusion processes have a correspondence with second-order partial differential operators. Indeed, a diffusion process is uniquely determined by a second-order partial differential operator L, called the **infinitesimal generator** of the process, that acts on suitable functions f, usually in $C_b^2(\mathbb{R}^d)$, in the following way:

$$Lf(x) = \sum_{i=1}^{d} b_i(x)\frac{\partial}{\partial x_i}f(x) + \frac{1}{2}\sum_{i,j=1}^{d}(\sigma(x)\sigma(x)^T)_{i,j}\frac{\partial^2}{\partial x_i \partial x_j}f(x).$$

The probability measure $\mu_t = Law(X_t)$ satisfies, for all $f \in C_b^2(\mathbb{R}^d)$

$$\partial_t\langle \mu_t, f\rangle = \langle \mu_t, Lf\rangle.$$

That means that the time mariginals μ_t satisfy in a weak sense the PDE

$$\partial_t \mu_t = -\sum_{i=1}^{d}\frac{\partial}{\partial x_i}\left(b_i(x)\mu_t\right) + \frac{1}{2}\sum_{i,j=1}^{d}\frac{\partial^2}{\partial x_i \partial x_j}\left((\sigma(x)\sigma(x)^T)_{i,j}\mu_t\right).$$

1.2 Diffusion processes with jumps in $\mathbb{R}^d$

Let us now enlarge the class of Markov processes considered, by defining diffusion with jumps. The description of the "jumps" of a diffusion is given by means of Poisson point processes.

Definition 1.4 Let $(U, \mathcal{U}, \nu)$ be a measure space with ν a σ-finite measure, we call **Poisson random measure** a family of random variables $\{\mathcal{N}_A\}_{A\in\mathcal{U}}$, defined on some probability space $(\Omega, \mathcal{F}, \mathbf{P})$, such that

- for all $A \in \mathcal{U}$, $\mathcal{N}_A$ is a Poisson random variable with intensity $\nu(A)$;
- for all disjoint sets $A_1, A_2, \ldots, A_k \in \mathcal{U}$, the random variables $\mathcal{N}_{A_1}, \mathcal{N}_{A_2}, \ldots, \mathcal{N}_{A_k}$ are mutually independent;
- for all $\omega \in \Omega$, $\mathcal{N}_\cdot(\omega)$ is a measure on $(U,\mathcal{U})$.

A **Poisson point process** $\mathcal{N}$ with intensity $\nu(du)dt$ is a Poisson random measure on the product space $U \times \mathbb{R}^+$. It is possible to define **integrals w.r.t. Poisson point processes**, that are processes of the form

$$\left\{\int_0^t \int_U f(u,s)\mathcal{N}(du,ds)\right\}_{t\geq 0}$$

for a certain measurable function f. Loosely speaking, an integral w.r.t. a Poisson point process is an object that sums elements at a precise time, where the value of the element and the time at which it occurs are prescribed by the Poisson point process itself. To re-establish the notation of SDE, a diffusion with jumps X at each time $t \geq 0$ can be written in the following form:

$$X_t = X_0 + \int_0^t b(X_s)ds + \int_0^t \sigma(X_s)dB_s + \int_0^t \int_0^\infty \int_{[0,1]} \psi(X_s,h)\mathbb{1}_{(0,\lambda(X_s)]}(u)\mathcal{N}(dh,du,ds)$$

where

- X_0 is a $\mathcal{F}_0$-measurable r.v.;
- b is the **drift** coefficient;
- $\int_0^t \sigma(X_t,t)dB_t$ is a stochastic integral w.r.t. d_1-dimensional Brownian motion $B = (B_t)_{t\geq 0}$;
- σ is a $d_1 \times d$ matrix, called the **diffusion** coefficient;
- $\mathcal{N}$ is a Poisson random process with characteristic measure $\nu(dh) \otimes du \otimes dt$;
- ψ is the amplitude of the jumps;
- λ is the rate of the jumps.

For a heuristic description, we choose to write the Poisson integral highlighting the difference between rate of the jumps (that basically represents the rate of occurrence of a jumps at a given position) and the amplitude of the jump itself. This is a form particularly useful when dealing with applications, where often we know separately the probability of the occurrence of a jump at a given time or position and the amplitude of it. The correspondence between the process and an operator, in this case, occurs with an integro-differential

operator. Indeed, the **infinitesimal generator** of a diffusion process with jumps takes the following form:

$$Lf(x) = \sum_{i=1}^{d} b_i(x)\frac{\partial}{\partial x_i}f(x) + \frac{1}{2}\sum_{i,j=1}^{d}(\sigma(x)\sigma(x)^T)_{i,j}\frac{\partial^2}{\partial x_i \partial x_j}f(x) + \lambda(x)\int_{[0,1]} f(x+\psi(x,h)) - f(x)\nu(dh).$$

Again, the probability measure $\mu_t = Law(X_t)$ satisfies, for all $f \in C_b^2(\mathbb{R}^d)$

$$\partial_t\langle \mu_t, f\rangle = \langle \mu_t, Lf\rangle.$$

Remark 1.1 Of course, since the paths of a diffusion process are a.s. continuous, we can identify the process X with a random variable on the path space $\mathbf{C}(\mathbb{R}^+, \mathbb{R}^d)$. The same happens for a diffusion with jumps, that has paths in the Skorokhod space $\mathbf{D}(\mathbb{R}^+, \mathbb{R}^d)$ of càdlàg functions. Therefore, it seems natural to interpret a stochastic process as a **random variable in the space of its trajectories** ($\mathbf{C}(\mathbb{R}^+, \mathbb{R}^d)$, $\mathbf{D}(\mathbb{R}^+, \mathbb{R}^d)$, $\mathbf{C}([0,T], \mathbb{R}^d)$, ...). For this reason, in the sequel, when we say that a sequence of continuous stochastic processes $X^n := (X_t^n)_{t\geq 0}$ converges weakly to a continuous process X, we mean that they converge in distribution as random variables on their path space.

2 Propagation of chaos

The idea of **propagation of chaos** was introduced by Kac in 1954, in the work "Foundations of Kinetic Theory" [5], where he presented a Markovian model of gas dynamics. Indeed, the Boltzmann equation for a rarefied gas with binary collisions governs the evolution of the molecules' density in the following non-linear way:

$$\partial_t u_t(x,v) + v\cdot\nabla_x u_t(x,v) = \int_{\mathbb{R}^3\times S^2}\left(u_t(x,\bar{v})u_t(x,\bar{v}') - u_t(x,v)u_t(x,v')\right)|(v'-v)\cdot n|dv'dn$$

where

$$\bar{v} = v + ((v'-v)\cdot n)n$$
$$\bar{v}' = v' + ((v-v')\cdot n)n$$

are the modified vectors of speed after a collision. Kac's aim was to obtain the spatially homogeneous Boltzmann equation as the limit of microscopic probabilistic description of molecules. Therefore he designed an interacting particle system, where the particles randomly collide with each other. The structure of the interaction and the exchangeability of the particles in the description of the model, i.e. the particle evolution is totally symmetric, are the key ingredients of the theory of propagation of chaos. Basically this implies that the density, in the limit for the number of particles going to infinity, evolves in a deterministic way. In Kac's case, this evolution is precisely given by the spatially homogeneous Boltzmann equation.

Of course, this link between *microscopic and macroscopic* descriptions of a model has found huge interest in application and recently in the modelling of complex systems, like social sciences, economics and neuroscience. In the following section we will describe what is a system of interacting diffusion (with or without jumps) in $\mathbb{R}^d$, what the propagation of chaos for this system is and the link with the macroscopic description of the model, that gives rise to a new class of Markov processes, the so-called nonlinear Markov processes. For a complete introduction on the topic, see Sznitman's lecture notes, [7].

2.1 Microscopic and macroscopic description of a model

We mentioned that the aim of propagation of chaos is to link microscopic and macroscopic descriptions. Here we define briefly the two frameworks.

2.1.1 Interacting particle systems

We start with the microscopic view, where we consider a system of N interacting particles evolving in $\mathbb{R}^d$. This is intended as the process $\mathbf{X}^N = (X^{1,N}, \dots, X^{N,N})$ with values in $\mathbb{R}^{N\times d}$, where each component satisfies a SDE of the same type, this gives the complete symmetry of the problem and that is crucial for propagation of chaos. For easyness of notation, we will give the first description of a diffusion process without jumps. Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in T}, \mathbf{P})$ be a filtered probability space and $\left\{B^i\right\}_{i=1,\dots,N}$ a family of independent d_1-dimensional Brownian motions. Then, intuitively speaking, $\mathbf{X}^N$ is the solution of the following system of SDE:

$$\begin{cases} dX_t^{1,N} = b(X_t^{1,N}, \mathbf{X}_t^N)dt + \sigma(X_t^{1,N}, \mathbf{X}_t^N)dB_t^1 \\ \dots \\ \dots \\ dX_t^{N,N} = b(X_t^{N,N}, \mathbf{X}_t^N)dt + \sigma(X_t^{N,N}, \mathbf{X}_t^N)dB_t^N \end{cases} \tag{1}$$

Here we have two functions, the drift coefficient $b\colon \mathbb{R}^d \times \mathbb{R}^{N\times d} \to \mathbb{R}^d$ and the diffusion coefficient $\sigma\colon \mathbb{R}^d \times \mathbb{R}^{N\times d} \to \mathbb{R}^d \times \mathbb{R}^{d_1}$, that are the same for every particle and that encode the interaction in the dependence on the entire process $\mathbf{X}^N$. Notice that, the form of the two functions b and σ must satisfy some condition, in particular it must express a **mean field** type of interaction, i.e. $b(X^{i,N}, \left(X^{j,N}\right)_{j=1,\dots,N})$ (resp. $\sigma(X^{i,N}, \left(X^{j,N}\right)_{j=1,\dots,N})$) must be invariant for permutation of the indexes of $\mathbf{X}^N$.

2.1.2 Nonlinear processes

The macroscopic description of the model corresponding to (1), is given by a Markov process X with values on $\mathbb{R}^d$, that satisfies what we call a *nonlinear SDE*. Again let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in T}, \mathbf{P})$ be a filtered probability space and $\{B\}_{i=1,\dots,N}$ a d_1-dimensional Brownian motion, then we say that the process X is the solution of the following *nonlinear SDE*:

$$\begin{cases} dX_t = \tilde{b}(X_t, \mu_t)dt + \tilde{\sigma}(X_t, \mu_t)dB_t \\ \mu_t = Law(X_t). \end{cases} \tag{2}$$

Here the nonlinearity is represented by the fact that the SDE depends on the position of the process but also on the **law of the process itself**. The two functions $\tilde{b}$ and $\tilde{\sigma}$ represent, in a certain sense, the link between (1) and (2). Indeed, the first argument of these functions is the position of the process (the same of the first argument of b and σ), while the second argument still represents the "interaction", but now it is a probability measure on $\mathbb{R}^d$ (instead of the whole vector of positions in the particle system). Precisely, the functions are defined in the following way $\tilde{b}\colon \mathbb{R}^d \times \mathcal{M}^1(\mathbb{R}^d) \to \mathbb{R}^d$ and $\tilde{\sigma}\colon \mathbb{R}^d \times \mathcal{M}^1(\mathbb{R}^d) \to \mathbb{R}^d \times \mathbb{R}^{d_1}$, where $\mathcal{M}^1(\mathbb{R}^d)$ is the space of probability measures on $\mathbb{R}^d$.
Notice that a nonlinear SDE is deeply different from a classical SDE and arguments for proofs of existence and uniqueness of solutions are not trivial extensions of classical stochastic calculus theory. We will not focus on this issues, for an introduction on this topic see again Sznitman [7] or Graham [3]. Let us add a remark: the *nonlinearity* of the process is reflected on the fact that its infinitesimal generator $L(\mu)$ depends on a measure itself. Therefore the correspondent PDE is nonlinear, in the sense that, for all f sufficiently smooth, $\mu_t = Law(X_t)$ solves

$$\partial_t \langle \mu_t, f \rangle = \langle \mu_t, L(\mu_t) f \rangle.$$

2.2 Chaoticity and propagation of chaos

We want to understand how it is possible to connect a microscopic and a macroscopic model of the previous type and what type of connection links those models. We start by giving the definition of a chaotic sequence of measure, that is a crucial concept in this topic.

Definition 2.1 (p-chaotic sequence of measures) Let E be a separable metric space and p_N a sequence of symmetric probabilities on E^N. p_N is p-chaotic, with p probability on E, if for any sequence $\phi_1, \dots, \phi_k \in C_b(E)$ and for all $k \geq 1$,

$$\lim_{N\to\infty} \langle p_N, \phi_1 \otimes \cdots \otimes \phi_k \otimes 1 \otimes \cdots \otimes 1 \rangle = \prod_{i=1}^{k} \langle p, \phi_i \rangle.$$

Clearly the condition of chaoticity is different from global independence of components, even if it says that there is an **asymptotic independence**. The idea of propagation of chaos relies exactly on this definition. It says that, if the interactions are not too strong, the condition of chaoticity propagates in time despite the interactions. This means that, if a particle system starts from a condition of i.i.d. (or chaotic) initial conditions, the trajectories mantain an asymptotic independence.

Definition 2.2 (Propagation of chaos for diffusion with jumps in $\mathbb{R}^d$) Let P^N be the law of an interacting particle system $\mathbf{X}^N$, where each component is in $\mathbb{R}^d$ and the system starts from an initial condition P_0^N. Let μ be the law of a process X moving in $\mathbb{R}^d$, starting from initial condition μ_0. Notice that P^N is a probability measure on $D(\mathbb{R}^+, \mathbb{R}^d)^N$, while μ is a probability measure on $D(\mathbb{R}^+, \mathbb{R}^d)$. We say that there is **propagation of chaos** if, whenever the sequence of initial conditions $(P_0^N)_{N\in\mathbb{N}}$ is μ_0-chaotic, then the sequence $(P^N)_{N\in\mathbb{N}}$ is μ-chaotic.

2.2.1 Propagation of chaos as a Law of Large Numbers

Due to the symmetry of the laws, there is an important interpretation of propagation of chaos as a sort of law of large number. Let us briefly expose this result, for a full characterization see [7]. For a fixed $t \geq 0$, we define the **empirical measure** of a particle system $\mathbf{X}_t^N$ as

$$\mu^N_{\mathbf{X}_t}(\cdot) = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_t^{i,N}}(\cdot).$$

It is a **random variable** with values in the space $\mathcal{M}_1(\mathbb{R}^d)$. Obviously, we can consider the whole trajectory of $\mathbf{X}^N$ and define the empirical measure as a measure on the space of trajectories $D(\mathbb{R}^+, \mathbb{R}^d)$.

Theorem 2.1 *P_N is μ-chaotic **is equivalent to***

$$\mu^N_X = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_i^N} \longrightarrow \delta_\mu,$$

where the convergence is in law.

Here we write the limit as δ_μ to underline that μ is a **deterministic** limit for a sequence of random variables. Indeed, while μ^N is a sequence of random variables with values in $\mathcal{M}_1(D(\mathbb{R}^+, \mathbb{R}^d))$, μ is a deterministic element of the same space.

2.2.2 An example: McKean-Vlasov particle system

Here we informally present an example of a particle system that propagates chaos, this is the model presented by McKean, [6]. Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in T}, \mathbf{P})$ be a filtered probability space and $\{B^i\}_{i=1,\dots,N}$ a family of independent d_1-dimensional Brownian motions. Let $b(\cdot,\cdot)\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ be a bounded Lipschitz function, μ_0 a probability measure on $\mathbb{R}^d$ and $(B^i)_{i=1,\dots,N}$ a family of independent Brownian motions. We define the particle system $\mathbf{X}^N = (X^{1,N}, \dots, X^{N,N})$ that satisfies, for all $i = 1, \dots, N$,

$$\begin{cases} dX_t^{i,N} = dB_t^i + \frac{1}{N}\sum_{j=1}^N b(X_t^{i,N}, X_t^{j,N})dt \\ Law(X_0^{i,N}) = \mu_0. \end{cases}$$

The term $\frac{\mathbf{1}}{\mathbf{N}}\sum_{\mathbf{j=1}}^{\mathbf{N}} \mathbf{b}(\mathbf{X_t^{i,N}}, \mathbf{X_t^{j,N}})$ represents the **interaction**. We see that the interaction term is invariant under permutation of the indexes and it is of **mean-field type**. It can be seen also as an integral w.r.t. the **empirical measure**:

$$\frac{1}{N}\sum_{j=1}^{N} b(x, X_t^{j,N}) = \int_{\mathbb{R}^d} b(x, \mathbf{y})\mu_{X_t^N}(\mathbf{dy}).$$

It has been proven that this system propagates chaos, therefore for all $T \geq 0$, there exists a probability measure μ_T on $D([0,T])$ such that

$$\mu^N_{\mathbf{X}_{[0,T]}} \xrightarrow{d} \delta_{\mu_{[0,T]}}.$$

Heuristically, we expect that, in some way, for all $t \geq 0$,

$$\int_{\mathbb{R}^d} b(x, \mathbf{y}) \mu_{X_t^N}(\mathbf{dy}) \to \int_{\mathbb{R}^d} b(x, \mathbf{y}) \mu_t(\mathbf{dy}).$$

Indeed, the law $\mu_{[0,T]}$ is precisely the law on $D([0,T], \mathbb{R}^d)$ of the **nonlinear process**:

$$\begin{cases} dX_t = dB_t + \int_{\mathbb{R}^d} \mathbf{b}(\mathbf{X_t}, \mathbf{y}) \mu_{\mathbf{t}}(\mathbf{dy}) dt \\ Law(X_0) = \mu_0 \\ \mu_t = Law(X_t) \text{ for } t > 0 \end{cases}$$

Notice that the marginal distribution of the nonlinear process μ_t satisfy in a weak sense the **nonlinear PDE**

$$\partial_t \mu_t = \frac{1}{2} \Delta \mu_t - \operatorname{div} \left(\int_{\mathbb{R}^d} b(\cdot, y) \mu_t(dy) \mu_t(\cdot) \right).$$

2.2.3 Strategy of proof for propagation of chaos

In literature we found mainly two approaches for the proof of propagation of chaos. Let us briefly summarize them.

1) The martingale approach consists in:

   - proving that the sequence of empirical mesures $\{\mu_N\}_{N \in \mathbb{N}}$ is **tight**, i.e. it admits a convergent subsequence;
   - identifying **every** limit point of $\{\mu_N\}_{N \in \mathbb{N}}$ with the law of the solution of the nonlinear SDE;
   - proving that the solution of the nonlinear SDE is unique.

2) The coupling approach consists in:

   - considering the processes $\mathbf{X}^N = (X^{1,N}, \dots, X^{N,N})$ the usual particle system and $x_N = (x_1, \dots, x_N)$ N independent copies of the solution of the **nonlinear process** both starting from the same vector of i.i.d. initial condition;
   - coupling the two vectors by building a **joint measure** whose marginals are the two laws;
   - by means of the coupling getting a quantitative bound of the form

     $$\mathbf{E}\left[\rho_T\left(\mu_N, \mu\right)\right] \leq \frac{C_T}{N^\gamma}$$

     for a certain exponent $\gamma > 0$ and a distance ρ_T between measures on the space of trajectories, that implies weak convergence of the sequence μ_N to the limit deterministic law μ. See Fournier and Guillin, [2], for a nice overview on rate of convergence w.r.t. moments conditions on the laws, dimension and choice of the distance ρ_T.

3 Recent applications of interacting particle systems

At its birth the mean-field approach was seen as an approximation of physical situations, where the existing local interactions were too difficult to understand and there was the need of an extreme simplification to interpret macroscopical behaviors. Recently, the research on complex systems has focused its interest on mean-field models as they reveal most of the main characteristics of social, economical or biological systems. Therefore, the study of mean field models has still an important role in applications, as we will explain in the following by briefly present particle systems for neuroscience.

3.1 Mean field models for neuroscience

Particle systems in neuroscience are used to model the brain's behavior, indeed the neuronal network has characteristic features that fit very well with the mean-field approach. First of all, the brain is a large size network of neurons, with high number of connections among them, therefore considering the interactions of mean-field type and the limit for size of the system going to infinity seems a reasonable approximation. Moreover, from the easier observation of microscopical behavior would be interesting to understand the *macroscopical one*, that sometimes seems to show very different peculiarities. For instance neurons' membrane potential has been modelled with particle systems with the aim of understanding the behavior (in particular the intrisic periodic phenomena even without external periodic inputs) of the overall system. The membrane potential of a neuron is a quantity that increases or decreases in time, depending on the influence of the other neurons, and that sometimes has a so-called *spike*. A *spike* is a sudden event that forces the membrane potential of the neuron to a resting potential. While the increase and the decrease due to the interactions can be considered as continuous components, the spikes represent jumps in the trajectory of the potential. Here we summarize two different models for the neuronal network. The second one is the main motivation of our study of particle systems with simultaneous jumps.

1) **Diffusion with jumps when reaching a threshold: leaky integrate and fire model**
   In this model there is interaction between neurons, the randomness comes from external noise and the spike of each neuron occurs when its membrane potential reaches a threshold.

2) **Random and simultaneous jumps: Poisson model**
   In this model the occurrence of a spike for a neuron encodes the randomness of the system, since it happens according a certain **rate** that increases as the membrane potential increases. The interaction is represented by the fact that, as soon as a neuron spikes, it causes a little increase in the membrane potential of the others.

3.2 Interacting particle systems with simultaneous jumps

In the following we present a quite general particle system with simultaneous jumps, inspired by Poisson models in neuroscience [1]. We define a system of N interacting

particles on the filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in T}, \mathbf{P})$, rich enough to carry the family $\{B^i, \mathcal{N}^i\}_{i=1,\dots,N}$ of independent Brownian motions and Poisson stationary processes. Each "particle" is associated to a **diffusion with jumps** $X_i^N(t)$ with values on $\mathbb{R}^d$. Therefore the system is represented by the process $X^N = \left(X^N(t)\right)_{t\in[0,T]}$ where

$$X^N(t) = \left(X_1^N(t), \dots, X_N^N(t)\right) \qquad \in \qquad \mathbb{R}^d \times \cdots \times \mathbb{R}^d$$

for all $t \in [0,T]$. The process solves the system of SDE: for all $i = 1, \dots, N$,

$$\begin{aligned} dX_i^N(t) = {} & F(X_i^N(t), \mu_X^N(t))dt + \sigma(X_i^N(t), \mu_X^N(t))dB_t^i \\ & \text{(drift and diffusion coefficients)} \\ & + \int_{[0,\infty)} \psi(X_i^N(t), \mu_X^N(t)) \mathbb{1}_{[0,\lambda(X_i^N(t),\mu_X^N(t)))}(u) \mathcal{N}^i(dt, du), \\ & \text{(main jump term)} \\ & + \frac{1}{N} \sum_{j\neq i} \int_{[0,\infty)} \Theta(X_j^N(t), X_i^N(t), \mu_X^N(t)) \mathbb{1}_{[0,\lambda(X_j^N(t),\mu_X^N(t)))}(u) \mathcal{N}^j(dt, du). \\ & \text{(collateral jumps terms)} \end{aligned}$$

Notice that each coefficient depends on the position of the particle and on all the other particles through the *empirical measure*, in a way that the coefficients are defined on $\mathbb{R}^d \times \mathcal{M}^1(\mathbb{R}^d)$, already in the particle system. We said that the peculiarity of this systems is the presence of simultaneous jumps, indeed each particle's law is influenced by N independent Poisson random processes and the rates of these N jumps depend only on one of the particles at each time. For instance, the i-th particle appears in the jump rate $(\lambda(X_i^N, \mu_X^N))$ of the i-th Poisson random integral. At the same time it forces also the i-th particle to jump of an amplitude given by ψ and all the other $N-1$ particles to jump of an amplitude given by $\frac{\Theta}{N}$. Existence and uniqueness of such a process is ensured, by classical results, for bounded-Lipschitz conditions on all the coefficients and square-integrable initial conditions.

Our aim is to prove that such an interacting system propagates chaos. We expect that because of the rescaling of the factor $\frac{1}{N}$ in the simultaneous collateral jumps. In particular we expect that those "collateral jumps" will end up in an additional nonlinear drift term when N goes to infinity. Therefore, the correspondent limiting *nonlinear process* should be the law of the following Markov process. Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in T}, \mathbf{P})$ be a filtered probability space and B a Brownian motion on it, the process $X = (X(t))_{t\in[0,T]}$ with values on $\mathbb{R}^d$ is the solution of the *nonlinear SDE*:

$$dX(t) = F(X(t),\mu_t)dt + \sigma(X(t),\mu_t)dB_t$$
(drift and diffusion coefficients)
$$+ \int_{[0,\infty)} \psi(X(s^-),\mu_s)\mathbb{1}_{[0,\lambda(X(s^-),\mu_s)]}(u)\mathcal{N}(du,dh,dt);$$
(main jump term)
$$+ < \mu_t, \lambda(\cdot,\mu_t)\Theta(\cdot,X(t^-),\mu_t) > dt$$
(new drift term coming from the collateral jumps)

where we have, for all $t \geq 0$, $\mu_t = Law(X(t))$. It is not the aim of these notes to discuss existence and uniqueness of such a nonlinear process, we just underline that in the case of bounded-Lipschitz coefficients, everything is well-posed also in this limit. Our main interest is to answer some questions about this model.

- Do simultaneous jumps interfere with propagation of chaos?
- Do classical methods apply to this model?
- Can we handle a procedure of coupling to get quantitative rate of convergence?

To answer these questions we build an approach by means of an **intermediate process without simultaneous jumps** to handle the proof of propagation of chaos by coupling and, consequentely, get rates of convergence. The **intermediate process** Y^N is again a system of N interacting particles, each of them associated to a **diffusion with jumps** $Y_i^N(t)$ with values on $\mathbb{R}^d$. Given again a filtered probability space $(\Omega,\mathcal{F},(\mathcal{F}_t)_{t\in T},\mathbf{P})$ and a family $\{B^i,\mathcal{N}^i\}_{i=1,\dots,N}$ of independent Brownian motions and Poisson stationary processes, the process solves the system of SDE: for all $i = 1,\dots,N$,

$$dY_i^N(t) = F(Y_i^N(t),\mu_Y^N(t))dt + \sigma(Y_i^N(t),\mu_i^N(t))dB_t^i$$
(drift and diffusion coefficients as for X^N)
$$+ \int_{[0,\infty)} \psi(Y_i^N(s^-),\mu_Y^N(s^-))\mathbb{1}_{[0,\lambda(Y_i^N(s^-),\mu_i^N(s^-)))}(u)\mathcal{N}^i(du,dh,dt),$$
(main jump term)
$$+\frac{\mathbf{1}}{\mathbf{N}}\sum_{\mathbf{j=1}}^{\mathbf{N}}\lambda(\mathbf{Y_j^N(t)},\mu_Y^N(t))\Theta(\mathbf{Y_j^N(t)},Y_i^N(t^-),\mu_Y^N(t))dt.$$
(collateral jumps terms transformed in an additional drift term)

By means of the introduction of the **intermediate process** we prove that the empirical laws of the two systems get closer as N goes to infinity, i.e.

$$\mathbf{E}\left[\rho_T\left(\mu_X^N,\mu_Y^N\right)\right] \leq \frac{C_T}{\sqrt{N}},$$

as $N \to \infty$, where ρ_T is the W_1 Vasserstein distance on $\mathcal{M}_1(D([0,T],\mathbb{R}^d))$. Moreover, **propagation of chaos** holds for the intermediate system Y^N, with μ as the law of the limit process.

References

[1] L. Andreis, P. Dai Pra, M. Fischer, *McKean-Vlasov limit for interacting systems with simultaneous jumps*. In preparation (2016).

[2] N. Fournier and A. Guillin, *On the rate of convergence in Wasserstein distance of the empirical measure*. Probability Theory and Related Fields, 162(3-4) (2015), 707–738.

[3] C. Graham, N. Cognome, *McKean-Vlasov Itô-Skorohod equations, and nonlinear diffusions with discrete jump sets*. Stochastic processes and their applications, 40(1) (1992), 69–82.

[4] N. Ikeda and S. Watanabe, "Stochastic differential equations and diffusion processes". Volume 24, Elsevier, 2014.

[5] M. Kac, *Foundations of kinetic theory*. In Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, volume 1955, pp. 171–197, 1954.

[6] H. P. McKean, *A class of Markov processes associated with nonlinear parabolic equations*. Proceedings of the National Academy of Sciences, 56(6) (1966), 1907–1911.

[7] A.-S. Sznitman, *Topics in propagation of chaos*. In Ecole d'Eté de Probabilités de Saint-Flour XIX 1989, pp. 165–251. Springer, 1991.

Cosheaves, an introduction

Pietro Polesello [(*)]

Abstract. It is well known that locally defined distributions glue together, that is, they define a sheaf. In fact, this follows immediately from the fact that test functions (i.e. smooth functions with compact support) form a cosheaf, which is the dual notion of a sheaf.
By definition, cosheaves on a space X and with values in category $\mathcal{C}$ are dual to sheaves on X with values in the opposite category $\mathcal{C}^{\mathrm{op}}$. For this reason, cosheaves did not attract much attention, being considered as part of sheaf theory. However, passing from $\mathcal{C}$ to $\mathcal{C}^{\mathrm{op}}$, may cause difficulties, as in general $\mathcal{C}$ and $\mathcal{C}^{\mathrm{op}}$ do not share the same good properties needed for sheaf theory (*e.g.* colimits are not exact in $\mathrm{Ab}^{\mathrm{op}}$, the opposite of category of abelian groups). Moreover, dealing with cosheaves may be more convenient, as they appear naturally in analysis (as the compactly supported sections of c-soft sheaves, such as smooth functions or distributions), in algebraic analysis (*e.g.* as the subanalytic cosheaf of Schwartz functions), in topology (in relation with Fox's theory of topological branched coverings), and in tops theory. Moreover, as sheaves are the natural coefficient spaces for cohomology theories, cosheaves play the same role for homology theories, such as Čech homology, and they are (hidden) ingredients of Poincaré duality (recently, ∞-cosheaves infiltrated Poincaré-Verdier duality in the context of Lurie's *higher topos theory*).
In this seminar, I will give a brief introduction to cosheaves, giving examples and explaining the relation with sheaves, with Lawvere's distributions and with Fox's branched coverings.

(This note is handwritten, contents follow on next page.)

References

[Br] G. E. Bredon, "Sheaf Theory". Second edition. Graduate Texts in Math. 170, Springer, 1997.

[Fox] R. H. Fox, *Covering spaces with singularities.* A symposium in honor of S. Lefschetz, Princeton University Press, Princeton, N.J. (1957), 243–257.

[KS1] M. Kashiwara and P. Schapira, "Sheaves on manifolds". Grundlehren der Mathematischen Wissenschaften 292, Springer-Verlag, Berlin, 1990.

[KS2] M. Kashiwara and P. Schapira, "Moderate and formal cohomology associated with constructible sheaves". Mém. Soc. Math. France (N.S.) 64, 1996.

[Lu] J. Lurie, "Higher topos theory". Annals of Mathematics Studies 170, Princeton University Press, Princeton, NJ, 2009.

[Sch] J.-P. Schneiders, "Quasi-abelian categories and sheaves". Mém. Soc. Math. France (N.S.) 76, 1999.

[W] Woolf, J., *The fundamental category of a stratified space.* J. Homotopy Relat. Struct. 4 (2009), no. 1, 359–387.

[(*)]Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: pietro@math.unipd.it . Seminar held on March 16th, 2016.

§ 0. INTRO

M : C^∞-manifold

$V \underset{\text{open}}{\subseteq} M \Rightarrow \mathcal{D}b(V) := \text{distributions on } V \in LCS$: Loc. Convex Vect. Spaces (topology defined by a family of semi-norms) (quasi-abelian cat. - see [Sch] - complete & cocomplete)

It is well-known that the assignment

$$\begin{array}{ccc} M \supseteq & V & \longmapsto \mathcal{D}_b(V) \\ \text{open} & U & \downarrow \rho : \text{restriction} \\ & U & \longmapsto \mathcal{D}_b(U) \end{array}$$

defines a sheaf (of LCS) on M.

This means the following: for any $\bigcup_{i\in I} V_i =: V$ open cover

"SEPARATED PRE-SHEAF CONDITION" i) $T \in \mathcal{D}b(V)$ s.t. $T|_{V_i} = 0 \ \forall i \in I \Rightarrow T = 0$

"GLUEING CONDITION" ii) $T_i \in \mathcal{D}b(V_i) \ \forall i \in I$ s.t. $T_i|_{V_{ij}} = T_j|_{V_{ij}}$ $\Rightarrow \exists\, T \in \mathcal{D}b(V)$ s.t. $T|_{V_i} = T_i \ \forall i \in I$

Equivalently: the sequence given by restrictions (of LCS)

$$0 \to \mathcal{D}b(V) \longrightarrow \prod_{i\in I} \mathcal{D}b(V_i) \underset{\rho_2}{\overset{\rho_1}{\rightrightarrows}} \prod_{i,j\in I} \mathcal{D}b(V_{ij}) \quad \text{is exact}$$

$$T \longmapsto T|_{V_i} \qquad T_i \longmapsto T_i|_{V_i},\ T_i \longmapsto T_i|_{V_j}$$

(i.e. $\mathcal{D}b(V) \xrightarrow{\cong} \ker(\rho_1, \rho_2)$)

$\Leftrightarrow$ (*) : $\mathcal{D}b(V) \xrightarrow{\cong} \varprojlim_{i\in I} \mathcal{D}b(V_i)$ (if the open cover $\bigcup_{i\in I} V_i$ is stable by finite $\cap$)

limit (of LCS) : $\ker(\rho_1, \rho_2)$ (+ induced topology of the product Top.)

Recall that, by definition:

$\mathcal{D}b(V) := \mathcal{Y}_c^\infty(V)'$ = topological dual of the LCS $\mathcal{Y}_c^\infty(V) := C^\infty$-densities on V with compact support ("$\mathbb{C}$-valued volume forms" (if V is oriented))

It is less known that

$$\begin{array}{ccc} M \supseteq_{\text{open}} V & \longmapsto & \mathcal{V}_c^\infty(V) \\ \cup\!| & & \uparrow \varepsilon : \text{extension by zero} \\ U & \longmapsto & \mathcal{V}_c^\infty(U) \end{array}$$

defines a <u>cosheaf</u> of LCS on M.

this means that the sequence given by extensions of LCS

$$\bigoplus_{i,j\in I} \mathcal{V}_c^\infty(V_{ij}) \underset{\varepsilon_2}{\overset{\varepsilon_1}{\rightrightarrows}} \bigoplus_{i\in I} \mathcal{V}_c^\infty(V_i) \longrightarrow \mathcal{V}_c^\infty(V) \to 0 \quad \text{is } \underline{\text{exact}}$$

$$\bigoplus_{i,j} \omega_{ij} \longmapsto \bigoplus_i \sum_j \varepsilon_{V_i}(\omega_{ij}) \quad \text{and} \quad \bigoplus_i \sum_j \varepsilon_{V_i}(\omega_{ji}) \qquad \bigoplus_i \omega_i \longmapsto \sum_i \varepsilon_V(\omega_i)$$

$$\left(\text{i.e. } \operatorname{coker}(\varepsilon_1,\varepsilon_2) \xrightarrow{\cong} \mathcal{V}_c^\infty(V)\right)$$

for any open cover $\bigcup_{i\in I} V_i =: V$.

$$\Longleftrightarrow \text{(**)} : \varinjlim_{i\in I} \mathcal{V}_c^\infty(V_i) \xrightarrow{\cong} \mathcal{V}_c^\infty(V) \qquad \left(\text{if the open cover } \bigcup_{i\in I} V_i \text{ is stable by finite } \bigcap\right)$$

$\varinjlim$ colimit of LCS : $\bigoplus_i \mathcal{V}_c^\infty(V_i) \Big/ \overline{\operatorname{Im}(\varepsilon_1-\varepsilon_2)}$ + quotient topology

$\neq$ colim in Top !

+ "diagonal" topology

Roughly, this means that for any open cover $\bigcup_{i\in I} V_i =: V$:

$$\begin{cases} i) \ \omega \in \mathcal{V}_c^\infty(V) \Rightarrow \omega = \sum_i \varepsilon_{V_i}(\omega_i) \text{ for } \omega_i \in \mathcal{V}_c^\infty(V_i) \ \forall i \in I \\ ii) \ \sum_i \varepsilon_{V_i}(\omega_i) = 0 \Rightarrow \omega_i = \sum_j \left[\varepsilon_{V_i}(\omega_{ij}) - \varepsilon_{V_i}(\omega_{ji})\right] \text{ for } \omega_{ij} \in \mathcal{V}_c^\infty(V_{ij}) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall i,j \in I \end{cases}$$

Note that, for $V_1 \cup V_2 = V$, i) and ii) follows from the Mayer-Vietoris sequence of LCS :

$$0 \to \mathcal{V}_c^\infty(V_1 \cap V_2) \xrightarrow{\varepsilon_1-\varepsilon_2} \mathcal{V}_c^\infty(V_1) \oplus \mathcal{V}_c^\infty(V_2) \longrightarrow \mathcal{V}_c^\infty(V_1 \cup V_2) \to 0$$

By taking the topological dual : (**) $\Rightarrow$ (*).

In fact: $(\cdot)^! := \mathrm{Hom}_{\mathrm{Top}}(\cdot, \mathbb{C})$ commutes with colimits, hence:

$$\Big(\varinjlim_{i\in I} \mathcal{V}_c^\infty(V_i)\Big)' \xleftarrow{\cong} \mathcal{V}_c^\infty(V)' = \mathcal{D}b(V)$$

$$\Big(\varinjlim_{i\in I} \mathcal{V}_c^\infty(V_i)\Big)' \xrightarrow{\cong} \varprojlim_{i\in I} \mathcal{V}_c^\infty(V_i)' = \varprojlim_{i\in I} \mathcal{D}b(V_i)$$

Moreover, the fact that (**) holds follows from the fact that $\mathcal{V}^\infty$ is a sheaf on M. Hence:

$$\mathcal{V}^\infty \text{ sheaf} \Rightarrow \mathcal{V}_c^\infty \text{ cosheaf} \Rightarrow \mathcal{D}b = \mathcal{V}_c^{\infty\prime} \text{ sheaf}.$$

We will see that this holds in a more general situation.
Let's start with the basic definitions and properties of cosheaves.

§ 1. Cosheaves

C: cocomplete category (i.e. admitting all small colimits)
such as Set, Ab, Mod(R), Top, LCS, etc.

X: top. space (more generally, a site)

$\mathcal{O}p(X) :=$ category of open subsets of X and inclusions

Def: i) a pre-cosheaf with values in C is a functor
$\mathcal{E}: \mathcal{O}p(X) \longrightarrow C$. For $U \subseteq V$, $\mathcal{E}(U) \xrightarrow{e_{UV}} \mathcal{E}(V)$: extension morphism.
We get a category $\mathrm{PCosh}(C_X) := \mathrm{Fct}(\mathcal{O}p(X), C)$.

ii) $\mathcal{E}$ is a cosheaf if for any $\bigcup_{i\in I} V_i =: V$ open cover stable by finite $\cap$, the natural morphism in C

$$(**) \quad \varinjlim_{i\in I} \mathcal{E}(V_i) \longrightarrow \mathcal{E}(V) \quad \text{is an isom.}$$

If (**) is epi ($r \xrightarrow{f} q$ epi if $\forall\ p \underset{g_2}{\overset{g_1}{\rightrightarrows}} R$ s.t. $g_1 \circ f = g_2 \circ f \Rightarrow g_1 = g_2$)
then $\mathcal{E}$ is called a separated pre-cosheaf.

We get a full subcategory $\mathrm{Cosh}(C_X) \subseteq \mathrm{PCosh}(C_X)$.

Remarks: 1) for any cosheaf $\mathcal{E}$, $\mathcal{E}(\emptyset)$ is initial in C (since $\emptyset$ is initial in $\mathcal{O}(X)$)

2) since C has coproducts $\sqcup$, $\mathcal{E}$ is a cosheaf $\Leftrightarrow$ the natural sequence given by extension morph.

$$\coprod_{i,j\in I} \mathcal{E}(V_{ij}) \underset{\varepsilon_2}{\overset{\varepsilon_1}{\rightrightarrows}} \coprod_{i\in I} \mathcal{E}(V_i) \longrightarrow \mathcal{E}(V)$$ is exact in C (i.e. $\mathrm{coker}(\varepsilon_1,\varepsilon_2) \xrightarrow{\sim} \mathcal{E}(V)$)

$\| 2$

$\varinjlim_{i\in I} \mathcal{E}(V_i)$

3) By standard "devissage of colimits" and by induction:
$\mathcal{E}$ is a cosheaf $\Leftrightarrow$

$$\begin{cases} \text{a) } (**) \text{ hom. for open covers stable by finite } \cup; \\ \text{b) } \begin{array}{ccc} \mathcal{E}(V_1\cap V_2) & \longrightarrow & \mathcal{E}(V_1) \\ \downarrow & & \downarrow \\ \mathcal{E}(V_2) & \longrightarrow & \mathcal{E}(V_1\cup V_2) \end{array} \text{ is cocartesian (i.e. a push-out) for any } V_1, V_2 \in \mathcal{O}(X). \end{cases}$$

3) b) is a "Mayer-Vietoris condition". Indeed, if $C = Ab$, 3) b) $\Leftrightarrow$

$$\mathcal{E}(V_{12}) \longrightarrow \mathcal{E}(V_1)\oplus\mathcal{E}(V_2) \xrightarrow{\varepsilon_{V_1} - \varepsilon_{V_2}} \mathcal{E}(V_1\cup V_2) \longrightarrow 0 \text{ is exact.}$$

4) If $C = Ab$, $\mathrm{Mod}(R)$, LCS, etc. and $\mathrm{for}: C \to \mathrm{Set}$ is the forgetful functor, then in general "for" does <u>not</u> commute with colimits. Hence, contrarily to sheaves, a cosheaf with values in C in general does not define an "underlying cosheaf of sets".

Dually, if "for" admits a left adjoint (e.g. free construction, discrete topology, etc.), then this has "for" as right adjoint, hence it preserves colimits $\Rightarrow$ it preserves cosheaves.

§ 2. Examples

1) $V \longmapsto \emptyset :=$ initial object in $\mathcal{C}$ is a cosheaf (initial in $\mathrm{Cosh}(C_X)$)

2) $V \longmapsto pt :=$ terminal object in $\mathcal{C}$ (supposed it exists) is a pre-cosheaf (terminal in $P\mathrm{Cosh}(C_X)$), and it is a cosheaf $\Leftrightarrow pt = \emptyset$, i.e. it is a zero object of $\mathcal{C}$

Similarly, the constant pre-cosheaf $V \longmapsto P \in \mathcal{C}$ is a cosheaf $\Leftrightarrow P = \emptyset$.

3) $V \longmapsto V \in \mathrm{Top}$ is the trivial cosheaf t_X of top. spaces:

$$\varinjlim_{i \in I} V_i \cong \mathrm{coker}\left(\coprod_{i,j \in I} V_{ij} \rightrightarrows \coprod_{i \in I} V_i \right) \xrightarrow{\cong} \bigcup_{i \in I} V_i = V$$

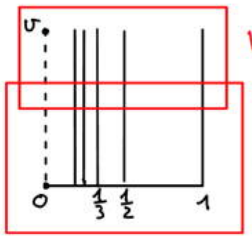
4) $V \longmapsto V \in \mathrm{Set}$ is the forgetful cosheaf (of sets) for_X. Indeed $\mathrm{for} : \mathrm{Top} \to \mathrm{Set}$ has a right adjoint (the trivial, or indiscrete, topology functor), hence it preserves colimits $\Rightarrow \mathrm{for}_X = \mathrm{for}|_{\mathcal{O}(X)}$: cosheaf

5) given $x_0 \in X$, $V \longmapsto \delta_{x_0}(V) := \begin{cases} pt & \text{if } x_0 \in V \\ \emptyset & \text{otherwise} \end{cases}$ is a cosheaf of sets

(δ_{x_0}: "Dirac cosheaf")

6) $V \longmapsto \#V := \{\text{connected components of } V\} \in \mathrm{Set}$ is a separated precosheaf of sets, but in general not a cosheaf.

(Ex)



X "deleted comb space". Then 3) 6)

$$\#V_{12} \underset{\{1\}}{\overset{\mathrm{id}}{\rightrightarrows}} \#V_1 \sqcup \#V_2 \longrightarrow \#X \qquad \text{is not exact}$$

$$\mathbb{N}_{\geq 1} \qquad \{0\} \sqcup \mathbb{N}_{\geq 1} \qquad \{1\} \qquad \{1\}$$

$$\mathrm{coker}(\ ) \cong \{0,1\}$$

However, if X is totally disconnected (i.e. the only connected subspace are the singletons), e.g. $X = \mathbb{Q}$ or discrete, then

$\#_x = for_x$ is a cosheaf. Moreover:

Prop: if X is loc. connected $\Rightarrow$ $\#_x$ is a cosheaf.

proof: the discrete topology functor

$$Set \longrightarrow LCT := \{\text{Loc. Connected Top. Spaces}\}$$

is right adjoint to the connected components functor

$$\# : LCT \longrightarrow Set$$

$\Rightarrow$ $\#$ preserves colimits $\Rightarrow$ $\#_x = \#|_{\mathcal{O}(X)}$ is a cosheaf. $\square$

7) $V \longmapsto \pi_0(V) := \{\text{path-component of } V\} \in Set$ is a cosheaf (0-th homotopy cosheaf) (exercise)

8) $V \longmapsto \underbrace{H_0(V)}_{\in Ab} := Z_0(V)/B_0(V)$: singular homology of V is a cosheaf (0-th homology cosheaf) as $H_0(V) \cong \mathbb{Z}[\pi_0(V)]$ and $\mathbb{Z}[\cdot]$ preserves colimits.

9) $X = M$ C^∞-manifold of dim n, F: sheaf on M of abelian groups $V \longmapsto H_c^n(V; F)$ is a cosheaf.

10) We will prove that, for any c-soft sheaf F of abelian groups on a loc. compact space X

$V \longmapsto \Gamma_c(V; F) := \{\text{sections in } \Gamma(V;F) \text{ with compact support}\}$

is a cosheaf. Replacing Ab by LCS, this generalizes the example of densities at the beginning.

11) One may define also differential (pre-) cosheaves, such as

$V \longmapsto S_p(V, A) := \{\text{singular } p\text{-chains with coefficients in } A \in Ab\}$

If $p = 0$, then $S_0(\cdot, A) \cong \mathbb{Z}[t_x] \otimes_{\mathbb{Z}} A$ is a cosheaf; if $p > 0$ this is in general not true (one has to use the baricentric subdivision!)

§ 3. SHEAVES AND COSHEAVES

First, note that: $\mathcal{C}$ is a cosheaf with values in C

$\Leftrightarrow$ $\mathcal{C}^{op}: \mathcal{O}(X)^{op} \to C^{op}$ is a sheaf

$\left(\text{since } \left(\varinjlim_{i\in I} \mathcal{C}(V_i)\right)^{op} \cong \varprojlim_{i\in I} \mathcal{C}^{op}(V_i)\right)$

hence $CoSh(C_X)^{op} \cong \{C^{op}\text{-valued sheaves on } X\}$. We get:

Prop: $V \longmapsto CoSh(C_V)$ is a stack. In particular: $\mathcal{C}_1, \mathcal{C}_2$ cosheaves

$\Rightarrow V \longmapsto \mathrm{Hom}_{Coh(C_X)}(\mathcal{C}_1|_V, \mathcal{C}_2|_V)$ is a sheaf of set: $\mathrm{Hom}(\mathcal{C}_1, \mathcal{C}_2)$

$\{\text{cosheaves of sets on } X\}^{op} \cong \left\{\begin{array}{l}\text{sheaves on } X \text{ with values in} \\ \underline{\text{complete atomic Boolean algebras}}\end{array}\right\}$

$\sim \{\text{power sets}\} \cong Set^{op}$

Lindenbaum–Tarski duality

However, in general C and C^{op} are not both well-suited for sheaf theory (such as $C = Ab$: $Ab^{op} \cong$ compact Hausdorff abelian groups by Pontryagin duality $\Rightarrow$ in such category colimits are not exact $\Rightarrow$ the "associated sheaf functor" does not exist) and it may be convenient to work with cosheaves.

———— ∘ ————

As for the example of densities, by taking duals of cosheaves, one gets sheaves. Precisely,

Prop: $\mathcal{C}$: pre-cosheaf with values in C. Then:

$\mathcal{C}$ cosheaf $\Leftrightarrow \underbrace{\mathrm{Hom}_C(\mathcal{C}(\cdot), A)}_{\mathrm{Hom}(\mathcal{C}, A)}$ is a sheaf of sets $\forall\, A \in C$

(resp. separated pre-cosheaf) (resp. separated sheaf)

proof: by applying $\mathrm{Hom}_C(\cdot, A)$ to $\varinjlim_{i\in I} \mathcal{E}(V_i) \xrightarrow{\alpha} \mathcal{E}(V)$

we get $\mathrm{Hom}_C(\varinjlim \mathcal{E}(V_i), A) \xleftarrow{\alpha_A^t} \mathrm{Hom}_C(\mathcal{E}(V), A)$

$\downarrow \beta_A$

then: $\cong \searrow \varprojlim \mathrm{Hom}_C(\mathcal{E}(V_i), A)$

α epi. $\iff \alpha_A^t$ injective $\forall A \in C \iff \beta_A$ injective $\forall A \in C$

(i.e. $\mathrm{Hom}_C(\mathcal{E}(\cdot), A)$ separated pre-sheaf)

α iso $\overset{\text{YONEDA}}{\iff} \alpha_A^t$ iso $\forall A \in C \iff \beta_A$ iso $\forall A \in C$.

(i.e. $\mathrm{Hom}_C(\mathcal{E}(\cdot), A)$ sheaf) □

Some classical sheaves arise as dual of cosheaves.

(Ex) 1) Given a set S, $\mathrm{Hom}(\mathcal{L}or_x, S)$ is the sheaf of maps $X \to S$ and $\mathrm{Hom}(\delta_{x_0}, S)$ is the <u>skyscraper</u> sheaf with stalk S at x_0.

2) Given an abelian group A, $\mathrm{Hom}(H_{0,x}, A) \cong H^0(\cdot, A)$ is the <u>0-th singular cohomology</u> sheaf with coefficients in A. If X is locally path-connected, $H^0(\cdot, A) \cong A_X$ the constant sheaf with stalk A.

3) $X = M$ C^∞-manifold of dim n, A: commutative ring. Then $\mathrm{Hom}(H_c^n(\cdot, A_X), A)$ is the <u>orientation</u> sheaf of M.

<u>RK</u>: 1) If X is locally path-connected, $H_{0,X}$ should be thought as constant cosheaf of abelian groups. Recalling that $H_{0,X} \cong \mathbb{Z}[\pi_{0,X}]$, $\pi_{0,X}$ should be thought as constant cosheaf of sets, if X is loc. path-connected.

2) If F is a sheaf with values in a complete category C, and $A \in C$ is "injective", i.e. $\mathrm{Hom}_C(\cdot, A)$ preserves finite limits, then the pre-cosheaf of sets $\mathrm{Hom}(F(\cdot), A)$ defines a cosheaf on the site X_{ef} ($Op(X)$ with only loc. finite coverings).

§ 4. COSHEAVES AND C-SOFT SHEAVES

A less trivial relation between sheaves and cosheaves is obtained by taking compact supports (cf. §3 Ex 10). This is in fact part of Poincaré–Verdier duality (see [KS1]).

Prop: X loc. compact (in part., Hausdorff) space, F pre-sheaf on X of abelian groups. Then the pre-cosheaf of abelian groups

$$\begin{array}{ccc} X \supseteq_{\text{open}} V & \longmapsto & \Gamma_c(V;F) =: F_c(V) \\ \cup & & \uparrow \varepsilon: \text{extension by zero} \\ U & \longmapsto & \Gamma_c(U;F) =: F_c(U) \end{array}$$

is a cosheaf F_c when F is c-soft (i.e. $\forall K \subseteq_{\text{compact}} X$ $\rho_K : \Gamma(X,F) \to \Gamma(K,F) \cong \varinjlim_{U \supseteq K,\ U \text{ open}} \Gamma(U,F)$ is surjective).

Rk: 1) if F is flabby (i.e. $\forall V \subseteq_{\text{open}} X$ $\rho_V : \Gamma(X,F) \to \Gamma(V,F)$ is surjective) $\Rightarrow$ F is c-soft.

(Ex) $\mathcal{E}^\infty_M$, $\mathcal{V}^\infty_M$, $\mathcal{D}b_M$ are c-soft sheaves (M: C^∞-manifold)

$\mathcal{B}_M := \{\text{Sato's hyperfunctions}\}$ is flabby (M: analytic manifold)

2) for any pre-sheaf F, F_c is a flabby pre-cosheaf:

$\forall V \subseteq_{\text{open}} X$ $\varepsilon_V : F_c(V) \to F_c(X)$ is mono. Indeed

$$\begin{array}{ccc} \Gamma_c(V,F) & \xrightarrow{\varepsilon_V} & \Gamma_c(X,F) \text{ commutes} \Rightarrow \varepsilon_V \text{ injective.} \\ \cap! & & \cap! \\ \Gamma(V,F) & \xleftarrow[\rho_V]{} & \Gamma(X,F) \end{array}$$

3) F: c-soft sheaf $\Rightarrow$ F_c: cosheaf $\Rightarrow$ $\mathrm{Hom}_{\mathbb{Z}}(F_c, A)$: sheaf

generalizes the sequence:

$\mathcal{V}^\infty$: c-soft sheaf $\Rightarrow$ $\mathcal{V}^\infty_c$: cosheaf $\Rightarrow$ $\mathcal{D}b_M = (\mathcal{V}^\infty_c)'$: sheaf

proof of the Prop: recall that $F_c(V) := \Gamma_c(V, F) = \Gamma_c(X, F_V)$
where $F_V := \ker\left(F \to j_* j^{-1} F\right)$ for $j : X \setminus V \hookrightarrow X$
and that $\Gamma_c(X, \cdot)$ preserves small filtrant limits.
Hence: 3) a) if $V = \bigcup_i V_i$ is stable by finite $\cap$

$$\Rightarrow \varinjlim_i F_c(V_i) = \varinjlim_i \Gamma_c(X, F_{V_i}) = \Gamma_c(X; \varinjlim_i F_{V_i}) = \Gamma_c(X; F_V) = F_c(V)$$

(the i is filtrant!)

3) b) is obtained by applying $\Gamma_c(X, \cdot)$ to the Mayer-Vietoris exact seq of c-soft sheaves

$$0 \to F_{V_1 \cap V_2} \to F_{V_1} \oplus F_{V_2} \to F_{V_1 \cup V_2} \to 0 \qquad \square$$

In fact, $F \mapsto F_c$ defines a functor $(\cdot)_c : Sh_{c\text{-soft}}(Ab_X) \to Cosh_{flabby}(Ab_X)$

th. (see [Br]): $(\cdot)_c$ is an equivalence of categories.

proof: let us define a quasi-inverse. For this we need

Def (see [Br]): $\mathcal{E}$: cosheaf of abelian groups, $s \in \mathcal{E}(U)$. the <u>support</u> of s is the subset $\operatorname{supp} s \subseteq U$ defined by:

$$x \notin \operatorname{supp} s \iff s \in \operatorname{Im}\left(\varepsilon_{VU} : \mathcal{E}(V) \to \mathcal{E}(U)\right) \quad \exists\, x \notin V \subseteq_{\text{open}} U$$

Let $\mathcal{E}$ be a flabby cosheaf. One proves that $\operatorname{supp} s$ compact and $\operatorname{supp} s \subseteq V \iff s \in \operatorname{Im}\left(\varepsilon_{VU} : \mathcal{E}(V) \to \mathcal{E}(U)\right)$.

For $Z \subseteq V$ locally closed (i.e. $Z = \text{open} \cap \text{closed}$) set:

$$\Gamma^Z(V; \mathcal{E}) := \operatorname{coker}\left(\left\{s \in \mathcal{E}(V) \mid \operatorname{supp} s \subseteq V \setminus Z\right\} \to \mathcal{E}(V)\right)$$

Note that, if Z closed in V,

$$\Gamma^Z(V, \mathcal{E}) = \operatorname{coker}\left(\mathcal{E}(V \setminus Z) \to \mathcal{E}(V)\right)$$

and $\forall\, A \in Ab \quad \operatorname{Hom}_{Ab}(\Gamma^Z(V; \mathcal{E}), A) = \Gamma_Z(V, \operatorname{Hom}(\mathcal{E}, A))$

(recall that $\Gamma_Z(V; F) := \ker\left(F(V) \to F(V \setminus Z)\right) = \left\{s \in F(V) \mid \operatorname{supp} s \subseteq Z\right\}$)

Let $\mathcal{E}^t$ the sheaf associated to the pre-sheaf

$$V \longmapsto \Gamma^V(X; \mathcal{E})$$

then one checks that $(\cdot)^t$ is quasi-inverse of $(\cdot)_c$ □

RK: If $\mathcal{E}$ is a flabby cosheaf of abelian groups and $A \in Ab$ is injective (i.e. $\mathrm{Hom}(\cdot, A)$ exact), then $\mathrm{Hom}(\mathcal{E}, A)$ is a flabby sheaf of abelian groups, hence c-soft $\Rightarrow$ $\mathrm{Hom}(\mathcal{E}, A)_c$ is a flabby cosheaf.
If $\mathcal{E} = F_c$ for a c-soft sheaf, one define a morph. of cosheaves $\mathrm{Hom}(F_c, A)_c \to \mathrm{Hom}(F, A)$ generalizing $\mathcal{D}b_{M,c} = (\mathcal{V}^\infty_{M,c})'_c \xrightarrow{\cong} (\mathcal{V}^\infty_M)'$.

§5. Lawvere's distributions

We saw from some examples that cosheaves on X in some sense look like measures on X. Recall Riesz's theorem:

$$\left\{\text{regular Borel measures on } X\right\} \xrightarrow{\cong} \left\{\begin{array}{l}\mathbb{R}\text{-linear maps } C^0_c(X) \xrightarrow{\alpha} \mathbb{R} \\ \text{s.t. } \alpha(f) \geq 0 \text{ for } f \geq 0\end{array}\right\}$$

(loc. compact)

$$\mu \longmapsto \left(f \longmapsto \int_X f \, d\mu\right)$$

$$X \supseteq_{\text{open}} V \longmapsto e(V) := \sup\left\{e(f) \mid f \in C^0_c(X), \operatorname{supp} f \leq V, 0 \leq f \leq 1\right\} \longleftarrow\!\shortmid e$$

There is an analogous result for cosheaves.
C : cocomplete category; $Sh(X) := Sh(Set_X) := \left\{\text{sheaves of sets on } X\right\}$

Recall that, for a category D, the Yoneda embedding is

$$Y : D \hookrightarrow Fct(D^{op}, Set) \qquad d \longmapsto Y_d := \mathrm{Hom}_D(\cdot, d)$$

and, if D is cocomplete, Y_d preserves colimits (but Y does not preserves colimits, in general!).

For $\mathcal{D} = \mathcal{O}p(X)$, the colimit of $\{V_i\}_{i\in I}$ is $\bigcup_{i\in I} V_i$, hence $\mathcal{O}p(X)$ is cocomplete and

$$Y : \mathcal{O}p(X) \hookrightarrow \overbrace{Fct(\mathcal{O}p(X)^{op}, Set)}^{PSh(X)}$$

factors through $Sh(X) \hookrightarrow PSh(X)$, i.e. we get

$$Y : \mathcal{O}p(X) \hookrightarrow Sh(X)$$

Prop. Y is a cosheaf (with values in $Sh(X)$).

proof: by a Prop. in §3, it is enough to prove that $\forall$ F sheaf of sets, $Hom(Y,F)$ is a sheaf, i.e. that for $\bigcup_{i\in I} V_i = V$ stable by finite $\cap$:

$$Hom(Y,F)(V) = Hom_{Sh(X)}(Y_V, F) = Hom_{PSh(X)}(Y_V, F)$$
$$\cong F(V) \cong \varprojlim F(V_i) \cong \varprojlim Hom_{PSh(X)}(Y_{V_i}, F)$$
$$\cong Hom_{PSh(X)}\Big(\underbrace{\text{“}\varinjlim\text{”}}_{\text{colimit in } PSh(X)} Y_{V_i}, F\Big) \cong Hom_{Sh(X)}\Big(\underbrace{\varinjlim}_{\text{colimit in } Sh(X)} Y_{V_i}, F\Big)$$

□

RK: we used that $Hom(Y,F) \cong F$ for any sheaf F.

By applying $Fct(\cdot, C)$ to Y, we get a functor

$$PCosh(C_X) \xleftarrow[Y_* := \circ Y]{} Fct(Sh(X), C)$$

which admits a left adjoint $Y^+ :=$ left Kan extension of Y_*.

Def: a C-valued Lawvere's distribution on X is a cocontinuous functor $Sh(X) \to C$ (i.e. preserving colimits). We set:
$Law(C_X) :=$ full subcategory of $Fct(Sh(X), C)$: "dual of $Sh(X)$"

Prop: the pair of adjoint functors (Y^+, Y_*) restricts to an equivalence of categories $Cosh(C_X) \cong Law(C_X)$.

One has to think Y^+ as "integration along cosheaves" and Y_* as "restriction to $\mathcal{O}p(X)$ via Y".

(Ex) $y^+(\delta_{x_0})$ is the "stalk at x_0" functor: $F \mapsto F_{x_0}$

$y^+(\#_X)$: $F \mapsto \#(E_F)$ ← étalé space of F

↳ loc. connected

RK: 1) For $C = Ab$, we get $Cosh(Ab_X) \cong Law(Ab_X)$. This restricts to an equivalence $Cosh_{flabby}(Ab_X) \cong \underbrace{Law^{ex}(Ab_X)}$ {exact Lawvere's distributions}

then $(\cdot)_c : Sh_{c\text{-}soft}(Ab_X) \xrightarrow{\cong} Law^{ex}(Ab_X)$: "dual of $Sh_!(X)$"

is the starting point of Lurie's version of Poincaré-Verdier duality: $(\cdot)_c : \infty\text{-}Sh(X) \xrightarrow{\cong} (\infty\text{-}Sh(X))^*$ (see [Lu]).

2) Recall that the forgetful functor $for: Mod(R) \to Set$ (R comm. ring) preserves limits and filtrant colimits and has a left adjoint $R(\cdot): Set \to Mod(R)$ which preserves colimits.

$S \mapsto R(S)$: free R-mod generated by S

then $(R(\cdot), for)$ define adjoint functors $(R(\cdot)^+, for)$

$$\begin{array}{ccc} Fct(Op(X)^{op}, Set) & \underset{R(\cdot)\circ\cdot}{\overset{for\circ\cdot}{\rightleftarrows}} & Fct(Op(X)^{op}, Mod(R)) \\ \uparrow & & \uparrow\downarrow + : \text{left adjoint} \\ Sh(X) & \xleftarrow{\quad for \quad} & Mod(R_X) := \{\text{sheaves of } R\text{-modules}\} \\ & \xrightarrow[R(\cdot)^+]{} & \end{array}$$

In particular, for $U \subseteq X$ open, $R(Y_U)^+$ is the sheaf of R-mod associated to $V \mapsto R\left(Hom_{Op(X)}(V,U)\right) = \begin{cases} R(pt) \cong R & \text{if } V \subseteq U \\ R(\emptyset) := 0 & \text{otherwise} \end{cases}$

i.e. $R(Y_U)^+ \cong R_U := Ker\left(R_X \to j_* j^{-1} R_X\right)$ for $j : X \setminus U \hookrightarrow X$.

We get a cosheaf $RY : Op(X) \xrightarrow{Y} Sh(X) \xrightarrow{R(\cdot)^+} Mod(R_X)$ $U \mapsto R_U$, which is faithful $\left(Hom_{Op(X)}(V,U) \hookrightarrow Hom_{R\text{-mod}}(R_V, R_U) \cong \Gamma(V, R_U)\right)$

Let A be an R-abelian category. By applying $\mathrm{Fct}(\cdot, A)$ to RY we get a functor $P\mathrm{Cosh}(A_X) \leftarrow \mathrm{Fct}(\mathrm{Mod}(R_X), A)$ which admits a left adjoint RY^+. (RY_* these restrict to an equiv.

$$\mathrm{Cosh}(A_X) \cong \mathrm{Fct}^{\mathrm{cocont}}(\mathrm{Mod}(R_X), A)$$

↖ cocontinuous functors

A similar construction for X_{sa} := subanalytic site of an analytic manifolds allows to define the Whitney functor $\cdot \otimes^{\omega} \mathcal{C}^{\infty}_X$ from the sub-analytic cosheaf $\mathcal{C}^{\infty,\omega}_{X_{sa}}$ of Whitney functions (see [KS2]).

§ 6. COSHEAVES AND COMPLETE SPREADS

From now on $C = \mathrm{Set}$ and $\mathrm{Cosh}(X) := \mathrm{Cosh}(\mathrm{Set}_X)$.

Recall that there is an equivalence of categories

$$\mathrm{Sh}(X) \cong \{\text{local homeom. } Y \xrightarrow{p} X\}$$

$$F \longmapsto (E_F := \text{étalé space of } F \to X)$$

$$\{\text{c°-sections of } p\} \longleftarrow\!\shortmid Y \xrightarrow{p} X$$

there is a similar result for cosheaves (of sets):

Th: $\mathrm{Cosh}(X) \cong \{\text{Fox's complete spreads } Y \xrightarrow{p} X\}$

$$\mathcal{C} \longmapsto (D\mathcal{C} := \text{display space of } \mathcal{C} \to X)$$

$$p^{\#}_{*}Y \longleftarrow\!\shortmid Y \xrightarrow{p} X$$

(for a proof, see for example [W]). Let's explain the objects involved.

Def: $f: Y \to X$ map (X, Y: top. spaces)

i) f is a spread if the topology of Y is generated by the connected components of $f^{-1}(U)$ for $U \underset{\text{open}}{\subseteq} X$

(in particular, f is continuous and Y loc. connected).

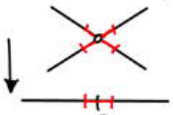
ii) f is a <u>complete spread</u> if f is a spread and $\forall x \in X$

$f^{-1}(x) \xrightarrow{\cong} (f_* \#_Y)_x := \varinjlim_{U \ni x} \#_Y(f^{-1}(U))$: costalk of the cosheaf $f_* \#_Y$

$y \longmapsto \left(C_{f^{-1}(U),Y}\right)_{U \ni x}$

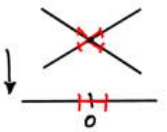
<u>RK</u> : 1) f spread + Y T_0 $\left(\text{for } y_1, y_2 \ \exists\ U \text{ open } y_1 \in U \not\ni y_2\right)$, or f complete $\Rightarrow$ the fibers of f are relatively Hausdorff and totally disconnected $\left(\#f^{-1}(x) = f^{-1}(x)\right)$. Compare to local homeom., which have discrete fibers.

2) Condition ii) means that the fiber of f at x "behave nicely near x". Contrarily to coverings (with X connected + loc. connected) whose fibers have the same cardinality, the cardinality of the fibers of a complete spreads may vary.

(Ex) a) i) 

spread not complete

$p^{-1}(0) = \emptyset \neq \left(p_* \#_Y\right)_0 = \{\bullet, \bullet, \bullet\}$

ii) 

complete spread

$p^{-1}(0) \xrightarrow{\cong} \left(p_* \#_Y\right)_0 = \{\bullet\}$

b) $Z \overset{j}{\subset} X$ inclusion

i) Z open $\Leftrightarrow$ j loc. homeom.

ii) Z loc. connected $\Leftrightarrow$ j spread

iii) Z closed & "loc. connected in X" $\Leftrightarrow$ j complete spread

3) f spread + proper (i.e. closed map + relatively Hausdorff compact fibers) $\Rightarrow$ f complete spread. Compare to: f loc. homeom + proper $\Rightarrow$ f covering.

4) If $\mathcal{E}$ is a cosheaf, its <u>display space</u> $D\mathcal{E} := \coprod_{x \in X} \mathcal{E}_x$, endowed with a natural topology and the projection $D\mathcal{E} \to X$ is a complete spread.

(Ex) $D\#_Y \cong Y$ + topology generated by the connected components of the open subsets of Y

If Y is loc. connected $\Rightarrow$ $\#_Y$ cosheaf and $D\#_Y \cong Y$.

§ 7. Topological branched coverings

Complete spreads were introduced by Fox in the 50's to give a purely topological notion of branched/ramified coverings of algebraic/complex geometry (see [Fox]).

Def: A <u>ramified/branched covering</u> is $f: Y \to X$ s.t.

i) f is a complete spread;

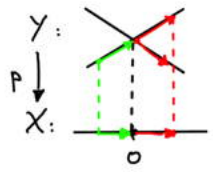
ii) set $X_0 := \{x \in X \mid \exists\ U \overset{\text{open}}{\ni} x \text{ s.t. } f: f^{-1}(U) \to U \text{ covering}\} \subseteq_{\text{open}} X$;
then $\overline{f^{-1}(X_0)} = Y$ (X_0: <u>ordinary points</u>; $X \setminus X_0$: <u>singular points</u>)

iii) $\overline{X_0} = X$

(+ some loc. connectedness property of X_0, $f^{-1}(X_0)$)

If moreover $X = \bigsqcup_{i \in I} X^{(i)}$ is a stratification (i.e. $X^{(i)} \cap \overline{X^{(j)}} \neq \emptyset \iff X^{(i)} \subseteq \overline{X^{(j)}}$) s.t. $Y = \bigsqcup_{i \in I} f^{-1}(X^{(i)})$ is a stratification and $\forall\ i \in I$ $f|_{f^{-1}(X^{(i)})}$ is a covering, one call f a <u>stratified branched covering</u> (see [W]).

(Ex)



stratified branched covering

strata: $\{0\}$, $X_0 := X \setminus \{0\}$.

γ: "entering" path γ: "exit" path

Def: An <u>entering path</u> is $\gamma: [0,1] \to X$ s.t. if $t_0 \leq t_1$ in $[0,1]$ and $\gamma(t_0) \in X^{(i_0)}$, $\gamma(t_1) \in X^{(i_1)} \Rightarrow X^{(i_0)} \subseteq \overline{X^{i_1}}$.

Roughly, an entering path may stay in the same stratum or enter into a "smaller" one.

th: $f: Y \to X$ is a stratified branched covering $\iff$ f has the unique lifting property for "families of entering paths".

RK: Under the equivalence between cosheaves and complete spreads, the stratified branched coverings correspond to constructible cosheaves i.e. cosheaves $\mathcal{C}$ s.t. $\forall i \in I$ $\mathcal{C}|_{X(i)} \cong S_i^Y$ for a set S_i, where $S_i^Y(U) = S_i^{\bigsqcup X(U)}$.
If the stratification is fixed, these may be classified by their "monodromy" (see [W]).

REFERENCES

[Br]: G. E. Bredon, Sheaf theory. Second edition (1997).

[Fox]: R. H. Fox, Covering spaces with singularities (1957).

[KS1]: M. Kashiwara and P. Schapira, Sheaves on manifolds (1990).

[KS2]: ———, Moderate and formal cohomology associated to constructible sheaves (1996).

[Lu]: J. Lurie, Higher topos theory (2009).

[Sch]: J.-P. Schneiders, Quasi-abelian categories and sheaves (1999).

[W]: J. Woolf, the fundamental category of a stratified space (2009)

Cheapest Routes with Integer Linear Programming

Michele Barbato (*)

Abstract. Combinatorial optimization deals with the optimization of a function over a finite, but huge, set of elements. It has a great impact on real life, as several problems arising in logistics, scheduling, facility location, to cite a few, can be stated as combinatorial optimization problems. Often, problems of this kind can be expressed as integer linear programs (ILP), *i.e.,* problems in which the function to be optimized is linear and so are the constraints that define the discrete feasibility set. We provide an introduction to ILP through examples by motivating related theoretical questions (*e.g.,* the polyhedral study). We will consider as initial case of study the Traveling Salesman Problem (TSP). The TSP consists in finding the cheapest route that visits a prescribed set of cities exactly once, before returning to the starting point. As such, the TSP is a prototype of several other problems arising in logistics. Subsequently, we will describe the double TSP with multiple stacks, that combines the construction of cheapest routes with loading constraints. We will reveal links between this problem and the TSP, as well as the limitations that a purely routing-based approach has for this problem.

1 Traveling Salesmen in Combinatorial Optimization

A *combinatorial optimization problem* consists in minimizing a function $f: S \to \mathbb{R}$ where S is a set of finite cardinality. In this writing we focus on the special case in which $S \subseteq \{0,1\}^d$ for some $d \in \mathbb{Z}_+$ and f is a linear function. This setting is powerful enough to tackle a wide variety of problems arising from real-world applications.

We illustrate the approach above on the celebrated Traveling Salesman Problem. In order to present this problem, we need some definitions. Let $G = (V, A)$ be the complete digraph on n vertices, with $V = \{1, \dots, n\}$ and $A = \{(i,j): i \neq j \in V\}$. The set A is the *arc set* of G. Interpreting G as a network, A models the direct connections between pairs of nodes in the network. We call any $c \in \mathbb{R}_+^{|A|}$ a *cost vector* of G. It can be thought as the vector containing the distances between pairs of nodes in the network. A *Hamiltonian circuit* of G is a connected subgraph of G containing all vertices of G and such that every

(*)Laboratoire Informatique de Paris Nord (LIPN), Université Paris 13 - 99, avenue Jean-Baptiste Clément, F-93430 Villetaneuse, France; E-mail: michele.barbato@lipn.univ-paris13.fr .
Seminar held on April 13th, 2016.

vertex has exactly one entering and one leaving arc. It can be interpreted as a single route that visits each node in the network exactly once before to come back to the starting node. The *Traveling Salesman Problem (TSP)* is to find the least cost Hamiltonian circuit of G with respect to the cost vector c. If we define $c(H) \coloneqq \sum_{a \text{ arc of } H} c_a$ for every subgraph H of G, the TSP can be expressed as:

$$\begin{aligned} \min \quad & c(H) \text{ s.t.} \\ & H \text{ is a Hamiltonian circuit of } G \end{aligned}$$

The Hamiltonian circuits of $G = (V, A)$ are in one-to-one correspondence with the solutions to the following system of constraints [6]:

$$\sum_{j \in V \setminus \{i\}} x_{ij} = 1 \qquad \forall i \in V \tag{1}$$

$$\sum_{j \in V \setminus \{i\}} x_{ji} = 1 \qquad \forall i \in V \tag{2}$$

$$\sum_{i \in S, j \in \bar{S}} x_{ij} \geq 1 \qquad \forall \varnothing \neq S \subseteq V \setminus \{1\} \tag{3}$$

$$x_{ij} \in \{0, 1\} \tag{4}$$

Given a solution $x^\star$ to (1)–(4), we reconstruct the corresponding Hamiltonian circuit H of G as follows: (i, j) is an arc of H if and only if $x^\star_{ij} = 1$.

Hence the TSP is the problem $\min\{cx : x \text{ satisfies (1)–(4)}\}$. Note that this is exactly the setting given at the beginning of this writing for combinatorial optimization problems. We call systems of constraints combining linear inequalities with integrality requirements for the variables, *integer linear programming formulations*. For instance, we say that (1)–(4) is an integer linear programming formulation for the TSP.

An Interlude on Polyhedra

A geometrical interpretation of (1)–(4) is that its solutions are binary points contained in a geometrical region of $\mathbb{R}^{|A|}$ delimited by linear hyperplanes (the linear constraints (1)–(3)). We can exploit this geometrical perspective to design effective algorithms to solve the TSP.(†) To illustrate this idea, we introduce some new definitions.

Definition 1.1 A *polyhedron* of $\mathbb{R}^m$ is $P = \{x \in \mathbb{R}^m : Ax \leq b\}$, where $A \in \mathbb{Q}^{\ell \times m}$ and $b \in \mathbb{Q}^\ell$ for some positive integers ℓ and m. A bounded polyhedron is called *polytope*. The *dimension* of a polyhedron P, denoted $\dim(P)$, is the dimension of the smallest affine space containing P. A *vertex* v of a polyhedron P is a point of P that cannot be expressed as a strict convex combination of two points of P other than v.

Definition 1.2 A *valid inequality* for a polyhedron P is a linear inequality $ax \leq \delta$ verified by all points of P. A *face* of a polyhedron P is any set of the form $P \cap \{x : ax = \delta\}$ with

(†) Please, note that the TSP is theoretical hard to solve. Nevertheless, nowadays the framework presented in this writing lets us solve non-trivial instances of the TSP with up to 85900 vertices in reasonable time. This can be accomplished with the state-of-the-art solver CONCORDE [1] for example.

$ax \leq \delta$ a valid inequality of P. A face F of a polyhedron P is *proper* if $F \subset P$. A *facet* of a polyhedron P is a proper face maximal with respect to the inclusion.

Definition 1.3 The *integer hull* P_I of a polyhedron P is the convex hull of its integer points, that is $P_I = conv(P \cap \mathbb{Z}^m)$.

The following result is well-known, see *e.g.,* [4].

Proposition 1.4 *Let $P \subseteq \mathbb{R}^m$ be a polytope. Then*

- *P_I is a polytope;*
- *$\min\{cx : x \in P\}$ has one of its optimal solutions on a vertex of P, for every $c \in \mathbb{R}^m$.*

We define $\mathcal{L}_n$ to be the *linear relaxation* of (1)–(4) *i.e.,* the polytope obtained from that system by replacing (4) by $0 \leq x_{ij} \leq 1$. Let us call TSP_n the integer hull of $\mathcal{L}_n$. Note that the set of vertices of TSP_n coincides with the set of points verifying (1)–(4). Then Proposition 1.4 implies that the TSP is equivalent to $\min\{cx : x \in TSP_n\}$. The optimization over $\mathcal{L}_n$ of a linear function is theoretically simple, meaning that it can be performed in time polynomial in the TSP instance size [10]. For the optimization of a linear function over TSP_n we do not have found, until now, similar positive results.

However, one of the greatest advances in combinatorial optimization has been the design of algorithms able to solve large TSP instances in reasonable time. One of the most effective of these methods is the branch-and-cut method, that we present in next section.

Branch-and-Cut Algorithm

The branch-and-cut algorithm is based on the following simple ideas. The starting observation is that $\min\{cx : x \in \mathcal{L}_n\} \leq \min\{cx : x \in TSP_n\}$ because $\mathcal{L}_n \supseteq TSP_n$. In addition, if $ax \leq \delta$ is a valid inequality for TSP_n and $\mathcal{F}$ is a polytope containing TSP_n, we have

$$\min\{cx : x \in \mathcal{F}\} \leq \min\{cx : x \in \mathcal{F} \cap \{x : ax \leq \delta\}\} \leq \min\{cx : x \in TSP_n\}$$

Such a $ax \leq \delta$ is also called *strengthening inequality*, since it can be used to tighten the polytope $\mathcal{F}$ so that it is closer to TSP_n.

The last observation is that the whole set of vertices of TSP_n can be listed in at most 2^n steps, because each vertex is a n-dimensional binary vector.

The following description of the branch-and-cut method is taken from [4]. The algorithm constructs a list L of optimization problems of the linear function cx over different polytopes. These optimization problems will be called *linear programs*. We indicate the i-th linear program to be solved by LP_i and its optimal value by z^i. We will use LP_i also to indicate the polytope over which we optimize the i-th linear program. The algorithm also keeps track of the minimum upper bound $\bar{z}$ for the TSP instance: an *upper bound* for the TSP is the value of $c\tilde{x}$ with $\tilde{x}$ solution to (1)–(4). We indicate with $x^\star$ the solution corresponding to $\bar{z}$.

The first linear program in L is $LP_0 = \min\{cx : x \in \mathcal{L}_n\}$. Initially $\bar{z} = +\infty$ and $x^\star$ is undefined. The algorithm is now as follows:

(a) (TERMINATION STEP). If $L = \varnothing$, the solution $x^\star$ is optimal. STOP.

(b) (NODE SELECTION). Choose a $LP_i \in L$ and remove it from L.

(c) (BOUNDING STEP). Solve LP_i. If infeasible, go to TERMINATION STEP. Else let x^i be an optimal solution of LP_i of value z^i.

(d) (PRUNING STEP).

- If $z^i \geq \bar{z}$ go to TERMINATION STEP.
- If x^i is feasible for the TSP, set $\bar{z} = z^i$ and $x^\star = x^i$. Go to TERMINATION STEP.
- If none of the previous applies in this step, proceed with the following.

(e) Decide whether to add some strengthening inequalities $a^i x \leq \delta^i$ for $i = 1, \ldots, s$. If so go back to the BOUNDING STEP, with $LP_i \leftarrow LP_i \cap \{a^i x \leq \delta : i \leq s\}$. Otherwise, go to the BRANCHING STEP below.

(f) (BRANCHING STEP). Select $j \in \{1, \ldots, n\}$ such that $x^i_j \notin \{0, 1\}$ and from LP_i construct the linear programs $LP_{i_1} \cap \{x_j = 1\}$ and $LP_{i_2} \cap \{x_j = 0\}$. Add LP_{i_1} and LP_{i_2} to L and go to TERMINATION STEP.

The above framework can be easily adapted to combinatorial optimization problems other than the TSP.

Motivations for the Polyhedral Study

In the branch-and-cut algorithm it is very important to use appropriately the strengthening inequalities in the 5th step. Indeed, one of the assumptions under which it is effective is that the LP_i's can be solved quickly in practice. Adding too many inequalities in step (e) could slow down the entire algorithm.

In general, it is convenient to add strengthening inequalities that can produce a considerable improvement in the objective function of the linear programs to be solved. From this observation it follows that it is quite important to have informations on the strength of a valid inequality.

Well-known results from polyhedral theory [4] guarantee that those inequalities defining facets of TSP_n are the strongest in this sense, at least in theory. Hence, much of the literature on the TSP focuses on the identification of the inequalities defining facets of TSP_n. For instance, constraints (3) define facets for TSP_n, see [9]. Besides the maximality property of a facet, one can identify inequalities defining facets for a polytope using a dimensional argument. Indeed, it is well-known that if F is a facet of a polytope P, then $\dim(F) = \dim(P) - 1$. For instance, constraints (3) define facets for TSP_n, see [9].

An Alternative Formulation for the TSP

The literature on the TSP (and other combinatorial optimization problems) has also focused on the ways it can be formulated as an integer linear program. Indeed, a combinatorial optimization problem admits several integer linear programming formulations, that can be more or less informative. The correctness of the branch-and-cut algorithm holds for any of the formulations, but the performance can vary drastically. This is the case also for the TSP, see *e.g.,* [10, 16] for exhaustive lists of formulations and a comparative analysis of their behavior in exact algorithms.

The best formulation for the TSP from a computational standpoint is in fact given by (1)–(4). However, in other situations (as, for example, the one presented in next section) it can be reasonable to resort to other formulations. Here we present an alternative formulation for the TSP over the digraph G. It can be seen that the Hamiltonian circuits of G are in one-to-one correspondence with the solutions to the following system of inequalities [17].

$$\sum_{j \in V \setminus \{i\}} x_{ij} = 1 \qquad \forall i \in V \tag{5}$$

$$\sum_{j \in V \setminus \{i\}} x_{ji} = 1 \qquad \forall i \in V \tag{6}$$

$$y_{ij} + y_{ji} = 1 \qquad \forall i \neq j \in V \setminus \{1\} \tag{7}$$

$$y_{ij} + y_{jk} + y_{ki} \leq 2 \qquad \forall \text{ distinct } i, j, k \in V \setminus \{1\} \tag{8}$$

$$x_{ij} \leq y_{ij} \qquad \forall i \neq j \in V \setminus \{1\} \tag{9}$$

$$x_{ij} \in \{0, 1\} \qquad \forall i, j \in V \tag{10}$$

$$y_{ij} \in \{0, 1\} \qquad \forall i, j \in V \setminus \{1\} \tag{11}$$

Let $(x^\star, y^\star)$ be a solution to (5)–(11). Then the corresponding Hamiltonian circuit H of G contains the arc (i, j) if and only if $x^\star_{ij} = 1$. This is exactly the same correspondence used in formulation (1)–(4). So why do we introduce new variables y_{ij}? The answer is that $y^\star_{ij} = 1$ is equivalent to say that in H vertex i is in the path from vertex 1 to vertex j. If we think that 1 is the starting vertex of all Hamiltonian circuits of G, then $y^\star_{ij} = 1$ means that i *precedes* j in the Hamiltonian circuit corresponding to (x^*, y^*). Hence, $y^\star$ describes a linear ordering on $\{1, \ldots, n\}$ induced by H. From now on, we will call variables x_{ij} *arc variables* and variables y_{ij} *precedence variables*.

The presence of the precedence variables lets us replace the large set of inequalities (3) in formulation (1)–(4) with the small set of inequalities (7)–(8). Indeed, observe that (3) contains a number of inequalities exponential in n (one inequality for each nonempty subset of $V \setminus \{1\}$), whereas (7)–(8) has "only" $O(n^3)$ members. On the other hand, as it is, formulation (5)–(11) performs quite poorly in a branch-cut-algorithm.

A polyhedral study of formulation (5)–(11) (*i.e.,* finding valid inequalities, studying the facets of the integer hull of its linear relaxation) seems to be necessary to improve the computational results obtainable using this formulation in combination with such a method.

Let $\mathcal{PL}_n$ denote the linear relaxation of formulation (5)–(11). Let us call $PTSP_n$ the integer hull of $\mathcal{PL}_n$. Several families of inequalities valid for $PTSP_n$ are known, see

e.g., [8]. However, as far as we know, no condition is known for them to be facet-defining. To start a thorough study of the facets of $PTSP_n$ it could be useful, as pointed out in previous section, to have a closed formula for $\dim(PTSP_n)$. In fact, one can prove the following result.

Proposition 1.5 ([2]) *The dimension of* $PTSP_n$ *is* $\frac{3n^2-9n+4}{2}$ *for* $n \geq 5$.

The simple argument used in the proof of previous proposition is to exhibit a set of $\frac{3n^2-9n+4}{2}+1$ affinely independent points satisfying (5)–(11). We omit this proof since quite long.

The Double TSP with Multiple Stacks

In this section, we study a generalization of the TSP introduced in [15], namely the *double TSP with multiple stacks*. In this problem, n items have to be picked up in one city, stored in a vehicle having s identical stacks of finite capacity, and delivered to n customers in another city. We will assume that the pickup and the delivery cities are very far from each other, thus the pickup phase has to be entirely completed before the delivery phase starts. The pickup (resp. delivery) phase consists in performing a Hamiltonian circuit, *i.e.*, starting from a depot, the n pickup (resp. delivery) locations have to be visited in sequence exactly once before coming back to the depot. Each time a new item is picked up, it is stored on the top of an available stack of the vehicle according to its capacity and no rearrangement of the stacks is allowed. During the delivery circuit the stacks are unloaded following a last-in-first-out policy, that is, only items currently on the top of their stack can be delivered. Each item must be delivered to a corresponding customer, that is items and customers are paired. The goal is to find the pickup and delivery circuits which minimize the total traveled distance, consistent with the last-in-first-out rule.

The double TSP with multiple stacks is at least as hard as the TSP since, when the vehicle has only one stack, it corresponds to the TSP: indeed, in this case, due to the last-in-first-out policy, the delivery circuit is nothing but the pickup circuit performed in the reverse order. However, this problem seems to be more difficult than the TSP to solve in practice.

Since its first appearance, the double TSP with multiple stacks has received increasing attention. Both exact algorithms and heuristics have been designed for this problem over the past few years. Regarding the exact algorithms, in [11] and [12], the authors design a procedure to iteratively generate the k-best TSP pickup and delivery solutions and to find the best combination satisfying the last-in-first-out consistency. Several exponential and polynomial size mixed integer linear programming formulations have been proposed and tested in branch-and-cut frameworks [13, 14]. In [5], the authors adapt a branch-and-cut algorithm for the pickup and delivery TSP with multiple stacks to the double TSP with multiple stacks.

From a computational point of view, these algorithms clearly show that the double TSP with multiple stacks is extremely hard to solve with exact methods. In particular, the difficulty of the problem increases with the capacity of the stacks [14].

In the remainder of this report we will assume that the capacity of the stacks is infinite.

Formal Description and Integer Linear Programming Formulation

Since the double TSP with multiple stacks is a routing problem, it is quite natural to model it by using digraphs. In addition, since in the problem the number of items equals the number of customers we will use the same digraph to model both the pickup and delivery cities.

More precisely, we denote by G_n the complete digraph having $V = \{0, \dots, n\}$ as vertex set and $A = \{(i, j) : i \neq j \in V\}$ as arc set. Vertex 0 represents the depot of a city. For every $i = 1, \dots, n$, vertex i is the location where there is item i, if G_n models the pickup city; instead if G_n represents the delivery city, vertex i is the location of customer i. In the following we assume that item i must be delivered to customer i.

To continue, we need to recall that each Hamiltonian circuit of G_n induces a linear ordering on the vertices $1, \dots, n$. We have already seen this in the arc-precedence variable formulation for the TSP in previous section. Here, the starting vertex is 0.

In the following definition the reader should think that s is the number of stacks defining an instance of the double TSP with multiple stacks. The definition formalizes the dispositions of the items in the s stacks that, consistently with the last-in-first-out rule, let to perform a pickup and a delivery route.

Definition 1.6 Given a positive integer s, an *s-loading plan* for two Hamiltonian circuits is a partition of $\{1, \dots, n\}$ into s sets such that each set can be ordered as a subsequence of both linear orderings induced by one circuit and the reverse of the other one.

Definition 1.7 A couple of Hamiltonian circuits is *s-consistent* if it admits an s-loading plan.

In other words, the s-consistent couples of Hamiltonian circuits are the feasible solutions to the double TSP with s stacks.

We now describe the double TSP with multiple stacks in terms of graphs. An instance of this problem with n items is defined on the digraph G_n by two cost vectors c^P and c^D on its arcs, and a number s of stacks. The digraph G_n models both cities; vertex 0 is the depot and the other ones are the locations where the items have to be picked up or delivered. The item collected at vertex i must be delivered to the customer in vertex i, for every $i = 1, \dots, n$. The vectors c^P and c^D represent the distances between the locations of the pickup and delivery cities, respectively. The pickup and the delivery circuits are two Hamiltonian circuits of G_n. In the double TSP with multiple stacks, one seeks a solution of minimum cost, that is, a pair of s-consistent Hamiltonian circuits C_1 and C_2 whose total cost $c^P(C_1) + c^D(C_2)$ is minimum.

One could wonder why we do not take into account the construction of an s-loading plan consistent with an optimal solution in the above description. After all, a company facing a problem similar to the double TSP with multiple stacks in real-life would like

to know the disposition of the items inside its trucks! Fortunately enough there is a complete characterization of s-consistent couples of Hamiltonian circuits, that also yields a polynomial time algorithm to reconstruct a corresponding s-loading plan.

Proposition 1.8 (3, 19) *Two Hamiltonian circuits of G_n are s-consistent if and only if no $s+1$ vertices of $V \setminus \{0\}$ appear in the same order in both circuits.*

In [3] the authors give an algorithm that verifies the condition of Proposition 1.8 in $O(n \log n)$ time and whose output is exactly an s-loading plan for the two input circuits, if this exists.

Now let us consider (x^P, y^P, x^D, y^D), a vector such that (x^T, y^T) verifies (5)–(11) for $T = P, D$. Essentially, (x^P, y^P, x^D, y^D) represents a pair of Hamiltonian circuits of G_n. We will assume that variables indexed with P describe Hamiltonian circuits in the pickup graph, and variables indexed with D describe Hamiltonian circuits in the delivery graph. Now, remember that $y^T_{ij} = 1$ means that i precedes j in the Hamiltonian circuit corresponding to (x^T, y^T) for $T = P, D$. Hence, using Proposition 1.8, we have that (x^P, y^P, x^D, y^D) represents a s-consistent pair of Hamiltonian circuits if and only if it satisfies

$$\sum_{i=1}^{s} (y^P_{j_i j_{i+1}} + y^D_{j_i j_{i+1}}) \geq 1 \qquad \text{for all distinct } j_1, \ldots, j_{s+1} \in \{1, \ldots, n\}. \tag{12}$$

An integer linear programming formulation for the double TSP with multiple stacks now follows from the observation that its solutions are in one-to-one correspondence with the binary vectors (x^P, y^P, x^D, y^D) satisfying (12) and such that (x^T, y^T) satisfies (5)–(11).

We call *DTSPMS polytope* the convex hull of the solutions of this formulation for the double TSP with multiple stacks, denoted $DTSPMS_{n,s}$. In next section we provide some results on the structure of the DTSPMS polytope.

Polyhedral and Computational Results

In this section we list some properties of the DTSPMS polytope. More precisely, we link the $DTSPMS_{n,s}$ with $PTSP_{n+1}$. For the sake of brevity we omit the proofs. These results and their proofs can be found in [2].

The first result is that the dimension of $DTSPMS_{n,s}$ is twice the dimension of $PTSP_{n+1}$.

Proposition 1.9 *For $n \geq 5$ and $s \geq 2$, we have* $\dim(DTSPMS_{n,s}) = 2 \dim(PTSP_{n+1})$.

More importantly, every facet-defining inequality of $PTSP_{n+1}$ induces two facet-defining inequalities of $DTSPMS_{n,s}$.

Proposition 1.10 *For $n \geq 5$ and $s \geq 2$, if $ax + by \geq c$ defines a facet of $PTSP_{n+1}$, then $ax^T + by^T \geq c$ defines a facet of $DTSPMS_{n,s}$, for $T = P, D$.*

As a consequence of a recent result on extended formulations [7], Proposition 1.10 yields a super-polynomial number of facets of the DTSPMS polytope yet they are not sufficient

to characterize the convex hull. We show this with an example. For every Hamiltonian circuit H let us indicate its reverse with $H^{\leftarrow}$. Given H a Hamiltonian circuit, we indicate by (x^H, y^H) the vertex of $PTSP_n$ corresponding to H.

Consider the following example of the double TSP with two stacks. Suppose that there are 5 items. Let c^P and c^D be such that $P^\star = 0,1,2,3,4,5,0$ and $D^\star = 0,1,2,5,4,3,0$ are Hamiltonian circuits which minimize c^P and c^D, respectively. Suppose also that costs are *symmetric*, that is, $c^P_{ij} = c^P_{ji}$ and $c^D_{ij} = c^D_{ji}$ for all $i \neq j \in V$.

The couples $(P^\star, D^\star)$ and $(P^\star, D^{\star\leftarrow})$ are not 2-consistent as it can be seen by applying the result of Proposition 1.8. The point $S^\star = (x^{P^\star}, y^{P^\star}, \frac{1}{2}x^{D^\star} + \frac{1}{2}x^{D^{\star\leftarrow}}, \frac{1}{2}y^{D^\star} + \frac{1}{2}y^{D^{\star\leftarrow}})$ does not belong to $DTSPMS_{5,2}$. However, $S^\star$ belongs to the intersection of $PATSP_6 \times PATSP_6$ with the linear relaxation of our formulation for the double TSP with two stacks and is actually one of its vertices; the point $S^\star$ is also an optimal solution to the minimization problem of $c^P x^P + c^D x^D$ over this polytope. Moreover, no facet given by Proposition 1.10 can cut off $S^\star$ since both $(x^{P^\star}, y^{P^\star})$ and $(\frac{1}{2}x^{D^\star} + \frac{1}{2}x^{D^{\star\leftarrow}}, \frac{1}{2}y^{D^\star} + \frac{1}{2}y^{D^{\star\leftarrow}})$ belong to $PTSP_6$.

More generally, for symmetric costs, the value of the linear relaxation of our formulation cannot be better than the value obtained by independently optimizing the pickup and delivery circuits, that is, by independently solving two symmetric Traveling Salesman problems. The following family of valid inequalities strengthens the linear relaxation of our formulation by cutting off such extreme points.

Proposition 1.11 *Let C be a circuit of $G_n \setminus \{0\}$ with $|C| \geq s+1$. Then the inequality*

$$y^P(C) + y^D(C) \geq \left\lceil \frac{|C|}{s} \right\rceil \tag{13}$$

is valid for $DTSPMS_{n,s}$.

Proof. For each arc a of C, consider the inequality (12) associated with the s consecutive arcs of C, starting from a. Summing these $|C|$ inequalities and dividing by s yields

$$y^P(C) + y^D(C) \geq \frac{|C|}{s}. \tag{14}$$

The vertices of $DTSPMS_{n,s}$ being integer, we round up the right hand side of (14) to get (13). □

Inequality (13) is a *circuit inequality of order* $|C|$. When $s = 2$ and $|C|$ is odd, we call it an *odd circuit inequality.*

The point $S^\star$ of the example above is cut off by the odd circuit inequality associated with the circuit $H = 0,5,4,3,2,1,5,0$.

The validity of inequalities (13) can also be proven using the results of Sassano on the set covering polytope [18]. The *set covering polytope* associated with a 0/1 matrix A is the convex hull of the 0/1 solutions to $Ax \geq \mathbf{1}$. Let S be the set of points $(y^P, y^D) \in$

$\{0,1\}^{n(n-1)} \times \{0,1\}^{n(n-1)}$ satisfying inequalities (12) together with

$$y_{ij}^T + y_{ji}^T \geq 1 \qquad \text{for all } i \neq j \in V \setminus \{0\} \text{ and } T = P, D, \tag{15}$$

$$y_{ij}^T + y_{jk}^T + y_{ki}^T \geq 1 \qquad \text{for all } i \neq j \neq k \neq i \in V \setminus \{0\} \text{ and } T = P, D. \tag{16}$$

Clearly, $conv(S)$ is a set covering polytope, and $proj_{(y^P, y^D)}(DTSPMS_{n,s})$ is one of its faces because (15) is a relaxation of (7) for $T = P, D$. Then, for each odd circuit C, the inequality $y^P(C) \geq \left\lceil \frac{|C|}{s} \right\rceil$ is a so-called *s-rose inequality of order* $|C|$ of the restriction of $conv(S)$ to the variable set y^P [18]. Using the Lifting Theorem 4.1 of [18], one can lift this inequality into the circuit inequality associated with C.

Computational Consequences

For the sake of brevity we do not report computational results obtained on the double TSP with multiple stacks using the results of this section. They can be found in [2]. However, we can summarize them as follows. Using some valid inequalities for $PTSP_{n+1}$ introduced in [8] and the odd circuit inequalities (13), a branch-and-cut algorithm presented in [2] has solved to optimality for the first time instances of the double TSP with two stacks with up to 18 items within a limit of three hours of CPU time. In addition, the same algorithm has solved to optimality all instances with less than 18 items, also resulting faster than previous exact methods for the double TSP with multiple stacks.

References

[1] D. Applegate, R. Bixby, V. Chvátal, and W. Cook, *Concorde: A code for solving traveling salesman problems.* `http://www.math.princeton.edu/tsp/concorde.html` (2003).

[2] M. Barbato, R. Grappe, M. Lacroix, and R. W. Calvo, *Polyhedral results and a branch-and-cut algorithm for the double traveling salesman problem with multiple stacks.* Discrete Optimization, 21 (2016), 25–41.

[3] M. Casazza, A. Ceselli, and M. Nunkesser, *Efficient algorithms for the double traveling salesman problem with multiple stacks.* Computers & Operations Research, 39(5) (2012), 1044–1053.

[4] M. Conforti, G. Cornuéjols, and G. Zambelli, "Integer programming". Volume 271. Springer, 2014.

[5] J.-F. Côté, C. Archetti, M. G. Speranza, M. Gendreau, and J.-Y. Potvin, *A branch-and-cut algorithm for the pickup and delivery traveling salesman problem with multiple stacks.* Networks, 60(4) (2012), 212–226.

[6] G. Dantzig, R. Fulkerson, and S. Johnson, *Solution of a large-scale traveling-salesman problem.* Journal of the Operations Research Society of America, 2(4) (1954), 393–410.

[7] S. Fiorini, S. Massar, S. Pokutta, H. R. Tiwary, and R. D. Wolf, *Exponential lower bounds for polytopes in combinatorial optimization.* J. ACM, 62(2) (May 2015), 17:1–17:23.

[8] L. Gouveia and P. Pesneau, *On extended formulations for the precedence constrained asymmetric traveling salesman problem.* Networks, 48(2) (2006), 77–89.

[9] M. Grötschel and M. W. Padberg, *Lineare Charakterisierungen von Travelling Salesman Problemen.* Zeitschrift für Operations Research, 21(1) (Feb. 1977), 33–64.

[10] G. Gutin and A. P. Punnen, "The traveling salesman problem and its variations". Volume 12. Springer Science & Business Media, 2006.

[11] R. M. Lusby and J. Larsen, *Improved exact method for the double TSP with multiple stacks.* Networks, 58(4) (2011), 290–300.

[12] R. M. Lusby, J. Larsen, M. Ehrgott, and D. Ryan, *An exact method for the double TSP with multiple stacks.* International Transactions in Operational Research, 17(5) (2010), 637–652.

[13] M. A. A. Martínez, J.-F. Cordeau, M. Dell'Amico, and M. Iori, *A branch-and-cut algorithm for the double traveling salesman problem with multiple stacks.* INFORMS Journal on Computing, 25(1) (2013), 41–55.

[14] H. L. Petersen, C. Archetti, and M. G. Speranza, *Exact solutions to the double travelling salesman problem with multiple stacks.* Networks, 56(4) (2010), 229–243.

[15] H. L. Petersen and O. B. G. Madsen, *The double travelling salesman problem with multiple stacks - Formulation and heuristic solution approaches.* European Journal of Operational Research, 198(1) (2009), 139–147.

[16] R. Roberti and P. Toth, *Models and algorithms for the asymmetric traveling salesman problem: an experimental comparison.* EURO Journal on Transportation and Logistics, 1(1-2) (2012), 113–133.

[17] S. C. Sarin, H. D. Sherali, and A. Bhootra, *New tighter polynomial length formulations for the asymmetric traveling salesman problem with and without precedence constraints.* Oper. Res. Lett., 33(1) (Jan. 2005), 62–70.

[18] A. Sassano, *On the facial structure of the set covering polytope.* Mathematical Programming, 44(1-3) (1989), 181–202.

[19] S. Toulouse and R. Wolfler Calvo, *On the complexity of the multiple stack tsp, kstsp.* In Proceedings of the 6th Annual Conference on Theory and Applications of Models of Computation, TAMC '09, pp. 360–369, Berlin, Heidelberg, 2009. Springer-Verlag.

Isoperimetric inequalities in Carnot-Carathéodory spaces

Valentina Franceschi (*)

Abstract. One of the most ancient mathematical problems is Dido's problem, appearing in Virgil's *Aeneid*: what is the shape to give to a rope in order to enclose a maximal region of land? The expected solution is of course the circle. Despite the ancient origins, a rigorous mathematical formulation and solution is quite recent, dating back to the 1950s when Caccioppoli and De Giorgi introduced the notion of perimeter in the n-dimensional Euclidean space. The latter notion led to the study of isoperimetric inequalities and to the solution of Dido's problem generalized to n dimensions. Mathematicians then generalized isoperimetric inequalities to different frameworks, such as riemannian manifolds and metric spaces.
After an overview of the classical definitions, in this article we present isoperimetric inequalities in a class of metric spaces arising from the study of hypoelliptic differential operators, called Carnot-Carathéodory spaces. We conclude presenting the main conjecture in this framework (Pansu's conjecture) ad some related results.

1 Dido's problem

One of the most ancient mathematical problems is the isoperimetric problem. Its first appearance leads back to Virgil's Aeneid:

"Devenere locos ubi nunc ingentia cernis
Moenia sergentemque novae Karthaginis arcem,
mercatique, solum, facti de nomine Byrsam,
taurino quantum possent circumdare tergo."

(Virgil, Aeneid, 1st book, *verses 365–369)*

"They came to this place, and bought
land, where you now see the vast walls,
and resurgent stronghold, of new Carthage,
as much as they could enclose with the
strips of hide from a single bull, and
from that they called it Byrsa."

According to the mythology associated with Virgil's saga, Dido, the queen of Tiro, took refuge in North Africa, after being exiled by her brother Pygmalion. Here, she purchased from a local king the land along the North African coastline that could be enclosed by the hide of a bull. Queen Dido sliced the hide into very thin strips and tied them together

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: valentina.franceschi88@gmail.com . Seminar held on May 4th, 2016.

in order to build a rope. Then, she used the rope to enclose a portion of land facing the coast: according to the legend, this region became the city of Carthago. *Dido's problem* is the following: what shape should I give to the rope in order to enclose maximal area? Dido's answer, which turns out to be the correct one, is the *circumference*.

Figure 1. The city of Carthago and Dido cutting the hide of a bull (Engraving by Mathhäus Merian the Elder, 1630).

Formulated in mathematical terms, Dido's problem is the following:

Among all simple closed planar curves of given length L,
the circle of circumference L encloses maximal area.

The problem can be naturally formulated in the n-dimensional Euclidean space $\mathbb{R}^n$, considering among all surfaces having a prescribed *surface area*, the one that maximizes the *enclosed volume*, thought of as the *n-dimensional Lebesgue measure* $\mathcal{L}^n$ of the region enclosed(†). Needless to say, the expected solution to the isoperimetric problem in $\mathbb{R}^n$ is the $(n-1)$-dimensional sphere.

1.1 The definition of Perimeter

Despite the ancient origins of Dido's problem, the first attempts to rigorously prove the isoperimetric property of the circle are quite recent. This is due to the complexity of *introducing the notion of surface area* in the widest possible class of sets in $\mathbb{R}^n$, namely in the class of Lebesgue measurable sets. The mathematics involved to this purpose represents a starting point for several and very important branches of Mathematics, such as *Geometric Measure Theory* and *Calculus of Variations*. In this survey I will just cite some of the proofs.

The first rigorous solution of the isoperimetric problem in the Euclidean plane dates back to the 1902 paper [17] by Hurwitz, in the class of all closed curves in the plane that are piecewise C^1 smooth. In fact, if $\gamma : [0,1] \to \mathbb{R}^2$ is such a curve, its *length* is defined as

$$\ell(\gamma) = \int_0^1 |\gamma'(t)|\, dt. \tag{1}$$

(†)The definition of the Lebesgue measure dates back to 1902

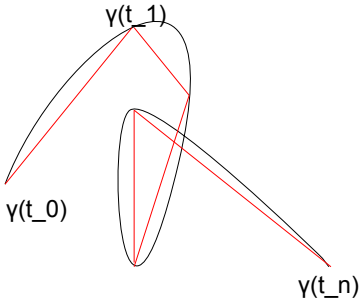
More generally, given a function $f : [a, b] \to \mathbb{R}^2$ its length can be defined as the maximal length of the polygonal curves approximating f: the *total variation* of f in $[a, b]$ is defined as

$$\bigvee_a^b f = \sup \Big\{ \sum_{i=0}^{n-1} |f(t_{i+1}) - f(t_i)|,\ a = t_0 < t_1 < \cdots < t_n = b \Big\} \tag{2}$$

When $\bigvee_a^b f < \infty$ we say that f has bounded total variation, and in this case, we set

$$\ell(f) = \bigvee_a^b f \tag{3}$$

Notice that definition (3) corresponds to definition (1) for a piecewise C^1 curve.



Given a two-dimensional manifold $M \subset \mathbb{R}^3$, one may consider the same idea as in (2) to define its *surface area*. Namely, one can consider the maximal area of the polyhedral domains inscribed in M, where a *polyhedral domain* is a set $\Pi \subset \mathbb{R}^3$ such that $\partial\Pi$ is contained in the union of a finite number of hyperplanes. However, the latter quantity does not describe the correct notion of surface area in $\mathbb{R}^3$, as Schwarz proved in the 1890 paper [28], showing he following example.

Example 1.1 (Schwarz, 1980) Consider the right circular cylinder of radius 1 and height $h > 0$ represented by:

$$(\cos u, \sin u, v), \quad 0 \le u \le 2\pi, 0 \le v \le h.$$

The surface area of its lateral surface, S, should measure $2\pi h$.

Let now be $m, n \in \mathbb{N}$ and divide S into m bands of height h/m and $2n$ congruent slices of width π/n. Consider the inscribed polygonal domain $\Pi(mn)$ defined as the union of the $2mn$ congruent triangular surfaces whose vertices are identified by intersection of the bands with the slices as in Figure 2.

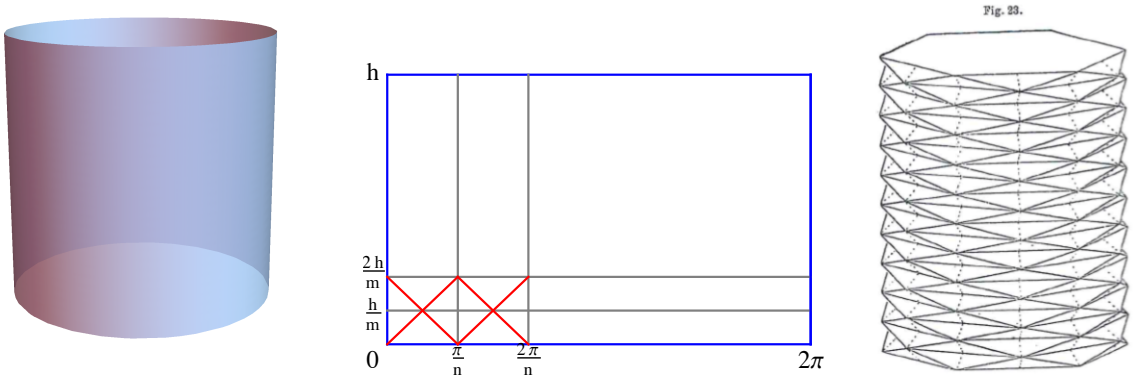
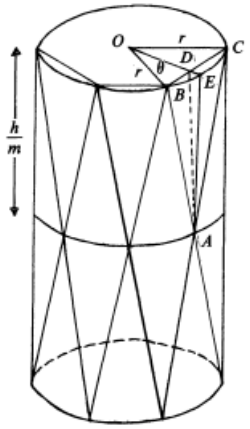


Figure 2. The cylinder and the triangular surfaces inscribed in the cylinder, represented on the right as in the paper [28].



Consider the face ABC. The angle θ is π/n. The edge BC hence measures $2|BD| = 2\sin\frac{\pi}{n}$.
Now, to find $|AD|$ we first calculate $|DE| = 1-\cos\frac{\pi}{n}$, hence $|DA|^2 = \frac{h^2}{m^2} + \big(1-\cos\frac{\pi}{n}\big)^2$. Therefore

$$\text{area}(ABC) = \frac{1}{2}2\sin\frac{\pi}{n}\sqrt{\frac{h^2}{m^2} + \big(1-\cos\frac{\pi}{n}\big)^2}$$

The area surface of the polyhedral domain $\Pi(mn)$ is

$$2n\sin\frac{\pi}{n}\sqrt{h^2 + m^2\big(1-\cos\frac{\pi}{n}\big)^2}.$$

Hence the limit depends on the mutual behavior of m and n going at infinity. For instance, if $m = n$, we obtain the expected number in the limit as $n\to\infty$:

$$2n\sin\frac{\pi}{n}\sqrt{h^2 + n^2\big(1-\cos\frac{\pi}{n}\big)^2} \sim 2\pi\sqrt{h^2 + n^2\big(\frac{\pi^2}{2n^2}\big)^2} = 2\pi h\sqrt{1+\frac{\pi^4}{4n^2h^2}} \sim 2\pi h(1+\frac{\pi^4}{8n^2h^2}) \to 2\pi h.$$

On the other hand, if $m = n^3$, we have

$$2n\sin\frac{\pi}{n}\sqrt{h^2 + n^6\big(1-\cos\frac{\pi}{n}\big)^2} \sim 2\pi\sqrt{h^2 + n^6\big(\frac{\pi^2}{2n^2}\big)^2} \to \infty, \quad n\to\infty.$$

As Schwarz example shows, while dealing with k-dimensional manifolds for $k \geq 2$ we therefore need to think of a different notion of surface area. A first candidate is a notion known from the 1920s. Given $k \geq 0$ and $E \subset \mathbb{R}^n$, the *k-dimensional Hausdorff measure* of E is given by

$$\mathcal{H}^k(E) = \lim_{\delta\to 0^+} \mathcal{H}^k_\delta(E),$$

where

$$\mathcal{H}^k_\delta(E) = \frac{\omega_k}{2^k}\inf\left\{\sum_{i\in I}(\text{diam}(E_i))^k : \text{diam}(E_i) < \delta,\ E \subset \bigcup_{i\in I} E_i\right\}$$

for a finite or countable covering of E, $\{E_i\}_{i\in I}$. Here we used the following notation:

- $\omega_k = \mathcal{L}^k(\{p\in\mathbb{R}^k : |p| = 1\})$;
- for a set $A \subset \mathbb{R}^n$ we call diameter of A the quantity $\text{diam}(A) = \sup_{x,y\in A}|x-y|$, with the convention $\text{diam}(\varnothing) = 0$.

Unfortunately, the k-dimensional Hausdorff measure is not a good notion of surface area of a k-dimensional manifold to approach variational problems. In fact, it fails to be lower semicontinuous with respect to the convergence of sets induced by the Hausdorff distance. Given $A, B \subset \mathbb{R}^n$ closed sets, the *Hausdorff distance* between A and B is

$$d_H(A,B) = \max\Big\{\sup_{x\in A}\text{dist}(x,B), \sup_{y\in b}\text{dist}(y,A)\Big\}$$

where $\operatorname{dist}(x, B) = \inf_{y \in B} |x - y|$. We say that A is the Hausdorff limit of a sequence of sets $(A_n)_{n \in \mathbb{N}}$ if A is closed and $d_H(A, A_n) \to 0$, as $n \to \infty$. The Hausdorff limit is unique because d_H is a distance between closed sets. In the following example we construct a sequence of sets A_h, $h \in \mathbb{N}$ such that

$$\lim_{h \to \infty} d_H(A_h, A) = 0 \quad \text{and} \quad \mathcal{H}^1(A) > \liminf_{h \to \infty} \mathcal{H}^1(A_h),$$

that implies that the Hausdorff measure is not lower semicontinuous with respect to the Hausdorff convergence.

Example 1.2 For any $h \in \mathbb{N}$, consider $[0, 1]$ to be divided into h^2 intervals and define A_h to be the union of the intervals which correspond to the integer multiples of h.

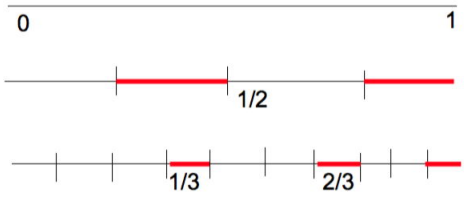


Figure 3. In red, the sets A_1, A_2 and A_3.

We have

$$d_H(A_h, [0,1]) = \frac{n-1}{n^2} \to 0, \quad \mathcal{H}^1(A_h) = \frac{1}{h} \to 0 \quad \text{as } h \to \infty.$$

Hence $A = [0, 1]$ and $\mathcal{H}^1(A) = 1 > 0 = \liminf_{h \to \infty} \mathcal{H}^1(A_h)$.

It is in the 1950s when Caccioppoli and De Giorgi introduced the notion of perimeter of a Lebesgue measurable set in $\mathbb{R}^n$. In 1953, Renato Caccioppoli introduces in [1] the following definition of surface area of a 2-dimensional manifold in $\mathbb{R}^3$.

Definition 1.3 (Caccioppoli, 1953) The *perimeter* of a Lebesgue measurable set $E \subset \mathbb{R}^3$ is

$$\inf \left\{ \liminf_{j \to \infty} P(\Pi_j) : \mathcal{L}^3(\Pi_j \Delta E) \to 0, \ j \to \infty \right\} \tag{4}$$

where Π_j are polygonal domains and $\Pi_j \Delta E$ denotes the symmetric difference $\Pi_j \Delta E = (\Pi_j \setminus E) \cup (E \setminus \Pi_j)$.

Caccioppoli's definition of perimeter, generalized to the n-dimensional space, was then characterized by E. De Giorgi in 1954 in the seminal paper [4]. Here, the author proposes a definition of perimeter, starting from a generalization of Gauss-Green formulas for smooth sets and he proves equivalence with Caccioppoli's definition. As a striking consequence, De Giorgi definition has natural and powerful applications to geometric variational problems, such as the isoperimetric problem.

Definition 1.4 (De Giorgi, 1954) Let $E \subset \mathbb{R}^n$ be a Lebesgue measurable set. The *perimeter of* E is

$$P(E) = \sup\Big\{ \int_E \mathrm{div}\varphi(x)\, dx \ : \ \varphi \in C^1_c(\mathbb{R}^n, \mathbb{R}^n),\ \sup_{\mathbb{R}^n} |\varphi| \le 1 \Big\}. \tag{5}$$

We say that E is a *set of finite perimeter* if $P(E) < \infty$.

Definition (5) can be motivated by the following observation: given $\varphi \in C^1(\mathbb{R}^n)$ and a set $E \subset \mathbb{R}^n$ with C^1-boundary, by the divergence theorem we have

$$\int_E \mathrm{div}\varphi\, dx = \int_{\partial E} \varphi \cdot \nu^E\, d\mathcal{H}^{n-1}$$

where N^E denotes the outer unit normal to the boundary of E and, given $p, q \in \mathbb{R}^n$, $p \cdot q$ denotes the standard scalar product in $\mathbb{R}^n$. Assuming $|\varphi| \le 1$, we hence get

$$\int_E \mathrm{div}\varphi\, dx \le \mathcal{H}^{n-1}(\partial E).$$

1.1.1 De Giorgi definition

De Giorgi, in fact, introduces the definition of perimeter in an equivalent formulation, using the regularity theory for the heat equation. Namely, given a Lebesgue measurable set $E \subset \mathbb{R}^n$, he considers the Cauchy problem for the heat equation

$$\text{(CP)} \qquad \begin{cases} \frac{\partial}{\partial t} u(x,t) - \Delta_x u(x,t) = 0 \quad x \in \mathbb{R}^n,\ t > 0 \\ u(x,0) = \chi_E(x) \end{cases}$$

If u solves (CP), then it is a C^∞-smooth function, and

$$\|u(\cdot,t) - \chi_E\|_{L^1(\mathbb{R}^n)} \to 0, \text{ as } t \to \infty,$$

where $L^1(\mathbb{R}^n)$ is the space of summable functions on $\mathbb{R}^n$: $L^1(\mathbb{R}^n) = \{f : \mathbb{R}^n \to \mathbb{R} : \int |f| < \infty\}$ and $\|f\|_{L^1(\mathbb{R}^n)} = \int |f|$. De Giorgi proves that the function

$$t \mapsto \int_{\mathbb{R}^n} |\nabla_x u(x,t)|\, dx$$

is strictly monotone decreasing and he defines the perimeter of E as

$$I(\chi_E) = \lim_{t \to 0} \int_{\mathbb{R}^n} |\nabla_x u(x,t)|\, dx$$

showing that $I(\chi_E) < \infty$ if and only if there exists a Radon measure μ_E satisfying

$$\int_E \mathrm{div}\varphi\, dx = \int_{\mathbb{R}}^n \varphi \cdot d\mu_E.$$

In this case

$$|\mu_E| = I(\chi_E) = P(E).$$

Moreover, if we consider the gradient of the characteristic function χ_E ($\chi_E(x) = 1$ if $x \in E$, 0 otherwise) in the sense of distributions, we obtain

$$\int_E \mathrm{div}\varphi \, dx = \int_{\mathbb{R}^n} \chi_E \mathrm{div}\varphi \, dx = -\int_{\mathbb{R}^n} \varphi \cdot \nabla\chi_E \, dx$$

hence the perimeter defined in (5) corresponds to the total variation of the Radon measure $\nabla\chi_E$, see Miranda [18]. In conclusion, in addition to the equivalence of De Giorgi and Caccioppoli definition of perimeter, we have

$$P(E) = I(\chi_E) = |\nabla\chi_E|(\mathbb{R}^n).$$

1.2 Isoperimetric inequality in $\mathbb{R}^n$

The notion of perimeter introduced by Caccioppoli and De Giorgi allows to prove the following isoperimetric inequality: for any Lebesgue measurable set $E \subset \mathbb{R}^n$ with finite measure we have

$$P(E) \geq n\omega_n^{\frac{1}{n}} \mathcal{L}^n(E)^{\frac{n-1}{n}}. \tag{6}$$

Equality case occurs if and only if E is a Euclidean ball. It is easy to check one implication: if $E = B(p, r)$, $p \in \mathbb{R}^n$, $r > 0$, we have: $P(E) = n\omega_n r^{n-1}$, $\mathcal{L}^n(E)^{\frac{n-1}{n}} = \omega_n^{\frac{n-1}{n}} r^{n-1}$, and equality occurs. The isoperimetric inequality (6) is equivalent to say that that the ball has the least perimeter among all sets with the same measure, and it is the unique set having this property. The *isoperimetric problem*

$$\inf\{P(E) : \mathcal{L}^n(E) = v\}, \quad v > 0$$

is hence solved.

The proof of (6) consists of proving existence and uniqueness of isoperimetric sets.

(a) The proof of *existence* of isoperimetric sets is based on the two following properties of the perimeter.

- *Lower semicontinuity:* Let $E_h \subset \mathbb{R}^n$, $h \in \mathbb{N}$ be a sequence of sets with finite perimeter and let $E \subset \mathbb{R}^n$ be such that $\mathcal{L}^n(E \triangle E_h) \to 0$ as $h \to \infty$. Then

$$P(E) \leq \liminf_{h\to\infty} P(E_h).$$

- *Compactness:* Let $E_h \subset \mathbb{R}^n$, $h \in \mathbb{N}$ be a sequence of sets with finite perimeter such that

$$\sup_{h\in\mathbb{N}} P(E_h) < \infty.$$

  Then there exists a set of finite perimeter $E \subset \mathbb{R}^n$ such that $\mathcal{L}^n(E \triangle E_h) \to 0$ as $h \to \infty$.

(b) To prove that the *unique* isoperimetric set is the ball, classical symmetrization techniques, known as Steiner and Schwarz rearrangements are used.

2 Carnot-Carathéodory spaces

The aim of these section is to present the theory of perimeters in a different framework.

In the last two decades an intense investigation on *Analysis and Geometry in Metric Spaces* has been carried out by many authors and led to a generalization of classical theories to these structures: Sobolev spaces (see for instance Hajłasz and Koskela [16], quasiconformal mappings, functions of bounded variations and sets of finite perimeter, currents and rectifiable sets. A very general framework to study isoperimetric inequalities is therefore established. An important class of Metric Spaces is given by *Carnot-Carathéodory spaces*, whose definition is attributed to Gromov (see [14], and [15] for the english version).

2.1 Definition of Carnot-Carathéodory spaces

Let $X = \{X_1, \dots, X_r\}$ a family of vector fields on $\mathbb{R}^n$ with $r \le n$. Associated to the family $X = \{X_1, \dots, X_r\}$, we define a sub-bundle of the tangent bundle:

$$\Delta^X = \bigcup_{p \in \mathbb{R}^n} \Delta^X_p \qquad \Delta^X_p = \operatorname{span}\{X_1(p), \dots, X_r(p)\}$$

and we call it the *horizontal bundle*. A vector field $Y \in \Delta^X$ is called a *horizontal vector field*. We say that an absolutely continuous curve $\gamma : [0,1] \to \mathbb{R}^n$ is *horizontal* if $\dot\gamma(t) \in \Delta^X_{\gamma(t)}$ for every $t \in [0,1]$. Given, for every $p \in \mathbb{R}^n$, a scalar product g_p on $\mathbb{R}^n$ such that $X_1, \dots, X_r$ are orthonormal, we define the *length* of an horizontal curve γ as

$$\ell_X(\gamma) = \int_0^1 g_{\gamma(t)}(\dot\gamma(t), \dot\gamma(t))\, dt = \int_0^1 \sqrt{\sum_{i=1}^r a_i^2(\gamma(t))}\, dt \tag{7}$$

where $\dot\gamma(t) = \sum_{i=1}^r a_i(\gamma(t)) X_i(\gamma(t)) \in \Delta^X_{\gamma(t)}$. When any two points $p, q \in \mathbb{R}^n$ can be connected by means of horizontal curves, the *Carnot-Carathéodory* (also *sub-Riemannian* or *CC* for short) *distance* associated to X is given by

$$d^X_{cc}(p,q) = \inf\Big\{\ell_X(\gamma) : \gamma \text{ horizontal}, \gamma(0) = p,\ \gamma(1) = q\Big\}. \tag{8}$$

Manifolds endowed with a family of vector fields for which such a distance can be constructed, are known as *Carnot-Carathéodory spaces*.

Example 2.1 (Euclidean distance) The euclidean space $\mathbb{R}^n$ endowed with the family $X = \{\partial_{x_1}, \dots, \partial_{x_n}\}$ is a Carnot-Carathéodory structure and the CC distance is the euclidean one $d_E(p,q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$. In this case we use the notation $B_E(x,r) = B^X_{cc}(x,r)$ for a euclidean ball of center $x \in \mathbb{R}^n$ and radius $r > 0$.

2.1.1 Hörmander condition

Before a formal definition of Carnot-Carathéodory spaces was given, a sufficient condition to connect any two points by means of horizontal curves was proved independently by

Chow and Rashewsky, involving the rank of the Lie algebra generated by $X_1, \dots, X_r$. The same condition has a key role for the hypoellipticity of the operator

$$\mathcal{L} = \sum_{i=1}^{r} X_i^2,$$

known as *subelliptic Laplacian*, proved by Hörmander in 1967, and it hence takes the name of *Hörmander condition* or *rank condition*. A differential operator $\mathcal{L}$ is said to be *hypoelliptic* if the weak solutions to $\mathcal{L}u = f$ with $f \in C^\infty(\mathbb{R}^n)$ are $C^\infty(\mathbb{R}^n)$.

In this section we introduce Hörmander condition. We first recall some definitions.

Definition 2.2 (Lie Algebra) A real *Lie Algebra* is a real vector space V endowed with an operation

$$\{\ ,\ \} : V \times V \to V, \qquad (v, w) \mapsto \{v, w\}$$

which satisfies the following properties

(a) $\{\ ,\ \}$ is $\mathbb{R}$-bilinear;

(b) $\{v, w\} = -\{w, v\}$ for any $v, w \in V$ (*skew symmetry*);

(c) the following identity, called the *Jacobi identity* holds:

$$\{u, \{v, w\}\} + \{v, \{w, u\}\} + \{w, \{u, v\}\} = 0 \qquad u, v, w \in V.$$

We say that a vector space $W \subset V$ is a *Lie subalgebra* of V if it is closed under the operation $\{\ ,\ \}$ and properties 1-3 are satisfied. Given a subset $A \subset V$, we call *Lie subalgebra generated by A (in $(V, \{\ ,\ \})$)* the smallest Lie subalgebra of V containing A and we denote it by $\mathrm{Lie}(A)$.

There is a classical way to associate to a family of vector fields on $\mathbb{R}^n$ a Lie algebra. Given two smooth vector fields on $\mathbb{R}^n$

$$X = \sum_{i=1}^{n} a_i \partial_{x_i}, \qquad Y = \sum_{j=1}^{n} b_j \partial_{x_j},$$

we define their composition law as the composition of partial differential operators, which is denoted by $\circ$, namely:

$$X \circ Y = \sum_{i=1}^{n} \left(a_i (\partial_i b_j) \partial_{x_j} + a_i b_j \partial^2_{x_i x_j} \right)$$

where $\partial^2_{x_i x_j}$ is the second order derivative with respect to x_i and x_j. The *commutator* $[X, Y]$ between X and Y is defined as

$$[X, Y] = X \circ Y - Y \circ X.$$

The set of C^∞ vector fields on $\mathbb{R}^n$, $\mathfrak{X}(\mathbb{R}^n) = \{X = \sum_{i=1}^n a_i \partial_{x_i} : a_i \in C^\infty(\mathbb{R}^n)\}$, endowed with the bracket operation $[\ ,\ ]$ is a Lie-algebra. In particular the commutator of vector fields is

again a vector field. Henceforth, given a family of smooth vector fields $X = \{X_1, \dots, X_r\}$, we consider the Lie algebra generated by X in $(\mathfrak{X}, [\ ,\ ])$, and we denote it by $\mathrm{Lie}(X)$. It is easy to see that it coincides with the real span of the iterated brackets of the elements of X, namely:

$$\mathrm{Lie}(X) = \mathrm{span}\{[X_i, [...[X_j, X_k]]] : i, j, k = 1, \dots, r\}. \tag{9}$$

Definition 2.3 (Hörmander vector fields) We say that the smooth vector fields on $\mathbb{R}^n$ $X_1, \dots, X_r \in \mathfrak{X}(\mathbb{R}^n)$ satisfy the *Hörmander condition* if the Lie algebra that they generate has full rank on $\mathbb{R}^n$, namely if

$$\mathrm{rank}(\mathrm{Lie}(X_1, \dots, X_r))(x) = n \qquad x \in \mathbb{R}^n, \tag{10}$$

where $\mathrm{rank}(W)$ denotes the dimension of W as vector space.

Theorem 2.4 (Rashevsky 1938, Chow 1939) *Let $p, q \in \mathbb{R}^n$. If $X = \{X_1, \dots, X_r\}$ is a family of vector fields satisfying the Hörmander condition, there exists an absolutely continuous curve $\gamma : [0,1] \to \mathbb{R}^n$ such that*

$$\gamma(0) = p,\ \gamma(1) = q \qquad \dot\gamma(t) = \sum_{i=1}^{r} a_i(t) X_i(\gamma(t)) \text{ for some coefficients } a_i \text{ for a.e. } t \in [0,1].$$

Remark 2.5 Carnot-Carathéodory distances where already present in the literature of hypoelliptic operators. In the work by Fefferman and Phong [7], dated 1981, the study of *subelliptic operators* which are not assumed to be written as sum of squares is accomplished associating them with a suitable metric d. This idea gives an impulse to the study of *degenerate elliptic operators*, via associated Carnot-Carathéodory metrics, see Franchi and Lanconelli [11], and Sobolev and Poincaré inequalities are studied in view of a regularity theory for weak solutions and estimates of the fundamental solution. Isoperimetric inequalities follow as a result of this research branch.

2.2 Examples

2.2.1 Carnot Groups

We say that $\mathbb{G}$ is a *Lie group* if it is a smooth manifold endowed with a group operation $*$ such that the composition map $(x, y) \mapsto x * y$ and the inverse map $x \mapsto x^{-1}$ ($x * x^{-1} = x^{-1} * x = e$, unit element) are smooth on $\mathbb{G}$. Fixed $x \in \mathbb{G}$ we call *left translation by x* the map

$$\tau_x : \mathbb{G} \to \mathbb{G},\ \tau_x(y) = x * y$$

and *right translation by x* the map

$$\varrho_x : \mathbb{G} \to \mathbb{G},\ \varrho_x(y) = y * x.$$

The maps τ_x, ϱ_x are clearly C^∞ diffeomorphisms of $\mathbb{G}$ into itself for any $x \in \mathbb{G}$. We say that a vector field X on $\mathbb{G}$ is *left-invariant* if the following holds

$$(Xf) \circ \tau_x = X(f \circ \tau_x) \quad \text{for every } f \in C^\infty(\mathbb{G}),\ x \in \mathbb{G}. \tag{11}$$

The set of all left invariant vector fields is a Lie algebra, which is called the *Lie algebra of* $\mathbb{G}$ and it is denoted by $\mathrm{Lie}(\mathbb{G})$ or $\mathfrak{g}$.

Definition 2.6 (Carnot Group) A *Carnot group* of *step* s is a connected, simply connected Lie group whose Lie algebra $\mathfrak{g}$ admits a step s stratification, i.e., there exist linear subspaces $V_1, \dots, V_s$ such that

$$\mathfrak{g} = V_1 \oplus \cdots \oplus V_s,\ [V_1, V_i] = V_{i+1},\ V_s \neq \{0\},$$

where $[V_1, V_i]$ is the subspace of $\mathfrak{g}$ generated by the commutators $[X, Y]$ with $X \in V_1$ and $Y \in V_i$.

We call the *homogeneous dimension* of $\mathbb{G}$ the number

$$Q = \sum_{i=1}^{s} i \dim V_i, \tag{12}$$

and the *rank* of $\mathbb{G}$, denoted by r, the dimension of V_1, which is the number of Lie-generators of the algebra.

Remark 2.7 Any n-dimensional Carnot group can be identified with $\mathbb{R}^n$.

The Lie algebra $\mathfrak{g}$ of a Carnot group $\mathbb{G}$ is naturally endowed with a family of *dilations* modeled on its stratification:

$$\delta_\lambda^{\mathfrak{g}}\Big(\sum_{i=1}^{s} Y_i\Big) = \sum_{i=1}^{s} \lambda^i Y_i, \quad Y_i \in V_i, \quad \lambda > 0.$$

The group $\mathbb{G}$ inherits a family of *anisotropic dilations* parametrized by $\lambda > 0$ and defined as

$$\delta_\lambda^{\mathbb{G}}(x) = \delta_\lambda^{\mathbb{G}}\Big(\mathrm{Exp}\Big(\sum_{i=1}^{n} Y_i\Big)\Big) = \mathrm{Exp}\Big(\sum_{i=1}^{s} \lambda^i Y_i\Big).$$

The dilation on the group $\delta_\lambda^{\mathbb{G}}$ turns out to be of the following form

$$\delta_\lambda^{\mathbb{G}}(x_1, \dots, x_n) = (\lambda x_1, \dots, \lambda x_r, \lambda^{\sigma_{r+1}} x_{r+1}, \dots, \lambda^{\sigma_n} x_n) :$$

with $\sigma_j = i$ if $Y_j \in V_i$, $j = r+1, \dots, n$, $i = 2, \dots, r$. Moreover, the family $\delta_\lambda^{\mathbb{G}}$ is a family of automorphisms of $\mathbb{G}$, namely

$$\delta_\lambda^{\mathbb{G}} x * \delta_\lambda^{\mathbb{G}} y = \delta_\lambda^{\mathbb{G}}(x * y);$$

and $(\delta_\lambda^{\mathbb{G}})^{-1} = \delta_{1/\lambda}^{\mathbb{G}}$.

A *sub-Riemannian structure on* $\mathbb{G}$ is given considering the first layer V_1 of the stratification of the Lie algebra as the horizontal bundle. Consider a basis for the Lie algebra $\mathfrak{g} = V_1 \oplus \cdots \oplus V_s$,

$$X_1, \dots, X_r, X_1^{(2)}, \dots, X_{r_2}^{(2)}, \dots, X_1^{(s)}, \dots, X_{r_s}^{(s)}$$

where $X_1, \dots X_r$ generates V_1, $X_1^{(j)}, \dots, X_{r_j}^{(j)}$ generates V_j for $j = 2, \dots, s$, $r_1 + \cdots + r_s = n$ and such that at the origin it is the canonical orthonormal basis of $\mathbb{R}^n$ in the coordinate system

$$x = (x_1, \dots, x_{r_1}, x_{r_1+1}, \dots, x_{r_1+r_2}, \dots, x_{r_1+\cdots+r_{s-1}+1}, \dots, x_n).$$

Namely,

$$X_1^{(j)}(0) = \frac{\partial}{\partial x_{r_1+\cdots+r_{j-1}+1}}, \ \dots \ , X_{r_i}^{(j)}(0) = \frac{\partial}{\partial x_{r_1+\cdots+r_j}}, \quad j = 1, \dots s.$$

We refer to $X_1, \dots, X_r$ as the family of *canonically generating vector fields* and we use the notation

$$X_{\mathbb{G}} = \{X_1, \dots, X_r\}.$$

We call the *Carnot-Carathéodory distance of the Carnot group* $\mathbb{G}$, and denote it by $d_{cc}^{\mathbb{G}}$, the one defined in (8) and associated to a family of canonically generating vector fields:

$$d_{cc}^{\mathbb{G}}(p, q) = \inf \Big\{ \int_0^1 \sqrt{\sum_{i=1}^{s} a_i(\gamma(t))^2}\, dt : \gamma(0) = p, \ \gamma(1) = q, \ \dot\gamma = \sum_{i=1}^{s} a_i X_i \Big\}.$$

We use the notation $B_{\mathbb{G}} = B_{cc}^X$ where X is a family of canonical generators for $\mathbb{G}$. The following properties of $d_{cc}^{\mathbb{G}}$ hold:

- The topology induced on $\mathbb{G}$ by $d_{cc}^{\mathbb{G}}$ is the topology of the manifold;
- $d_{cc}^{\mathbb{G}}$ is *left invariant*:
$$d_{cc}^{\mathbb{G}}(\tau_x y, \tau_x z) = d_{cc}^{\mathbb{G}}(y, z);$$
- $d_{cc}^{\mathbb{G}}$ is 1-*homogeneous with respect to intrinsic dilations*
$$d_{cc}^{\mathbb{G}}(\delta_\lambda^{\mathbb{G}} x, \delta_\lambda^{\mathbb{G}} y) = \lambda d_{cc}^{\mathbb{G}}(x, y), \qquad x, y, z \in \mathbb{G} \ \lambda > 0.$$

2.2.2 Heisenberg groups

The *n-dimensional Heisenberg group*, denoted by $\mathbb{H}^n$, is $\mathbb{C}^n \times \mathbb{R}$ endowed with the following group operation:

$$(z, t) * (z', t') = (z + z', t + t' + 2\mathrm{Im}(z\bar{z'})),$$

where $\bar{z'}$ denotes the conjugate of z'. Identifying $\mathbb{C}^n$ with $\mathbb{R}^{2n}$ through $z = x + iy \mapsto (x, y) = (x_1, \dots, x_n, y_1, \dots, y_n)$, the operation can be also written as

$$(x, y, t) * (x', y', t') = \Big(x + x', y + y', t + t' + 2\sum_{i=1}^{n} \big(x_i' y_i - x_i y_i'\big)\Big). \tag{13}$$

To find a family of canonically generating vector fields of the Lie algebra $\mathfrak{h}$ of $\mathbb{H}^n$, we look for a family of left invariant vector fields $X = \{X_1, \dots, X_n, Y_1, \dots, Y_n, T\}$ which correspond to the canonical basis of $\mathbb{R}^{2n+1}$ at the origin

$$X_i(0) = \partial_{x_i}, \qquad Y_i(0) = \partial_{y_i}, \qquad T(0) = \partial_t, \ i = 1, \dots, n.$$

This leads to

$$X_i(x,y,t) = \partial_{x_i} + 2y_i\partial_t, \quad Y_i(x,y,t) = \partial_{y_i} - 2x_i\partial_t, \quad T(0) = \partial_t,\ i = 1,\dots,n.$$

Notice that the only nonzero commutator of the family X is $[X_i, Y_i] = -4\partial_t = -4T$. Brackets of order bigger than 2 are zero. Hence, $\mathfrak{h} = \mathrm{Lie(X)}$ with $X = \{X_i, Y_i : i = 1,\dots,n\}$ and the family X satisfies the Hörmander condition (10): $\mathrm{rank(Lie}(X)) = 2n+1$. The horizontal bundle is therefore given by $\Delta = \mathrm{span}\{X_i, Y_i : i = 1,\dots,n\}$. Moreover the Lie algebra $\mathfrak{h}$ admits the stratification

$$\mathfrak{h} = \Delta \oplus [\Delta,\Delta], \quad \Delta = \mathrm{span}\{X_i, Y_i : i = 1,\dots n\},$$

so that $\mathbb{H}^n$ is a Carnot group of step 2 and rank $2n$. The homogeneous dimension of $\mathbb{H}^n$ is

$$Q = 2n+2$$

and the dilations of the group are

$$\delta^{\mathbb{H}}_\lambda(z,t) = (\lambda z, \lambda^2 t), \quad \lambda > 0,\ (z,t) \in \mathbb{H}^n.$$

***Derivations of the Heisenberg group.* 1.** While talking about sub-Riemannian structures, it is often said that the Heisenberg group is the *"easiest" example*, apart from the euclidean space. In fact, we can view the Heisenberg Lie algebra $\mathfrak{h}$ as the unique three dimensional nilpotent Lie algebra, with a step 2 stratification $\mathfrak{h} = V_1 \oplus V_2$, and rank 2 such that

$$[V_1, V_1] = V_2,\ [V_1, V_2] = \{0\}$$

In particular, if $V_1 = \mathrm{span}\{e_1, e_2\}$, $V_2 = \mathrm{span}\{\epsilon\}$ it is sufficient to impose

$$[e_1, e_2] = \epsilon. \tag{14}$$

The group law of the corresponding Carnot group is induced by relation (14).

2. In [29, Chapter XII] the Heisenberg group is introduced from *harmonic analysis*. We recall here the main steps of his argument. Consider the *Siegel domain*

$$\mathcal{U} = \{(\zeta, w) \in \mathbb{C}^2 : \mathrm{Im}(w) > |\zeta|^2\}$$

and associate to each $(z,t) \in \mathbb{H}^1$ the mapping

$$(\zeta, w) \mapsto (\zeta + z, w + t + 2i\zeta * \overline{z} + i|z|^2) = \mathcal{T}_{z,t}(\zeta, w). \tag{15}$$

The transformation (15) maps $\mathcal{U}$ into itself, preserves the boundary $\partial\mathcal{U}$, and defines an action of $(\mathbb{H}^1, *)$ into the space $\mathcal{U}$, i.e., $\mathcal{T}_{z,t} \circ \mathcal{T}_{z't'} = \mathcal{T}_{(z,t)*(z',t')}$. Therefore, the set of transformations $\mathcal{T}_{z,t}$ endowed with the composition of maps, $\circ$, is a group of affine holomorphic bijections of $\mathcal{U}$: $\mathbb{H}^1$ is identified with the *group of translations of the Siegel domain* (i.e., $\mathcal{U}$ is invariant under these transformations). On the other hand, the action of $\mathbb{H}^1$ on the

origin $0 \in \mathcal{U}$, $\mathcal{T}_{z,t}(0,0) = (z, t + i|z|^2)$, identifies $\mathbb{H}^1$ with the *boundary of the Siegel domain* $\partial\mathcal{U}$.

3. In the end, we recall that the horizontal bundle of the Heisenberg group defines a *contact structure* on $\mathbb{R}^3$ as follows. The one-form $\theta = dt + xdy - ydx$ is a *contact form* on $\mathbb{R}^3$, i.e., $\theta \wedge d\theta \neq 0$. Given a contact form ω on a three-dimensional manifold M, the *contact structure* induced by ω on M is $\ker\omega = \{X \in TM : \omega(X) = 0\}$, which turns out to be a two dimensional subbundle of the tangent bundle. This construction allows to consider a map, $J : \ker\omega \to \ker\omega$, $J^2 = -I$ called *complex structure* which represents the starting point of contact geometry. The contact structure induced by θ is the horizontal bundle of the Heisenberg group and the complex structure is defined by $JX = Y$, $JY = -X$.

2.2.3 Grushin spaces

Let $\mathbb{R}^n = \mathbb{R}^h \times \mathbb{R}^k$, where $h, k \geq 1$ are integers and $n = h + k$. Let $\alpha \geq 0$ be a real number. A *Grushin space* is $\mathbb{R}^n$ endowed with the following structure on $\mathbb{R}^n$: $X_\alpha = \{X_1, \dots, X_h, Y_1, \dots, Y_k\}$,

$$\begin{aligned} X_i &= \partial_{x_i}, \quad i = 1, \dots, h, \\ Y_j &= |x|^\alpha \partial_{y_j}, \quad j = 1, \dots, k, \end{aligned} \tag{16}$$

where $|x|$ is the standard norm of $x \in \mathbb{R}^h$. When $h = k = 1$, $\mathbb{R}^2$ endowed with the family X_α is called the *Grushin plane* and it has been considered by Franchi and Lanconelli in [11] to prove Hölder regularity of the weak solutions of $Lu = 0$,

$$L = \frac{\partial^2}{\partial x^2} + |x_1|^{2\alpha} \frac{\partial^2}{\partial x_2^2},$$

using Moser's technique. The differential operator L is known in the literature as the *Grushin operator*, and it is hypoelliptic for $\alpha \in \mathbb{N}$.

We show a formula for the length of horizontal cuves in Grushin structures. For $(x, y) \in \mathbb{R}^n = \mathbb{R}^h \times \mathbb{R}^k$, $x \neq 0$, and $\alpha > 0$ consider the metric

$$ds_\alpha^2 = dx_1^2 + \cdots + dx_h^2 + \frac{1}{|x|^{2\alpha}}(dy_1^2 + \cdots + dy_k^2) \tag{17}$$

where dx_i, dy_j denote the elements of the canonical basis of the cotangent bundle to $\mathbb{R}^n$ in the coordinate system $(x_1, \dots, x_h, y_1, \dots, y_k)$. Then ds_α^2 makes $X_1, \dots, X_h, Y_1, \dots, Y_k$ orthonormal. Following (7), we define the α-length of an horizontal curve $\gamma : [0,1] \to \mathbb{R}^n = \mathbb{R}^h \times \mathbb{R}^k$ as

$$\ell_\alpha(\gamma) = \int_0^1 \sqrt{\sum_{i=1}^h \gamma_i'(t)^2 + \frac{1}{|(\gamma_1(t), \dots, \gamma_h(t))|^{2\alpha}} \sum_{j=1}^k \gamma_{1+j}'(t)^2}\, dt. \tag{18}$$

The Carnot-Carathéodory distance on $\mathbb{R}^n$ associated to the family X is denoted by d_α. The Grushin space $\mathbb{R}^n = \mathbb{R}^h \times \mathbb{R}^k$, with d_α can be endowed with a family of non-isotropic dilations parametrized by $\lambda > 0$

$$\delta_\lambda^\alpha(x, y) = (\lambda x, \lambda^{\alpha+1} y), \quad (x, y) \in \mathbb{R}_x^h \times \mathbb{R}_y^k = \mathbb{R}^n \tag{19}$$

such that $d_\alpha(\delta_\lambda^\alpha p, \delta_\lambda^\alpha q) = \lambda d_\alpha(p,q)$, for $p, q \in \mathbb{R}^n$. We define the *homogeneous dimension of the Grushin space* $\mathbb{R}^n = \mathbb{R}^h \times \mathbb{R}^k$ *with* d_α as

$$Q = h + (\alpha+1)k. \tag{20}$$

3 Perimeter in Carnot-Carathéodory spaces

The perimeter associated with a family of vector fields is defined following the De Giorgi definition of perimeter as follows.

Definition 3.1 [X-perimeter] For any $E \subset \mathbb{R}^n$ the *X-perimeter* of E is defined as

$$P_X(E) = \sup\Big\{ \int_E \operatorname{div}_X \varphi(x)\, dx : \varphi \in \mathcal{F}_r(\mathbb{R}^n)\Big\}, \tag{21}$$

where

$$\mathcal{F}_m(\mathbb{R}^n) = \Big\{\varphi \in C_c^1(\mathbb{R}^n;\mathbb{R}^m) : \max_{x\in\mathbb{R}^n} |\varphi(x)| = \max_{x\in\mathbb{R}^n} \sqrt{\sum_{j=1}^r \varphi_j^2(x)} \le 1\Big\}.$$

We say that E is a set of finite X-perimeter if $P_X(E) < \infty$.

Proposition 3.2 (Representation formula for α-perimeter) *Let $E \subset \mathbb{R}^n$ be bounded open set with Lipschitz boundary and N^E denote the outer unit normal to ∂E. Then*

$$P_\alpha(E) = \int_{\partial E} \sqrt{|N_x^E|^2 + |x|^{2\alpha}|N_y^E|^2}\, d\mathcal{H}^{n-1}.$$

3.1 Perimeter and Length

In the euclidean setting $(\mathbb{R}^2, d_E)$, the perimeter of a smooth set and the length of its boundary as a curve coincide. In a Carnot-Carathéodory structure there is no connection in general between the length of smooth curves and perimeter. We show it with the next example in the case of the Grushin plane $(\mathbb{R}^2, d_\alpha)$.

Example 3.3 ($P_\alpha(E) \neq \ell_\alpha(\partial E)$) We recall that, in the case of the Grushin plane $(\mathbb{R}^2, d_\alpha)$, the horizontal bundle is given by $\Delta = \operatorname{span}\{\partial_x, |x|^\alpha \partial_y\}$ where (x,y) denotes a point in $\mathbb{R}^2$. The metric

$$ds^2 = dx^2 + \frac{1}{|x|^{2\alpha}} dy^2,$$

defined for $x \neq 0$, is such that $ds^2(\partial_x, |x|^\alpha\partial_y) = 0$, $ds^2(\partial_x, \partial_x) = 1$, $ds^2(|x|^\alpha\partial_y, |x|^\alpha\partial_y) = 1$. Hence the length of a curve $\gamma = (\gamma_1, \gamma_2)$ parametrized on $[0,1]$, is defined as

$$\ell_\alpha(\gamma) = \int_0^1 \sqrt{\gamma_1'(t)^2 + \frac{\gamma_2'(t)^2}{\gamma_1(t)^{2\alpha}}}\, dt.$$

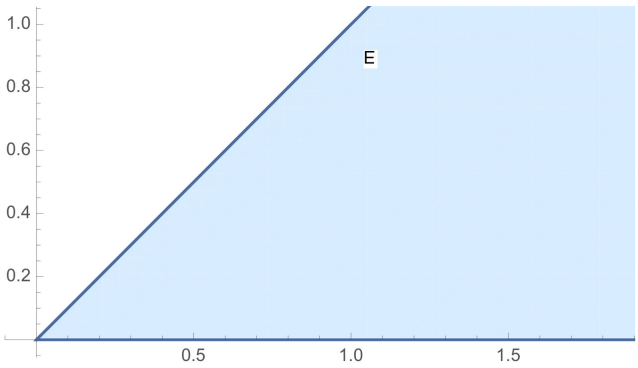


Figure 4. The curve γ and a set E having γ as a part of its boundary.

Let $\gamma : [0,1] \to \mathbb{R}^2$, $\gamma(t) = (t,t)$ and suppose $\gamma^* = \gamma([0,1]) \subset \partial E$ where $E \subset \{(x,y) \in \mathbb{R}^2 : x > y\}$ is a smooth set with finite α-perimeter. We have $\gamma'(t) = (1,1)$ and the outer unit normal to E is $N^E = (-1/\sqrt{2}, 1/\sqrt{2})$ at any point in γ^*. We have

$$\ell_\alpha(\gamma) = \int_0^1 \sqrt{1 + \frac{1}{t^{2\alpha}}}\, dt = \int_0^1 \frac{1}{t^\alpha}\sqrt{t^{2\alpha}+1}\, dt$$

Using the representation formula for the α-perimeter of a smooth set (see Proposition 3.2), we have for $\alpha > 0$,

$$\begin{aligned} P_\alpha(E; \{x \le y,\ 0 < x < 1\}) &= \int_{\partial E \cap \{x=y,\ 0<x<1\}} \sqrt{(N_1^E)^2 + |x|^{2\alpha}(N_2^E)^2}\, d\mathcal{H}^1 \\ &= \int_{\gamma^*} \sqrt{\frac{1}{2} + \frac{|x|^{2\alpha}}{2}}\, d\mathcal{H}^1 = \int_0^1 \sqrt{\frac{1}{2} + \frac{t^{2\alpha}}{2}}\sqrt{2}\, dt \\ &= \int_0^1 \sqrt{1+t^{2\alpha}}\, dt < \int_0^1 \frac{1}{t^\alpha}\sqrt{1+t^{2\alpha}} = \ell_\alpha(\gamma). \end{aligned}$$

Notice that when $\alpha = 0$ we find $P_\alpha(E) = \ell_\alpha(E)$: in this case, in fact, $\ell_\alpha = \ell_E$ and $P_\alpha = P$.

4 Sharp Isoperimetric Inequalities in Carnot-Carathéodory spaces

We finally pass to isoperimetric inequalities in Carnot-Carathéodory spaces. The *isoperimetric problem for the X-perimeter and the Lebesgue measure* is, given $v > 0$, the following minimization problem:

$$\inf\{P_X(E) : \mathcal{L}^n(E) = v\}.$$

Solutions are called *isoperimetric sets*.

As we previously noticed, the isoperimetric inequality (6) in $\mathbb{R}^n$ implies that the unique isoperimetric sets are euclidean balls. This is due to the fact that the constant $n\omega_n^{\frac{1}{n}}$ appearing in the isoperimetric inequality is *sharp*, in the sense that it is the smallest possible positive constant that can be plugged in the inequality and it characterizes equality case.

4.1 Non-sharp isoperimetric inequality

In the framework of Carnot-Caratheódory spaces, what is known, under suitable assumptions on the family X, is an *isoperimetric inequality in a non-sharp form* (see inequality (22) as it is presented in the following Proposition. The first proof of the isoperimetric inequality in Carnot-Carathéodory spaces is due to Pansu that proves it in [23] in the first Heisenberg group $\mathbb{H}^1$. Garofalo and Nhieu proved it in [13] for a family of vector fields with locally Lipschitz coefficients. Franchi, Gallot and Wheeden in [10] proved the isoperimetric inequality for the X-perimeter associated with a family of vector fields that includes Grushin spaces.

Proposition 4.1 *Let Q be the homogeneous dimension associated to the family X (under suitable assumptions on the family, for instance: of Grushin type - see (20), satisfying Hörmander condition - see (12). Then there exists a constant $C > 0$ such that*

$$\mathcal{L}^n(E) \leq C P_X(E)^{\frac{Q}{Q-1}} \tag{22}$$

for every set $E \subset \mathbb{R}^n$ with finite X-perimeter and finite Lebesgue measure.

Notice that the isoperimetric inequality in a non-sharp form does not imply a characterization of isoperimetric sets. Finding the best constants in (22), i.e., the smallest possible constant $C > 0$ that can be plugged in (22), is equivalent to characterize isoperimetric sets.

4.2 Pansu's conjecture

The only sub-Riemannian spaces where the isoperimetric problem has been solved are some types of Grushin structures. The first result is in the Grushin plane: in [21, Theorem 1.1], the authors prove existence of solutions to

$$\min\{P_\alpha(E) : E \subset \mathbb{R}^2, \mathcal{L}^2(E) = v\}, \text{ for } v > 0 \text{ fixed},$$

and they characterize them. Namely, minimizers are unique up to vertical translations and they are obtained through a dilation δ_λ^α of the following set

$$E^\alpha_{isop} = \Big\{(x,y) \in \mathbb{R}^2 : |y| < \varphi_\alpha(|x|) = \int_{\arcsin|x|}^{\frac{\pi}{2}} \sin^{\alpha+1}(t)\, dt,\ |x| < 1\Big\}. \tag{23}$$

In [8] we generalize this result to Grushin structures on $\mathbb{R}^n = \mathbb{R}^h \times \mathbb{R}^k$ for $k = 1$ and to H-type groups.

There is a famous conjecture about the shape of isoperimetric sets in the Heisenberg groups, which was formulated by Pansu in 1982 in $\mathbb{H}^1$, see [23], [24]. *Pansu's conjecture* is the following: up to a null set, a left translation, and a dilation, the only isoperimetric set in $\mathbb{H}^1$ is

$$E_{\text{isop}} = \{(z,t) \in \mathbb{H}^1 : |t| < \arccos|z| + |z|\sqrt{1-|z|^2},\ |z| < 1\}. \tag{24}$$

Only partial proofs of the Pansu's conjecture are known in the literature. The first results on the Heisenberg isoperimetric problem date back to 2008. In [19, Theorem 1.2],

the conjecture is confirmed in the class

$$\mathcal{R} = \{E \subset \mathbb{H}^n : \text{ if } (z,t) \in E, \text{ then } (\zeta,t) \in E \text{ for } |\zeta| = |z|\} \tag{25}$$

of *axially symmetric* sets. Namely, it is proved that the infimum

$$\mathrm{Isop}(\mathcal{R}) = \inf\Big\{\frac{P_H(E)^{2n+2}}{\mathcal{L}^{2n+1}(E)^{2n+1}} : E \in \mathcal{R}\Big\}$$

is attained. Moreover, up to a dilation, a vertical translation and a $\mathcal{L}^{2n+1}$-negligible set, any axially symmetric isoperimetric set (i.e., a set $E \in \mathcal{R}$ such that the infimum in $\mathrm{Isop}(\mathcal{R})$ is attained) coincides with E_{isop}. On the other hand, in [27, Theorem 7.2] it is proved that if $E \subset \mathbb{H}^1$ is an isoperimetric set, whose boundary is a C^2 smooth surface, then up to a dilation and a left translation, $E = E_{\mathrm{isop}}$.

In [22, Theorem 1.1] Pansu's conjecture is proved in $\mathbb{H}^1$ assuming convexity of the isoperimetric set.

In In [25, Theorem 3.1], the following geometric situation is considered. For any $r > 0$, let $D_r = \{(z,0) \in \mathbb{H}^n : |z \le r|\}$ be the closed Euclidean disk of radius r contained in $\{z = 0\}$, and $C_r = \{(z,t) \in \mathbb{H}^n : |z| \le r\}$ be the vertical cylinder over D_r. Let $E \subset \mathbb{H}^n$ be a finite H-perimeter set such that $D_r \subset E \subset C_r$ for some $r > 0$. The author uses a calibration argument to prove that $P_H(E) \ge P_H(E_{\mathrm{isop}})$, and equality holds if and only if $E = E_{\mathrm{isop}}$. In [9] we refine this argument to prove a stability result for the isoperimetric inequality in $\mathbb{H}^n$.

For a detailed review on the Heisenberg isoperimetric problem we refer to the book [3] and to the lecture notes [20].

References

[1] R. Caccioppoli, *Elementi di una teoria generale dell'integrazione k-dimensionale in uno spazio n-dimensionale.* Atti del Quarto Congresso dell'Unione Matematica Italiana, Taormina, 1951, vol. II, pp. 41–49, Casa Editrice Perrella, Roma, (1953).

[2] L. Capogna, D. Danielli, N. Garofalo, *An isoperimetric inequality and the geometric Sobolev embedding for vector fields.* Math. Res. Lett. 1 (1994), no. [2], 263–268.

[3] L. Capogna, D. Danielli, S. D. Pauls, J. Tyson, "Heisenberg group and the sub-Riemannian isoperimetric problem". Progress in Mathematics, 259. Birkhäuser Verlag, Basel, 2007. xvi+223 pp.

[4] E. De Giorgi, *Su una teoria generale della misura* $(r-1)$*-dimensionale in uno spazio ad r dimensioni.* Ann. Mat. Pura Appl. 4 36 (1954), 191–213.

[5] E. De Giorgi, *Nuovi teoremi relativi alle misure* $(r-1)$*-dimensionali in uno spazio ad r dimensioni.* Ricerche Mat. 4 (1955), 95–113.

[6] E. De Giorgi, *Sulla proprietà isoperimetrica dell'ipersfera, nella classe degli insiemi aventi frontiera orientata di misura finita.* Atti Accad. Naz. Lincei. Mem. Cl. Sci. Fis. Mat. Nat. Sez. I (8) 5 (1958), 33–44.

[7] C. Fefferman, D. H. Phong, *Subelliptic eigenvalue problems.* Conference on harmonic analysis in honor of Antoni Zygmund, Vol. I, II (Chicago, Ill., 1981), Wadsworth Math. Ser., Wadsworth, Belmont, CA, (1983), 590–606.

[8] V. Franceschi, R. Monti, *Isoperimetric Problem in H-type groups and Grushin spaces.* Rev. Mat. Iberoam., to appear.

[9] V. Franceschi, G. P. Leonardi, R. Monti, *Quantitative isoperimetric inequalities in $\mathbb{H}^n$.* Calc. Var. Partial Differential Equations, to appear.

[10] B. Franchi, S. Gallot, R. L. Wheeden, *Sobolev and isoperimetric inequalities for degenerate metrics.* Math. Ann. 300 (1994), 557–571.

[11] B. Franchi, E. Lanconelli, *Une métrique associée à une classe d'opérateurs elliptiques dégénérés.* Conference on linear partial and pseudodifferential operators (Torino, 1982). Rend. Sem. Mat. Univ. Politec. Torino 1983 Special Issue, 105–114 (1984).

[12] B. Franchi, R. Serapioni, F. Serra Cassano, *Meyers-Serrin Type Theorems and Relaxation of Variational Integrals Depending Vector Fields.* Houston Journal of Mathematics 22 (1996), 859–889.

[13] N. Garofalo, D.-M. Nhieu, *Isoperimetric and Sobolev inequalities for Carnot–Carathéodory spaces and the existence of minimal surfaces.* Comm. Pure Appl. Math. 49 (1996), 1081–1144.

[14] M. Gromov, "Structures métriques pour les variétés riemanniennes". Edited by J. Lafontaine and P. Pansu. Textes Mathématiques, 1, CEDIC, Paris, 1981. iv+152 pp.

[15] M. Gromov, "Metric structures for Riemannian and non-Riemannian spaces". Based on the 1981 French original. With appendices by M. Katz, P. Pansu and S. Semmes. Translated from the French by Sean Michael Bates. Progress in Mathematics, 152. Birkhuser Boston, Inc., Boston, MA, 1999. xx+585 pp.

[16] P. Hajłasz, P. Koskela, "Sobolev met Poincaré". Mem. Amer. Math. Soc. 145 (2000), no. 688, x+101.

[17] A. Hurwitz, *Sur quelques applications géomtriques des séries de Fourier.* Ann. Sci. École Norm. Sup. (3) 19 (1902), 357–408.

[18] M. Miranda, *Distribuzioni aventi derivate misure insiemi di perimetro localmente finito.* Annali della Scuola Normale Superiore di Pisa - Classe di Scienze, 18 (1964), no. [1], 27–56.

[19] R. Monti, *Heisenberg isoperimetric problem. The axial case.* Adv. Calc. Var. 1 (2008), no. 1, 93–121.

[20] R. Monti, *Isoperimetric problem and minimal surfaces in the Heisenberg group.* Lecture notes of the ERC School *Geometric Measure Theory and Real Analysis*, 57–130, Edizioni SNS Pisa, 2014–2015.

[21] R. Monti, D. Morbidelli, *Isoperimetric Inequality in the Grushin Plane.* J. Geom. Anal., 14 (2004), no. [2], 355–368.

[22] R. Monti, M. Rickly, *Convex isoperimetric sets in the Heisenberg group.* Ann. Sc. Norm. Super. Pisa Cl. Sci. (5) 8 (2009), no. [2], 391–415.

[23] P. Pansu, *An isoperimetric inequality on the Heisenberg group.* Conference on differential geometry on homogeneous spaces (Turin, 1983). Rend. Sem. Mat. Univ. Politec. Torino, Special Issue (1983), 159–174.

[24] P. Pansu, *Une inégalité isopérimétrique sur le groupe de Heisenberg.* C. R. Acad. Sci. Paris Sér. I Math. 295 (1982), no. [2], 127–130.

[25] M. Ritoré, *A proof by calibration of an isoperimetric inequality in the Heisenberg group* $\mathbb{H}^n$. Calc. Var. Partial Differential Equations, 44 (2012), no. [1-2], 47–60.

[26] M. Ritoré, C. Rosales, *Rotationally invariant hypersurfaces with constant mean curvature in the Heisenberg group* $\mathbb{H}^n$. J. Geom. Anal., 16 (2006), no. [4], 703–720.

[27] M. Ritoré, C. Rosales, *Area-stationary surfaces in the Heisenberg group* $\mathbb{H}^1$. Adv. Math. 219 (2008), no. [2], 633–671.

[28] H. A. Schwarz, *Sur une définition erronée de l'aire d'une surface courbe.* Gesammelte Mathematische Abhandlungen, Berlin (1890) II, 309–311.

[29] E. M. Stein, "Harmonic analysis: real-variable methods, orthogonality, and oscillatory integrals". With the assistance of Timothy S. Murphy. Princeton Mathematical Series, 43. Monographs in Harmonic Analysis, III. Princeton University Press, Princeton, NJ, 1993. xiv+695 pp.

Fractional Calculus: Numerical Methods and Models

Abdelsheed Ismail Gad Ameen (*)

Abstract. In this article, we first give a short introduction of fractional calculus (FC) and its geometrical, physical interpretation. Then, we discuss the differential equations of fractional order (Caputo type) which have recently proved to be valuable tools for modeling of many biological phenomena. Most fractional ordinary differential equations (FODEs) do not have exact analytic solutions so that numerical techniques must be used. Hence, we present the fractional Euler method to solve systems of nonlinear FODEs and show how to use this method for solving the Susceptible-Infected-Recovered (SIR) model of fractional order.

1 Preliminaries

In this section, we will present some necessary definitions and notations related to classical calculus. Often, these results can be carried over to the fractional case. Also, we will briefly review some of the important concepts that will be used in this report.

1.1 Integration and differentiation

The fundamental theorem of classical calculus ([1], Theorem 6.18) given a relation between integer order integration and differentiation.

Theorem 1.1 ***(Fundamental Theorem of Classical Calculus)*** *Let $f : [a, b] \to \mathbb{R}$ be a continuous function and let $F : [a, b] \to \mathbb{R}$ be defined by*

$$F(t) = \int_a^t f(s)ds.$$

Then, F is differentiable and

$$F' = f.$$

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: abdelsh@math.unipd.it . Seminar held on May 25th, 2016.

It is one of the goals of fractional calculus to retain this relation in a generalized sense. Throughout this work, It is convenient to use the following notations from now on.

Definition 1.1

(a) By D, we denote the operator that maps a differentiable function onto its derivative, i.e.

$$Df(t) := f'(t) = \frac{d}{dt}f(t).$$

(b) By I_a, we denote the operator that maps a function f, assumed to be (Riemann) integrable on the compact interval $[a, b]$, onto its primitive centered at a, i.e.

$$I_a f(t) := \int_a^t f(s)ds, \tag{1.1}$$

for $a \le t \le b$. If $a = 0$ we will simply write I instead of I_0.

(c) For $n \in \mathbb{N}$ we use the symbols D^n and I_a^n to denote the n–fold iterates of D and I_a, respectively, i.e. we set $D^1 := D$, $I_a^1 := I_a$, and $D^n := DD^{n-1}$ and $I_a^n := I_a I_a^{n-1}$ for $n \ge 2$.

A first result, which will be most important for the later generalization to non-integer integrals (i.e. fractional integrals), can be obtained from this definition. We now begin with the integral operator I_a^n. In the case $n \in \mathbb{N}$, it is well known (and easily proved by induction) (see e.g. [2]) that we can replace the recursive definition of Definition 1.1 (c) by the following explicit formula.

Lemma 1.1 *Let f be Riemann integrable on $[a, b]$. Then, for $a \le t \le b$ and $n \in \mathbb{N}$, we have*

$$I_a^n f(t) = \frac{1}{(n-1)!}\int_a^t (t-s)^{n-1} f(s)ds. \tag{1.2}$$

From this Lemma another consequence can be drawn. In terms of Definition 1.1 the fundamental theorem of classical calculus reads $DI_a f = f$, which implies by Definition 1.1 (c) that $D^n I_a^n f = f$. This leads to the following Lemma:

Lemma 1.2 *Let $m, n \in \mathbb{N}$ such that $m > n$, and let f be a function having a continuous nth derivative on the interval $[a, b]$. Then,*

$$D_a^n f(t) = D^m I_a^{m-n} f(t). \tag{1.3}$$

1.2 Laplace transform and some special functions

Laplace transform and the special functions as Gamma, Mittag-Leffler are most frequently used in the fractional calculus and especially in solving FODEs.

Definition 1.2 We define the Laplace transform of a function $f(t)$, denoted $\mathcal{L}\{f(t)\}$, $0 < t < \infty$ as

$$\mathcal{L}\{f(t);p\} = \int_0^\infty e^{-pt} f(t)dt.$$

The Laplace convolution of two functions $f(t)$ and $g(t)$ is defined as follows:

Definition 1.3 Let $f,\ g \in L_1(\mathbb{R})$. The Laplace convolution of f and g is denoted by $f * g$ and defined as

$$(f * g)(t) := \int_0^t f(t-u)g(u)du, \quad t > 0.$$

The Gamma function, denoted by $\Gamma(z)$, is a generalization of the factorial function $n!$, i.e. $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$. Thus, we have the following definition.

Definition 1.4 For $z \in \mathbb{C}\backslash\{0,-1,-2,-3,...\}$ Gamma function $\Gamma(z)$ is defined as

$$\Gamma(z) = \begin{cases} \int_0^\infty t^{z-1}\ e^{-t}dt, & \text{if} \quad Re(z) > 0 \\ \\ \Gamma(z+1)/z & \text{if} \quad Re(z) \le 0, \quad z \neq 0,-1,-2,-3,... \end{cases}$$

While the Mittag-Leffler function is a generalization of the exponential function (Podlubny [15], p.16).

Definition 1.5 For $z \in \mathbb{C}$ the Mittag-Leffler function $E_\alpha(z)$ is defined by

$$E_\alpha(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k + 1)}, \quad \alpha > 0$$

and the generalized Mittag-Leffler function $E_{\alpha,\beta}(z)$ by

$$E_{\alpha,\beta}(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k + \beta)}, \quad \alpha, \beta > 0. \tag{1.4}$$

2 Fractional Calculus

The main objects of classical calculus are derivatives and integrals of functions. If we start with a function $f(t)$ and put its derivatives on the left-hand side and on the right-hand side we continue with integrals, we obtain a both-side infinite sequence.

$$...\frac{d^2 f(t)}{dt^2},\ \frac{df(t)}{dt},\ f(t),\ \int_a^t f(s)ds,\ \int_a^t \int_a^{s_1} f(s)ds\ ds_1, ...$$

Fractional calculus tries to interpolate this sequence so this operation unifies the classical derivatives and integrals and generalizes them for arbitrary order. Most authors on this topic will cite a particular date as the birthday of so called "Fractional Calculus" [3, 4]. In a letter [5] from Leibniz to L'Hospital dated 3.8.1695, we can find the earliest remarks on the meaning of non-integer derivatives, especially the case 1/2. In this letter Leibniz's response: "An apparent paradox, from which one day useful consequences will be drawn". In these words fractional calculus was born. Following L'Hopital's and Liebniz's first inquisition, fractional calculus was primarily a study reserved for the best minds in mathematics. Consequently, a lot of contributions to the theory of fractional calculus up to the middle of the 20-th century, of famous mathematicians are known: Laplace (1812), Fourier (1822), Abel (1823-1826), Liouville (1832-1837), Riemann (1847), Grünwald (1867-1872), Letnikov (1868-1872), Heaviside (1892-1912), Weyl (1917), Erdèlyi (1939-1965) and many others (see [6]). However, this topic is a matter of particular interest just the last thirty years. For the first specialized conference on fractional calculus and its applications has been organized by B. Ross in June 1974 at the University of New Haven, USA. For the first monograph, the merit is ascribed to K.B. Oldham and J. Spanier [7], who, after a joint collaboration began in 1968, published a book devoted to fractional calculus in 1974. In 1987, the huge book by Samko, Kilbas and Marichev, referred to now as "encyclopedia" of fractional calculus, Miller and Ross ([8], 1993), and Podlubny ([15], 1999), etc.

2.1 Fractional integration and differentiation

We will focus on the Riemann-Liouville, the Caputo operators since they are the most used ones in applications. The results of this subsection are greater parts well known and can be found in various books (see e.g. [2, 7, 8]). Now, Let $L_1 = L_1[a,b]$ be the class of integrable functions on the interval $[a,b]$, $0 \le a < b < \infty$ with the norm defined by:

$$\|f(t)\| = \int_a^b |f(s)|ds, \quad t \in [a,b].$$

Definition 2.1 Let $\alpha \in \mathbb{R}_+$. The operator I_a^α, defined on $L_1[a,b]$ by

$$I_a^\alpha f(t) = \frac{1}{\Gamma(\alpha)} \int_a^t (t-s)^{\alpha-1} f(s)ds, \tag{2.1}$$

for $a \le t \le b$, is called the Riemann-Liouville fractional integral operator of order α. For $\alpha = 0$, we set $I_0^\alpha := I$, the *identity operator*. When $a = 0$, the fractional integral of order $\alpha > 0$ can be considered as the *Laplace convolution* between the causal function $\phi_\alpha(t)$ and $f(t)$, i.e. (see [9])

$$I^\alpha f(t) = f(t) * \phi_\alpha(t), \quad \alpha > 0$$

where

$$\phi_\alpha(t) = \begin{cases} \dfrac{t^{\alpha-1}}{\Gamma(\alpha)}, & \text{for} \quad t > 0, \\ 0, & \text{for} \quad t \le 0. \end{cases}$$

If $\alpha \in \mathbb{N}$ the Riemann-Liouville fractional integral coincides with the classical integral I_a^n in equation (1.2) except that the domain has been extended from Riemann integrable to Lebesgue integrable functions. With the existence of fractional integral of Definition 2.1 guaranteed (see e.g. [8]), we can give the following properties (see [2, 8]):

Lemma 2.1 *For $\alpha,\ \beta > 0$ and $f(t) \in L_1[a,b]$, we have*

$$I_a^\alpha I_a^\beta f(t) = I_a^{\alpha+\beta} f(t) = I_a^\beta I_a^\alpha f(t).$$

And,

$$(I_a^\alpha)^n f(t) = I_a^{\alpha n} f(t); \quad n = 1, 2, 3, \dots ,$$

which is a well known result in the integer case.

Lemma 2.2 *Let I_a^α be defined in L_1, then as $\alpha \to n$ we have*

$$I_a^\alpha f(t) \to I_a^n f(t), \text{uniformly in } L_1, \quad n = 1, 2, \dots ,$$

where $I_a f(t)$ defined by equation (1.1).

We now consider the following examples for fractional integration,

Example 2.1 For $\alpha > 0$ and $t > 0$, we have

$$I^\alpha t^\lambda = \frac{\Gamma(1+\lambda)}{\Gamma(\lambda+\alpha+1)} t^{\lambda+\alpha}, \quad \lambda > -1.$$

In particular, if $\lambda = 0$, then the fractional integral of a constant k of order α is

$$I^\alpha k = \frac{k}{\Gamma(\alpha+1)} t^\alpha.$$

Example 2.2 Let $f(t) = (t-a)^\lambda$ for some $\lambda > -1$ and $\alpha > 0$. Then,

$$I_a^\alpha f(t) = \frac{\Gamma(\lambda+1)}{\Gamma(\lambda+\alpha+1)} (t-a)^{\alpha+\lambda}.$$

Example 2.3 Let $\alpha > 0$, $\lambda > -1$, and $t > 0$ then we have

$$I^\alpha (t^\lambda + 1) = \frac{\Gamma(\lambda+1)}{\Gamma(\lambda+\alpha+1)} t^{\lambda+\alpha} + \frac{t^\alpha}{\Gamma(\alpha+1)}.$$

Until now we only considered the Riemann-Liouville integral operator. For a classical case we have the identity (1.3) (under certain conditions) and we can now motivate the definition of the fractional differential operator by generalizing this identity to non-integer order. There are different definitions for fractional derivatives, which do not coincide in general.

Definition 2.2 Suppose that $\alpha > 0,\ t > a,\ \alpha, a, t \in \mathbb{R}$. Then (see [8, 12])

$$(2.2) \qquad {}_aD_t^{\alpha} f(t) := \begin{cases} \frac{1}{\Gamma(n-\alpha)} \frac{d^n}{dt^n} \int_a^t \frac{f(s)}{(t-s)^{\alpha-n+1}} ds = \frac{d^n}{dt^n} I_a^{n-\alpha} f(t), & n-1 < \alpha < n \in \mathbb{N}, \\ \frac{d^n}{dt^n} f(t), & \alpha = n \in \mathbb{N}, \end{cases}$$

is called the Riemann-Liouville fractional derivative or the Riemann-Liouville fractional differential operator of order α. We note that this operator is the left-inverse operator of the fractional integral (2.1) (see [6]), i.e., $D^{\alpha} I^{\alpha} f(t) = f(t)$.

In 1967 a paper [13] by the Italian mathematician M. Caputo was published, where a new definition of a fractional derivative was used. Now, we state the definition and some properties of Caputo fractional derivative.

Definition 2.3 Suppose that $\alpha > 0,\ t > a,\ \alpha, a, t \in \mathbb{R}$. The fractional operator

$$(2.3) \qquad {}_a^C D_t^{\alpha} f(t) := \begin{cases} \frac{1}{\Gamma(n-\alpha)} \int_a^t \frac{f^{(n)}(s)}{(t-s)^{\alpha-n+1}} ds = I_a^{n-\alpha} D^n f(t), & n-1 < \alpha < n \in \mathbb{N}, \\ \frac{d^n}{dt^n} f(t), & \alpha = n \in \mathbb{N}, \end{cases}$$

is called Caputo fractional derivative or Caputo fractional differential operator of order α.

These definitions are more convenient in many applications in physics, engineering and applied science. But, in Caputo definition we find a link between what is possible and what is practical.

Remark 2.1 Here the symbols ${}_aD_t^{\alpha} f(t)$ and ${}_a^C D_t^{\alpha} f(t)$ are used for the Riemann-Liouville and Caputo fractional derivatives respectively (see [15]), a and t are called terminals (lower and upper correspondingly), if $a = 0$ then the symbols $D^{\alpha} f(t)$ and ${}^C D^{\alpha} f(t)$ are adopted.

Theorem 2.1

(*i*) *Let $\alpha,\ \beta \in (0,1)$ and $f(t)$ is absolutely continuous function on $[a,b]$. If $f'(t)$ is bounded and $\alpha + \beta \in (0,1)$, then*

$${}_a^C D_t^{\alpha}\ {}_a^C D_t^{\beta} f(t) = {}_a^C D_t^{\alpha+\beta} f(t) = {}_a^C D_t^{\beta}\ {}_a^C D_t^{\alpha} f(t).$$

(*ii*) *Let $\alpha \in (0,1)$. If $f(t)$ is absolutely continuous function on $[a,b]$, then*

(a) $I_a^{\alpha}\ {}_a^C D_t^{\alpha} f(t) = f(t) - f(a)$.

(b) ${}_a^C D_t^{\alpha} I_a^{\alpha} f(t) = f(t)$.

Now the following theorem shows the relation between the two definitions.

Theorem 2.2 *Let $t > 0,\ \alpha \in \mathbb{R},\ n-1 < \alpha < n \in \mathbb{N}$. Then the following relation between the Riemann-Liouville (2.2) and the Caputo (2.3) derivatives holds*

$$ {}^{C}D_t^{\alpha} f(t) = D_t^{\alpha} f(t) - \sum_{k=0}^{n-1} \frac{t^{k-\alpha}}{\Gamma(k+1-\alpha)} f^{(k)}(0). $$

Proof. A proof of this theorem is given in [6] using Taylor series expansion. Also, these two definitions coincides if and only if $f(t)$ together with its first $n-1$ derivatives vanish at $t = 0$. □

Recalling the fractional derivative of the power functions, thus we have

Corollary 2.1 *The following relation between the Riemann-Liouville and Caputo fractional derivatives holds*

$$ {}^{C}D^{\alpha} f(t) = D^{\alpha}\left(f(t) - \sum_{k=0}^{n-1} \frac{t^k}{k!} f^{(k)}(0) \right). $$

We now consider the following examples for fractional derivative (Caputo's sense),

Example 2.4 Let $\alpha \in (0,1]$ and $\lambda > 0$, then we have

$$ {}^{C}D_a^{\alpha}(t-a)^{\lambda} = \frac{\Gamma(\lambda+1)}{\Gamma(1+\lambda-\alpha)}(t-a)^{\lambda-\alpha}, $$

also,

$$ \lim_{\alpha\to 1} {}^{C}D_a^{\alpha}(t-a)^{\lambda} = \lambda(t-a)^{\lambda-1}, \quad \text{and} \quad \lim_{\alpha\to 0} {}^{C}D_a^{\alpha}(t-a)^{\lambda} = (t-a)^{\lambda}. $$

Example 2.5 Let $\alpha \in (0,1]$ and $\lambda > -1$, then we have

$$ \begin{aligned} {}^{C}D^{\alpha}(1+t^{\lambda}) &= I^{1-\alpha}(\lambda\, t^{\lambda-1}) = \lambda I^{1-\alpha} t^{\lambda-1} \\ &= \lambda \frac{\Gamma(\lambda)}{\Gamma(\lambda-\alpha+1)} t^{\lambda-\alpha} = \frac{\Gamma(\lambda+1)}{\Gamma(\lambda-\alpha+1)} t^{\lambda-\alpha}. \end{aligned} $$

And,

$$ \begin{aligned} D^{\alpha}(1+t^{\lambda}) &= DI^{1-\alpha}(1+t^{\lambda}) \\ &= D\left(\frac{t^{1-\alpha}}{\Gamma(2-\alpha)} + \frac{\Gamma(\lambda+1)}{\Gamma(\lambda-\alpha+2)} t^{\lambda-\alpha+1} \right) \\ &= \frac{(1-\alpha)t^{-\alpha}}{\Gamma(2-\alpha)} + \frac{(\lambda-\alpha+1)\Gamma(\lambda+1)}{\Gamma(\lambda-\alpha+2)} t^{\lambda-\alpha} \\ &= \frac{t^{-\alpha}}{\Gamma(1-\alpha)} + \frac{\Gamma(\lambda+1)}{\Gamma(\lambda-\alpha+1)} t^{\lambda-\alpha}, \end{aligned} $$

which verifies that $D^\alpha f(t) \neq {}^C D^\alpha f(t)$.

2.2 FODEs of Caputo-type

FODEs are generalizations of classical ordinary differential equations to an arbitrary (non-integer) order. The many important mathematical models are described by differential equations containing fractional-order derivatives. Such models are interesting for engineers, biologists and physicists but also for pure mathematicians. Their evolutions behave in a much more complex way than in the classical integer-order case and the study of the corresponding theory is a hugely demanding task.

The formal definition of a FODE involving Caputo fractional derivative is given as follow

Definition 2.4 Let $\alpha > 0,\ \alpha \notin \mathbb{N},\ n = \lceil \alpha \rceil$ and $f : A \subseteq \mathbb{R}^2 \to \mathbb{R}$. Then

$$ {}^C D^\alpha y(t) = f(t, y(t)) \tag{2.4}$$

is called fractional differential equation of Caputo type. As initial conditions for this type of FDE be

$$ D^k y(0) = y^{(k)}(0) = b_k, \quad (k = 0, 1, ..., n-1). \tag{2.5}$$

To illustrate the main advantage of considering the Caputo fractional derivative, Let the following initial value problems (IVPs)

$$\begin{aligned} &D^\alpha y(t) - \lambda y(t) = 0, \quad t > 0,\ n-1 < \alpha < n \in \mathbb{N},\ \lambda > 0 \\ &D^{\alpha-k-1} y(t)|_{t=0} = b_k, \quad k = 0, 1, ..., n-1 \end{aligned} \tag{2.6}$$

and

$$\begin{aligned} &{}^C D^\alpha y(t) - \lambda y(t) = 0, \quad t > 0,\ n-1 < \alpha < n \in \mathbb{N},\ \lambda > 0 \\ &y^{(k)}(0) = b_k, \quad k = 0, 1, ..., n-1. \end{aligned} \tag{2.7}$$

• In (2.6) the Riemann-Liouville fractional differentiation operator is applicable. In this case, also in the initial conditions fractional derivatives are required. Such initial value problems can successfuly be solved theoretically, but their solutions are practically useless, because there is no clear physical interpretation of this type of initial conditions (see [15], p.78).

• On the contrary, in (2.7) where the Caputo fractional differentiation operator is applicable, standard initial conditions in terms of derivatives of integer order are involved. These initial conditions have clear physical interpretation as an initial position $y(a)$ at the point a (where y is the unknown function), the initial velocity $y'(a)$, initial acceleration $y''(a)$ and so on. On the other hand, the Caputo fractional derivative is more restrictive, as it can be seen from (2.2) and (2.3), since it requires the existence of the $n-$derivative

of the function.

We note that the definitions of fractional derivative involves an integration, which is a non-local operator (as it is define on an interval) and we can understand by using the following formula (2.8) the importance of non-locality for fractional operator. This lemma to show that (2.4)-(2.5) can formulated as Volterra integral equation:

Lemma 2.3 ([14]) *Let $\alpha > 0$, $\alpha \notin \mathbb{N}$ and $n = \lceil \alpha \rceil$. The function $y \in C[0,h]$ is a solution of the FODE of Caputo type (2.4), combined with the initial conditions (2.5) if and only if it is a solution of the nonlinear Volterra integral equation of the second kind*

$$y(t) = \sum_{k=0}^{n-1} \frac{t^k}{k!} b_k + \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} f(s, y(s)) ds. \tag{2.8}$$

Remark 2.2 Let us consider formula (2.8) for some $\alpha \in (0,1]$ (i.e. $n = 1$) and for two different values of t, say t_1 and t_2 with $t_1 < t_2$, then we can write

$$\begin{aligned} y(t_2) - y(t_1) = & \frac{1}{\Gamma(\alpha)} \int_0^{t_1} [(t_2 - s)^{\alpha-1} - (t_1 - s)^{\alpha-1}] f(s, y(s)) ds \\ & + \frac{1}{\Gamma(\alpha)} \int_{t_1}^{t_2} (t_2 - s)^{\alpha-1} f(s, y(s)) ds. \end{aligned} \tag{2.9}$$

(I) In the classical case $\alpha = 1$, the term in brackets on the right-hand side of (2.9) is zero, hence the entire first integral vanishes and we have

$$y(t_2) - y(t_1) = \int_{t_1}^{t_2} f(s, y(s)) ds,$$

this implies that, if we already know the solution $y(t_1)$ of our given problem (2.4) with (2.5) at the point $t_1 > 0$, then we may compute the solution at the point $t_2 > t_1$ exclusively on the basis of $y(t_1)$ and the function f.

(II) In the fractional case $0 < \alpha < 1$, this situation is fundamentally different. Here the first integral on the right-hand side of (2.9) does not vanish in general. Hence, whenever we want to compute the solution $y(t_2)$ at some point t_2 it is necessary to take into account the entire history of y from the starting point 0 up to the point of interest t_2. This reflects the non-locality of the Caputo fractional differential operator.

It thus follows that integer-order equations are appropriate tools for the modelling of systems without memory whereas fractional-order equations are the method of choice for the description of systems with memory. Hence, one of the main advantage of FODEs over ODEs.

3 Fractional Euler method

In general, there are several ways to discretize FODE of Caputo-type; the most often used two techniques are based on the following ideas:

- Discretizing the Caputo derivative directly to get the numerical schemes (**direct methods**).
- Transforming the original fractional equation into the fractional integral equation, then applying the corresponding numerical methods to discretize the fractional integral to get the numerical schemes (**integration methods**).

In this section, we study the fractional Euler method (as an example of the integration methods) for the typical initial value problem (2.4)-(2.5). We now assume that a unique solution of (2.8) exists on some interval $[0, T]$ and we are interested in a numerical solution on the uniform grid $\{t_j = jh : j = 0, 1, ..., N\}$ with some integer N and step-size $h = T/N$. Assuming that we have already calculated the approximations $y_j \approx y(t_j),\ j = 1, 2, ..., k$, the basic idea is to obtain the solution y_{k+1} by replacing the integral on the right-hand side of equation (2.8) by the product rectangle rule

$$\int_0^{t_{k+1}} (t_{k+1} - s)^{\alpha-1} f(s, y(s))ds \approx \sum_{j=0}^{k} a_{j,k+1} f(t_j, y_j),$$

where

$$a_{j,k+1} = \int_{t_j}^{t_{j+1}} (t_{k+1} - s)^{\alpha-1} ds = \frac{(t_{k+1} - t_j)^\alpha - (t_{k+1} - t_{j+1})^\alpha}{\alpha}.$$

In the equispaced case, we have the following expression for weights

$$a_{j,k+1} = \frac{h^\alpha}{\alpha}((k + 1 - j)^\alpha - (k - j)^\alpha).$$

Thus, the explicit recursion

$$y_{k+1} = \sum_{j=0}^{n-1} \frac{t_{k+1}^j}{j!} b_k + \frac{1}{\Gamma(\alpha)} \sum_{j=0}^{k} a_{j,k+1} f(t_j, y_j). \tag{3.1}$$

In the limit case $\alpha \to 1$ the fractional Euler method reduces to the classical forward Euler method. As a consequence of Corollary 2.1 in [23], the error can be estimated as follows:

Theorem 3.1 *The approximation computed by the fractional Euler method satisfies the error bound*

$$|y(t_j) - y_j| = O(h)$$

uniformly for all j if $D^\alpha y \in C^1[0, T]$.

4 Fractional-order SIR model

There are many of models for describing epidemics with different properties with respect to mortality, immunity, time horizon and so on (e.g. [10, 11]). Here, one of these models is examined. Precisely, we considered a standard SIR model with vaccination, treatment and variable total population. We show that this model possesses non-negative solutions as desired in any population dynamics. Also, the stability of equilibrium points is studied. Graphical results are presented and discussed.

4.1 Model description

To derive this model we suppose the total population $N(t)$ is partitioned into three compartments which are Susceptible $S(t)$, Infectious $I(t)$ and Recovered $R(t)$. Let b denote the birth (recruitment) rate of the population, β is the disease transmission rate between infected and susceptible. We assume d to be the natural death rate, σ is the disease-induced death rate. Also, we assume there exists μ_1 and μ_2 which respectively denotes the proportion of the susceptible that is vaccinated per unit time and the proportion of the infectives that is treated per unit time.

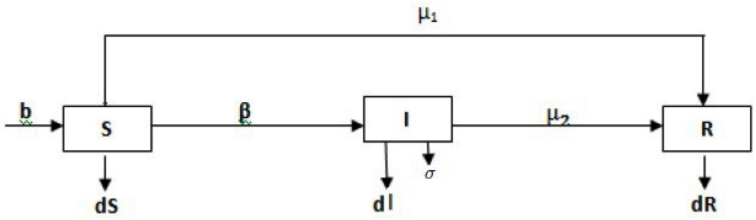


Figure 1. Flowchart showing the compartment model for SIR with μ_1 and μ_2.

The assumptions of the model leads to the following system of FODEs

$$\begin{cases} D^\alpha S(t) = b - \beta S(t)I(t) - (d+\mu_1)S(t), \\ D^\alpha I(t) = \beta S(t)I(t) - (\mu_2 + d + \sigma)I(t), \\ D^\alpha N(t) = b - d\ N(t) - \sigma I(t), \end{cases} \tag{4.1}$$

subject to

$$S(t_0) = S_0, \quad I(t_0) = I_0, \quad N(t_0) = N_0. \tag{4.2}$$

The main reason that leads to this extension (typically with α chosen close to 1) is to reduce the error that may arise from neglected parameters or simplifications in the classical model [11] (i.e. system of first order derivatives). When a disease outbreak occur, the predicted number of individuals who are infected and recovered due to the vaccination by the classical model might be significantly different (less or more) than the realistic data. Hence the fractional model (4.1) possess memory.

We intend to solve the model (4.1) by formula (3.1), which offer accurate solution during a long time interval. This may be important in order to show the effect of vaccination μ_1 and treatment μ_2 of the fractional order model.

4.2 Non-negative solutions

Let $\mathbb{R}^3_+ = \{X \in \mathbb{R}^3 \,|\, X \geq 0\}$ and $X(t) = (S(t), I(t), N(t))^T$, we now prove the main theorem.

Theorem 4.1 *There is a unique solution $X(t) = (S(t), I(t), N(t))^T$ for model (4.1) at $t \geq 0$ (where, $t_0 = 0$) and the solution will remain in $\mathbb{R}^3_+$.*

Proof. From Theorem 3.1 and Remark 3.2 of [16], we know that the solution on $(0, \infty)$ is existent and unique. Now, we will show that the feasible region $\mathbb{R}^3_+$ is positively invariant (non-negative solutions). Rearranging the last equation for the system (4.1) and we assume that $g(t) = b - \sigma I$ is a constant function of time. Then we get the fractional order differential equation representing the total population as follows:

$$D^\alpha N(t) + d\ N(t) = g(t) \tag{4.3}$$

Solving equation (4.3) using Laplace transform method [15] and taking the initial condition to be zero (to simplify), we have the following solution

$$N(t) = \int_0^t (t-\tau)^{\alpha-1} E_{\alpha,\alpha}(-d\ (t-\tau)^\alpha) g(\tau) d\tau\ \geq 0,$$

where $0 < \alpha < 1$, $d > 0$ and $E_{a,b}(x)$ is the two-parameter Mittag-Leffler function (see Definition 2.3). Since Mittag-Leffler function is an entire function [15] thus $E_{\alpha,\alpha}(-d(t-\tau)^\alpha)$ is bounded for all $t > 0$. Therefore, as $n \to \infty$ and $t \to \infty$, we have $N \leq \frac{b}{d}$. For $S(t)$, $I(t)$ by the same way we have $S(t) \geq 0$ and $I(t) = 0$, hence proved that the solution $X(t)$ is positive invariant. □

5 Equilibrium points and their asymptotic stability

To determine the stability analysis, we first evaluate the equilibrium points or steady states of the system (4.1). The equilibrium points involved determine the disease-free (where $I = 0$) and endemic (where $I \neq 0$).

To evaluate the equilibrium points, let

$$\begin{cases} D^\alpha S = 0, \\ D^\alpha I = 0, \\ D^\alpha N = 0, \end{cases}$$

then, the system (4.1) has two equilibrium points

(a) *At disease-free equilibrium:*
The disease-free equilibrium (DFE) of the system (4.1)

$$\varepsilon_1 = (S_{eq}, I_{eq}, N_{eq})_{I=0} = \left(\frac{b}{d+\mu_1},\ 0,\ \frac{b}{d} \right).$$

Using the next-generation operator approach [17, 18], we derive the expression of the basic reproduction number $\mathcal{R}_0$ (see e.g. [19]), allied to the DFE (i.e. ε_1). Following, [17, 18], the next generation matrix is given by (FV^{-1}). Then, we can compute the basic reproduction number as follow:

$$\mathcal{R}_0 = \rho(FV^{-1}) = \frac{b\beta}{(d+\mu_1)(\mu_2+d+\sigma)}, \tag{5.1}$$

where ρ denotes the eigenvalue of largest magnitude or spectral radius.

(b) *At endemic equilibrium:*
In the case where there is infection, we have

$$\varepsilon_2 = (S_{eq}, I_{eq}, N_{eq})_{I\neq 0} = \left(\frac{\mu_2 + d + \sigma}{\beta}, (\mathcal{R}_0 - 1)\frac{\mu_1 + d}{\beta}, \frac{b\beta(\mu_2 + d) + \sigma(\mu_1 + d)(\mu_2 + d + \sigma)}{d\beta(\mu_2 + d + \sigma)} \right)$$

We can note that the equilibrium points are the same for both integer and fractional system. But the stability region of the fractional-order system with order α, which is illustrated in Figure 2 (where σ, ω refer to the real and imaginary parts of the eigenvalues, respectively, and $j = \sqrt{-1}$), is greater than the stability region of the integer order case (see e.g. [20]).

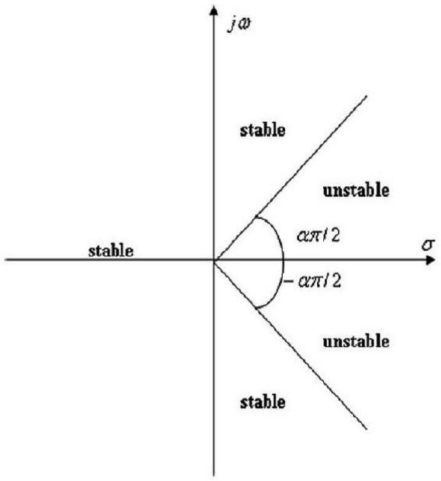


Figure 2. Stability region of the fractional-order system.

Therefore, we will now drive analytically the stability of different equilibria.

For ε_1, ε_3 and the expression (5.1) of $\mathcal{R}_0$, we have the following theorems:

Theorem 5.1 *The disease free equilibria ε_1 of the system (4.1) is locally asymptotically stable if $\mathcal{R}_0 < 1$.*

Proof. Determining the Jacobian matrix of the system (4.1) at ε_1 we have:

$$J_{\varepsilon_1} = \begin{bmatrix} -d-\mu_1 & \frac{-b\beta}{d+\mu_1} & 0 \\ 0 & \frac{b\beta}{d+\mu_1} - \mu_2 - d - \sigma & 0 \\ 0 & -\sigma & -d \end{bmatrix}.$$

The eigenvalues of J_{ε_1} are

$$\lambda_1 = -(d+\mu_1) < 0,\ \lambda_3 = -d < 0,\ \lambda_2 = \frac{b\beta}{d+\mu_1} - (\mu_2 + d + \sigma).$$

Now, we should give the following remark to continue with our proof.

Remark 5.1 Disease free equilibrium ε_1 of the system (4.1) is locally asymptotically stable if $|\arg(\lambda_i)| > \frac{\alpha\pi}{2},\ \forall\ i = 1,2,3$ (see e.g. [21, 22]).

If $\mathcal{R}_0 = \frac{b\beta}{(d+\mu_1)(\mu_2+d+\sigma)} < 1$, then $\frac{b\beta}{d+\mu_1} < (\mu_2 + d + \sigma) \Rightarrow \lambda_2 < 0$ and therefore, $|\arg(\lambda_i)| > \frac{\alpha\pi}{2},\ \forall\ i = 1,2,3$. Thus, disease free equilibrium ε_1 of the system (4.1) is locally asymptotically stable if $\mathcal{R}_0 < 1$. □

Theorem 5.2 *The endemic equilibrium point ε_2 is locally asymptotically stable if $\mathcal{R}_0 > 1$.*

Proof. The Jacobian matrix evaluated at the endemic equilibrium gives

$$J_{\varepsilon_2} = \begin{bmatrix} -\mathcal{R}_0(\mu_1 + d) & -(d + \mu_2 + \sigma) & 0 \\ (\mathcal{R}_0 - 1)(\mu_1 + d) & 0 & 0 \\ 0 & -\sigma & -d \end{bmatrix}$$

and its eigenvalues are

$$\lambda_1 = -d < 0,\ \lambda_{2,3} = \frac{-\mathcal{R}_0(\mu_1 + d) \pm \sqrt{\mathcal{R}_0^2(\mu_1 + d)^2 - 4(\mathcal{R}_0 - 1)(\mu_1 + d)(d + \mu_2 + \sigma)}}{2}.$$

This shows that if $\mathcal{R}_0 > 1$, then $\lambda_2 < 0$ and $\lambda_3 < 0$, hence it becomes asymptotically stable. □

6 Numerical results

The following values, for parameters (see [11]), are considering

$$b = 0.03,\quad d = 0.02,\quad \sigma = 0.1,\quad \beta = 0.75,\quad S_0 = 0.95,\quad I_0 = 0.05,\quad N_0 = 1. \tag{6.1}$$

From this values of parameters, we estimate that $\mathcal{R}_0 = \frac{0.0225}{(\mu_1+0.02)(\mu_2+0.12)}$. The approximate solutions displayed in Figs. 3-5 for step size $h = 0.1$ with different value of fractional order $0 < \alpha \leq 1$ and it is clear that varying the values of μ_1 and μ_2 will alter the number of susceptible and infected persons. If $\mu_1 = \mu_2 = 0$ (i.e. in the absence of vaccination and treatment), then $\mathcal{R}_0 = 9.3750 > 1$ and from the results the disease will persist, while in the beginning of time interval the number of susceptible decrease (see Fig. 3(a)), the number of infected increases (see Fig. 3(b)) and in Fig. 3(c) we can note that $N(t)$ never goes to extinction, this is the main reason for chosen these values of parameters (6.1). If $\mu_1 = \mu_2 = 1$ (i.e. in the presence of vaccination and treatment), $\mathcal{R}_0 = 0.0197 < 1$, the number of susceptible dramatically decreased due to the population have been already vaccinated (see Fig. 5(a)) and the infection will die out (see Fig. 5(b)). About the relevance of vaccination and treatment is obvious from Fig. 3. For the fractional order case, in Fig. 3(b) the climax of $I(t)$ is reduced. But the disease takes a longer time to be eradicated (see Fig. 5(b)). From the numerical results in Figs. 3-5, it is clear that the approximate solutions continuously depends on the time-fractional derivative α.

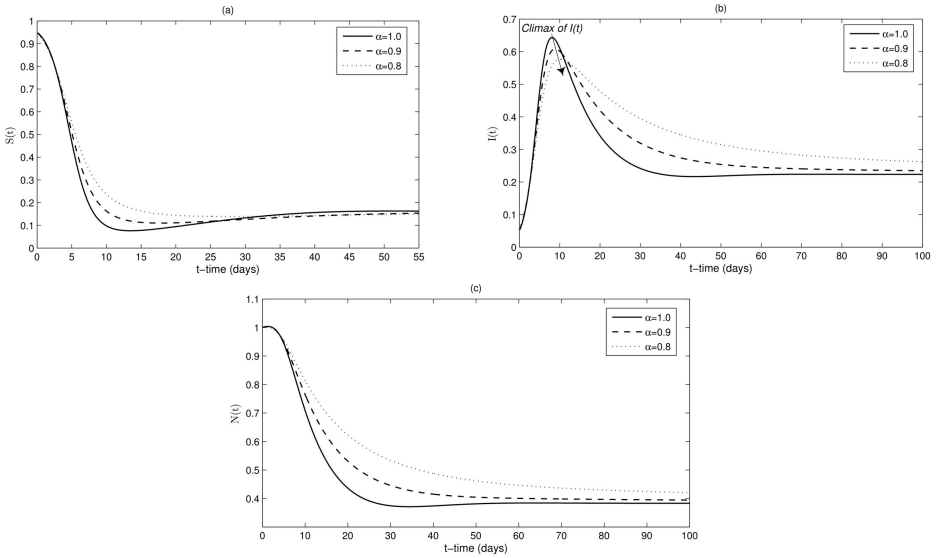


Figure 3. (a) $S(t)$, (b) $I(t)$, (c) $N(t)$ versus t with different values of α and $\mathcal{R}_0 > 1$.

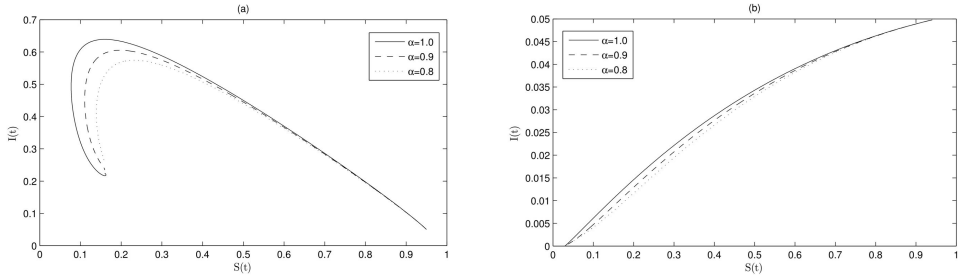


Figure 4. $I(t)$ versus $S(t)$ with different values of α where (a) $\mathcal{R}_0 > 1$ and (b) $\mathcal{R}_0 < 1$.

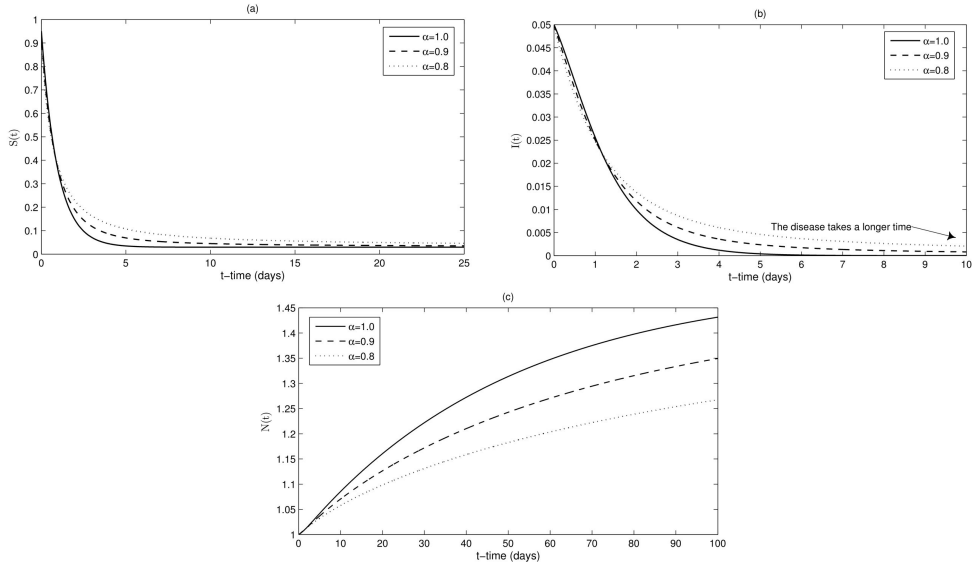


Figure 5. (a) $S(t)$, (b) $I(t)$, (c) $N(t)$ versus t with different values of α and $\mathcal{R}_0 < 1$.

References

[1] W. Rudin, "Principles of mathematical analysis". McGraw-Hill Book Company, Inc., New York-Toronto-London (1953).

[2] S. G. Samko, A. A. Kilbas ,O. I. Marichev, "Fractional Integrals and Derivatives: Theory and Applications". Gordon and Breach Science Publishers, Amsterdam (1993).

[3] Z. E. A. Fellah, C. Depollier, *Application of fractional calculus to the sound waves propagation in rigid porous materials.* Acta Acust united Ac 88 (2002), 34–39.

[4] R. L. Magin, *Fractional calculus in bioengineering.* Crit. Rev. Biomed. Eng. 32, 1–104 (2004).

[5] G. H. Pertz, C. J. Gerhardt, *Leibnizens gesammelte Werke, Lebinizens mathematische Schriften, Erste Abtheilung, Band II.* Pages 301–302. Dritte Folge Mathematik (Erster Band). A. Asher & Comp., 1849. Briefwechsel zwischen Leibniz, Hugens van Zulichem und dem Marquis de l'Hospital.

[6] R. Gorenflo, F. Mainardi, *Essential of fractional calculus.* MaPhySto Center (2000) `http://www.maphysto.dk/oldpages/events/LevyCAC2000/MainardiNotes/fm2k0a.ps`.

[7] K. B. Oldham, J. Spanier, "The fractional calculus". Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London (1974).

[8] K. S. Miller, B. Ross, "An introduction to the fractional calculus and fractional differential equations". A Wiley-Interscience Publication. John Wiley & Sons Inc., New York (1993).

[9] I. M. Gelfand, G. E. Shilov, "Generalized Functions". New York-London, Academic Press (1997).

[10] E. A. Bakare, A. Nwagwo, E. Danso-Addo, *Optimal control analysis of an SIR epidemic model with constant recruitment.* IJAMR, 3 (2014), 273–285.

[11] T. T. Yusuf, F. Benyah, *Optimal control of vaccination and treatment for an SIR epidemiological model.* WJMS, 8 (2012), 194–204.

[12] I. Podlubny, A. M. A. El-Sayed, *On Two Definitions of Fractional Calculus.* Solvak Academy of science-institute of experimental phys, UEF-03-96 ISBN: 80-7099-252-2 (1996).

[13] M. Caputo, *Linear models of dissipation whose Q is almost frequency independent II.* Geophys. J. Royal astr. Soc. 13 (1967), 529–539.

[14] A. A. Kilbas, S. A. Marzan, *Cauchy problem for differential equations with Caputo derivative.* (Russian), Dokl. Akad. Nauk 399 (2004), 7–11.

[15] I. Podlubny, "Fractional differential equations". Academic Press Inc., San Diego, 1999.

[16] W. Lin, *Global existence theory and chaos control of fractional differential equations.* J. Math. Anal. Appl., 332 (2007), 709–726.

[17] O. Diekmann, J. A. P. Heesterbeek, "Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation". Wiley, New York, 2000.

[18] P. Van den Driessche, J. Watmough, *Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission.* Math. Biosci., 180 (2002), 29–48.

[19] W. O. Kermack, A. G. McKendrick, *A Contribution to the mathematical theory of epidemics.* Proc. Roy. Soc. Ser A, 115 (1927), 700–721.

[20] H. A. A. El-Saka, *The fractional-order SIS epidemic model with variable population size.* J. Egyptian Math. Soc., 22 (2014), 50–54.

[21] E. Ahmed, A.M.A. El-Sayed, H.A.A. El-Saka, *On some Routh-Hurwitz conditions for fractional order differential equations and their applications in Lorenz, Rössler, Chua and Chen systems.* Phys. Lett. A 358 (2006), 1–4.

[22] M. El-Shahed, A. Alsaedi, *The fractional SIRC model and influenza A*. Math. Probl. Eng. 2011 (2011) 1-9.

[23] D. Baleanu, K. Diethelm, E. Scalas, J. J. Trujillo, "Fractional Calculus: Models and Numerical Methods". World Scientific Publishing Co Pte Ltd, 2012.

Polyhedral structures in algebraic geometry

Stefano Urbinati (*)

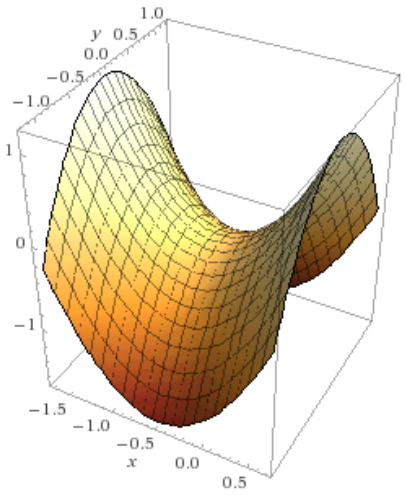
Abstract. Algebraic geometry studies the zero locus of polynomial equations connecting the related algebraic and geometrical structures. In several cases, nevertheless the theory is extremely precise and elegant, it is hard to read in a simple way the information behind such structures. A possible way of avoiding this problem is that of associating to polynomials some polyhedral structures that immediately give some of the information connected to the zero locus of the polynomial. In relation to this strategy I will introduce Newton-Okounkov bodies and Tropical Geometry.

1 Introduction

Algebraic geometry studies the solution sets of polynomial equations. As a motivation one could think to the study of manifolds.

The main objects in algebraic geometry are called algebraic varieties, shapes that locally can be described as the zero set of finitely many polynomials, i.e. where these polynomials vanish.

As an example, this is the graph of $x^2 - y^2 + x$ in $\mathbb{R}^3$:



(*)Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: urbinati@math.unipd.it . Seminar held on June 1st, 2016.

At a first glance, it seems that only restricting to polynomials we do loose a lot, on the other hand we gain several benefits:

- We can study things that are not smooth, in comparison to the study of differentiable manifolds.
- Polynomials can be defined over any ring and field. Thus we can talk about algebraic geometry over $\mathbb{C}, \mathbb{R}, \mathbb{Q}, \mathbb{Z}$, finite fields as Z_p and $\mathbb{Q}_p$ (p-adic field).
- Any function can be approximated by polynomials, thus any reasonable question shall be first studied for polynomials.
- All polynomials with fixed many variables form a ring. Also, locally an algebraic variety is given by a ring. Thus we have a dictionary

  $$\text{Algebra} \leftrightarrow \text{Algebraic Geometry}$$

  that allow us to translate any question from one language to the other and use the powerful tools of both.

Still, even being the simple objects, polynomials are really complicated and there are several theories associating to a given polynomial some polyhedral structures to better understand some of the underlying properties.

2 The Newton polytope

In this section we will introduce the key object linking algebraic geometry and the world of polyhedral geometry. In particular it takes name by Newton that first introduced it in two letters to Leibniz from 1976.

Now on we will work over the complex numbers.

Definition 1 Let $F(x_1, \dots, x_n) \in \mathbb{C}[x_1, \dots, x_n]$ a complex polynomial. Then

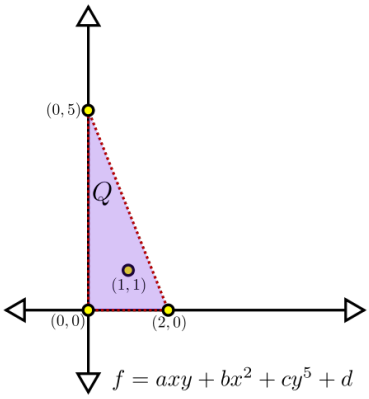
$$F(x_1, \dots, x_n) = \sum a_{i_1,\dots,i_n} x_1^{i_1} \cdot \dots \cdot x_n^{i_n}.$$

To such a polynomial we associate a polytope $\Delta(F) \subseteq \mathbb{R}^n$ defined by

$$\Delta(F) = \text{convex hull}\{(i_1, \dots, i_n) \in \mathbb{N}^n | a_{i_1,\dots,i_n} \neq 0 \text{ in } F\}.$$

We call it the Newton Polytope of F.

Example 2 Let us for example cosider the polynomial $f(x, y) = axy + bx^2 + cx^5 + d$ with $a, bc, d \neq 0$. Then we obtain:



Remark 3 This very simple construction turns out to have great applications as finding the solutions of a given equation or the geometric genus of a curve. We will just give few examples.

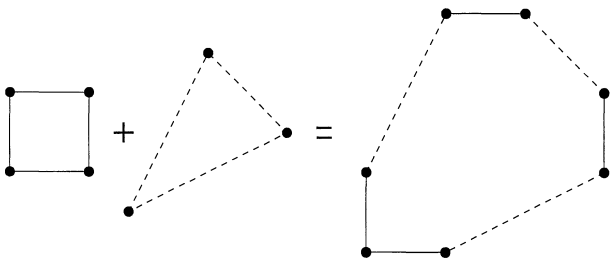
Let us begin with Khovanskii's theorem. Given a polynomial $F(x, y) \in \mathbb{C}[x, y]$, then it defines naturally an algebraic curve $C = \{(x, y)|F(x, y) = 0\} \subseteq \mathbb{C}^2$. Topologically, after desingularization and compactification, this curve can be seen as a compact surface in $\mathbb{R}^3$ with g-holes. The genus of the curve C is the number g and it is a natural invariant of the curve.

Theorem 4 (Khovanskii, 1977) *Consider a generic complex curve $F(x, y) = 0$ in the plane, with fixed Newton polyhedron Δ. Its genus is equal to the number of points with integer coordinates in the interior of the Newton polytope Δ.*

Thus, the curve $y^2 + Q_3(x) = 0$, where $Q_3(x)$ is a sufficiently generic polynomial of degree three, has genus 1, since its Newton polytope contains only one integral point (with coordinates $(1, 1)$).

Before moving on, I will introduce one of the most important operations that can be done with polytopes (and actually with any subset of points in $\mathbb{R}^n$).

Definition 5 By the *Minkowski sum* of two subsets of a linear space we mean the set of all sums of pairs of vectors, one from the first subset and another from the second one.



Remark 6 The product of a subset by a number is defined similarly. The Minkowski sum of convex bodies (convex polyhedra, convex polyhedra with vertices at points with integer coordinates) is a convex body (convex polyhedron, convex polyhedron with vertices at points with integer coordinates).

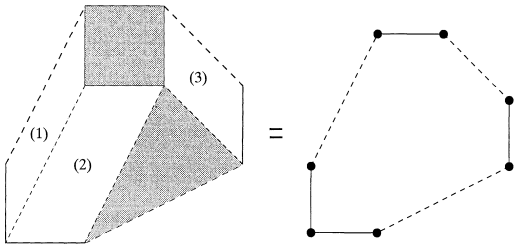
One of the key properties of Newton polytopes connected with Minkowski sums is the following:

Theorem 7 (Ostrowski, 1975) *Let $f, g, h \in \mathbb{R}[x_1, \ldots, x_n]$ with $f = g \cdot h$. Then $\Delta(f) = \Delta(g) + \Delta(h)$, where the latter is the Minkovski sum.*

We can now define the mixed volume of two polytopes. In particular for what we will need for the next theorem let us focus on the two dimensional case.

Definition 8 Let P and Q two polytopes in $\mathbb{R}^2$, then the mixed volume of P and Q is defined as

$$V(P,Q) = vol(P+Q) - vol(P) - vol(Q).$$



We can now state the following theorem.

Theorem 9 (Bernstein, 1975) *Let us consider $g, h \in \mathbb{C}[x, y]$ forming a system with finitely many solutions. Then the number of solutions in $(\mathbb{C}^*)^2$ is at most the mixed volume $V(\Delta(g), \Delta(h))$. If the two polynomials are general then we obtain an equality.*

3 Few words about tropical geometry

In this section I want to introduce the notion of tropical variety. It will be very informal and it is mostly supposed to give an idea on how to use it.

Tropical algebraic geometry is algebraic geometry over the tropical semiring.

Definition 10 The tropical semiring is $(\mathbb{C} \cup \infty, \oplus, \odot)$ where

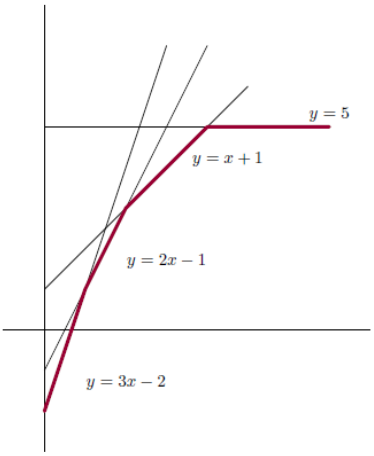
$$a \oplus b = \min\{a, b\} \qquad a \odot b$$

Let $x_1, \dots, x_n$ be variables. A tropical monomial is any tropical product of these variables, where repetitions are allowed.

For example: $x_2 \odot x_1 \odot x_3 \odot x_1 \odot x_4 \odot x_2 \odot x_3 \odot x_2 = x_1^2 x_2^3 x_3^2 x^4$.

A tropical polynomial is a finite linear combination of tropical monomials.

Example 11 $p(x) = -2 \odot x^3 \oplus -1 \odot x^2 \oplus 1 \odot x \oplus 5 = \min\{3x - 2, 2x - 1, x + 1, 5\}$
obtaining the graph



Definition 12 Given a tropical polynomial p, a tropical hypersurface $V(p)$ is

$$V(p) = \{x \in \mathbb{R}^n : \text{the minimum in } p(x) \text{ is attained at least twice}\} =$$

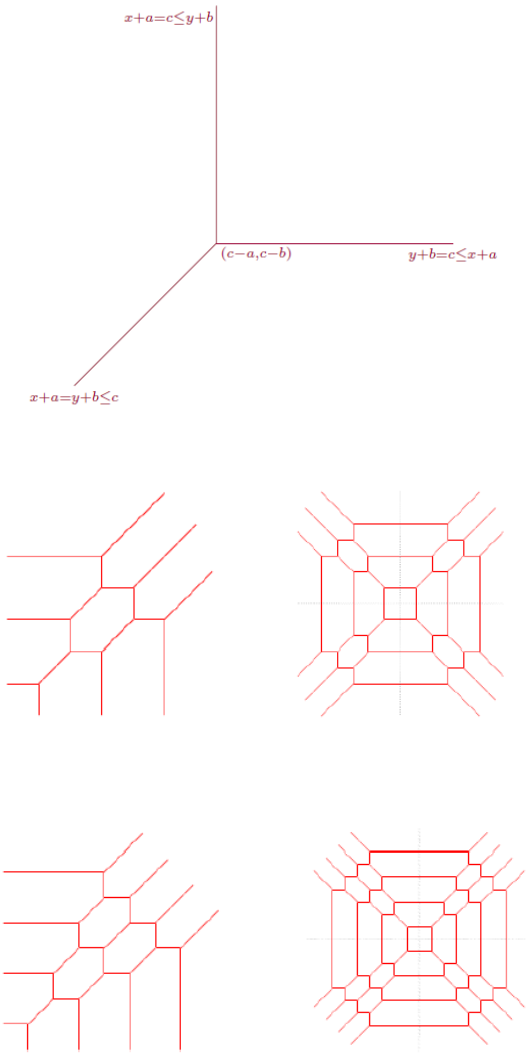
$$= \{x \in \mathbb{R}^n : \ p \text{ is not linear at } x\}.$$

Example 13 $p(x, y) = a \odot x \oplus b \odot y \oplus c$ with $a, b, c \in \mathbb{C}$,
$V(p) = \{(x, y) \in \mathbb{R}^2 : \min\{x + a, y + b, c\}$is attained at least twice$\}$ is a tropical line in the plane $\mathbb{R}^2$.

Of course tropical varieties can be much more complicated. In general the structure depends on the valuation given to the base field, that in several useful examples happens to be the field of Puiseux series. These are some possible shapes.

What is my interest in this type of geometry? From a wider point of view, tropical geometry mainly focuses in the study of subvarieties of open Tori equipped with a valuation. The valuation induces a map to $\mathbb{R}^n$ that after a limiting operation describes the tropicalization of the initial subvariety. I am interested in how the tropicalization gives information on the local behavior of bundles on a variety.

In 2005 J. Tevelev proved that the tropicalization of a closed subvariety Y of a n-dimensional torus induces a 'good' compactification of Y in a toric variety. With C.



Novelli, in [12], we show how this compactification reconstructs the notion of tropical line bundle introduced by C. Torchiani on tropical cycles, starting from the original variety and giving a birational nature to the whole structure.

4 Link between the polytope and the tropical algebra

The Newton polytope is a very classical and simple object that has several surprising properties. In this section we want to give a more general and modern structure generalizing the Newton polytope and connecting it to more recent theories.

Let me first review the definition of McMullen's polytope algebra.

Definition 14 The polytope algebra Π is a $\mathbb{R}$-algebra, with a generator $[P]$ for every

polytope P in $\mathbb{R}^n$, and $[\varnothing] = 0$. The generators satisfy the relations

(V) $[P \cup Q] + [P \cap Q] = [P] + [Q]$, whenever $P \cup Q$ is a polytope

(T) $[P + t] = [P]$, for all translations $t \in \mathbb{Q}^n$

(M) The multiplication in Π is given by $[P] \cdot [Q] := [P+Q]$, where $P+Q = \{p+q : p \in P, q \in Q\}$ is the Minkowski sum. The multiplicative unit is the class of a point: $1 = [\{0\}]$. A basic relation in Π states that $([P]-1)^{n+1} = 0$. This implies that the logarithm of a polytope P is well-defined:

$$\log([P]) = \sum_{r=1}^{n} \frac{(-1)^{r+1}}{r} ([P]-1)^r.$$

It is known that Π is a graded $\mathbb{Q}$-algebra, $\Pi = \bigoplus_{k=0}^{n} \Pi_k$. The k-th graded component Π_k is the $\mathbb{Q}$-vector space spanned by all elements of the form $(\log([P]))^k$, where P runs over all polytopes in $\mathbb{Q}^n$.

We now fix a polytope $P \subset \mathbb{Q}^n$, and we define $\Pi(P)$ to be subalgebra of Π generated by all classes $[Q]$, where Q is a Minkowski summand of P (i.e., $P = \lambda Q + R$, for some positive rational λ and some polytope R). Let denote Σ the normal fan of P, and let $X = X(\Sigma)$ be the corresponding projective toric variety, i.e. a complex variety that has $(\mathbb{C}^*)^n$ as a dense open subset. Note that the algebra $\Pi(P)$ depends only on the fan Σ, and hence it is an invariant of the toric variety X.

The following is the key property that interests my research.

To every rational polytope we can naturally associate a tropical hypersurface. The tropical hypersurface $\mathcal{T}(P)$ of a rational polytope $P \subseteq \mathbb{Q}^n$ is then defined to be the set of normal cones of P of dimension at most $n-1$.

As for polytopes, also tropical hypersurfaces form an algebra denoted by T with an operation similar to the mixed volume called *stable intersection*.

Theorem 15 (Jense-Yu, 2015) *There is an isomorphism of graded algebras*

$$\varphi : \Pi \to T$$

given by

$$\varphi([P]) = 1 \oplus \mathcal{T}(P) \oplus \frac{1}{2!}\mathcal{T}^2(P) \oplus \cdots \oplus \frac{1}{n!}\mathcal{T}^n(P)$$

for polytopes P and linearly extending to Π.

Under this map, $\log([P]) \mapsto \mathcal{T}(P)$, so tropicalization is the logarithm.

This is the basic situation involving the tropical and the polytope algebra. I will now give a more detailed idea of what I am doing.

5 Linear series and Newton-Okounkov bodies

In recent years there has been a successful try to generalize the notion of Newton polytope to a more general setting, i.e. involving linear series of divisors. This yelds to the definition of Okounkov bodies. It originates from papers due to A. Okounkov from the middle of the 1990s (for instance [13]). More recently, Lazarsfeld and Mustaţă [9] and independently Kaveh and Khovanskii [8] initiated an intensive research on the topic recording strong relations of the construction to properties of linear series that were not observed at first.

The idea of the construction of Okounkov bodies is to associate simple geometric objects to linear series on normal varieties. This idea comes as a generalization of the toric case, where to each divisor D is associated a polytope P_D. As for the toric, also in the general case Okounkov bodies are convex bodies which encode several properties of linear series, as for example, their volume. The idea of the construction is quite natural, even though it is in general very hard to determine them.

One possibility introduced for the case of surfaces in [10] to simplify the construction is that of finding *minimal elements* generating all the possible bodies. As for the surface case, the philosophy of Zariski decomposition plays an important role in the generalization of the construction. First, since the global sections of a divisor are asymptotically the same as those of its movable part, it is enough to consider the bodies associated to movable divisors. Second, for a Mori Dream Space X, movable divisors generate a cone $\mathrm{Mov}(X)$, that can be subdivided in chambers representing the nef cone of different flipped models of X ([7]). A possible approach to construct these minimal bodies is to find indecomposable polytopes associated to the extremal rays of the nef cone (since it is not round) of each model and ask whether these elements are enough to reconstruct all the possible bodies. This will yield a Minkowski decomposition of the polytope, and the set of indecomposable elements is called *Minkowski base*.

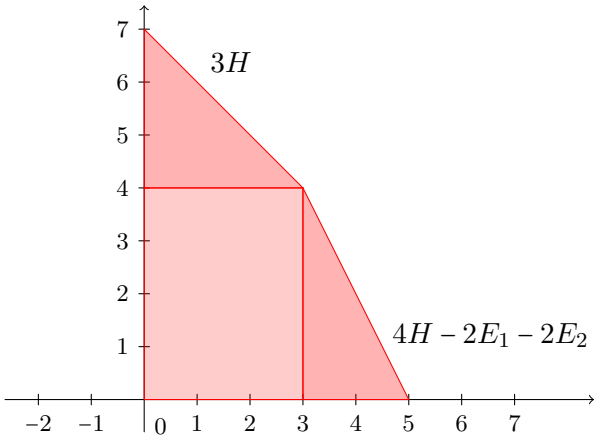
In [11], D. Schmitz and P. Łuszcz-Świdecka prove that for smooth projective surfaces whose pseudo-effective cone is rational polyhedral, the Okounkov body of a big divisor with respect to a general flag decomposes as the Minkowski sum of finitely many simplices and segments arising as Okounkov bodies of nef divisors.

Example 16 Consider the blow up of $\mathbb{P}^2$ in two points with exceptional divisors E_1, E_2. From the construction in the proof of the above result we obtain the set of divisors

$$\{H, H - E_1, H - E_2, 2H - E_1 - E_2\}$$

as the set MB in this case. Here H denotes the pullback of the class of a line in ¶^{2}. Let us consider the big and nef divisor $D = 7H - 2E_1 - 2E_2$. We can write $D = 3H + 2(2H - E_1 - E_2)$ and obtain the following decomposition of Okounkov bodies

$$\triangle(D) = 3 \triangle (H) + 2 \triangle (2H - E_1 - E_2).$$



In the higher dimensional case, as mentioned above, it is obvious that nef divisors will not be enough to obtain a similar result, since having a flip for the variety will imply a change of the nef cone but not of the Okounkov body of the divisors. The aim is instead to look at Minkowski indecomposable polytopes arising from divisors in the movable cone, $\mathrm{Mov}(X)$.

Among the prominent objects studied in relation with Okounkov bodies there are toric varieties. For instance in [1], it is shown that varieties having an ample divisor with a rational polyhedral Okounkov body with respect to some flag, admit a flat degeneration to a toric variety.

In [14], with Piotr Pokora and David Schmitz, as a first step in the generalization for higher dimensional varieties of the result in [11], we study the case of Okounkov bodies on toric varieties constructed with respect to torus-invariant flags. One of the main tools that we are going to use is given in [9], where the authors identify the polytopes arising as toric polytopes and those coming as Okounkov bodies. The main result we prove is:

Theorem *The set of all T-invariant divisors D on a smooth projective toric variety X such that there exists a small modification $f : X \dashrightarrow X'$ and divisor D' spanning an extremal ray of $Nef(X')$ such that $D = f^*(D')$ forms a Minkowski base with respect to T-invariant flags.*

Even more, we give a complete algorithm to find the Minkowski base for a projective toric variety of arbitrary dimension, whose elements will correspond to the extremal rays of the chambers decomposing the movable cone $\mathrm{Mov}(X)$. This also gives a complete description of the secondary fan (or GKZ decomposition) for the movable cone of the given toric variety. Moreover this correspondence implies that the Minkowski base is unique up to numerical equivalence and scaling.

References

[1] Anderson, D., *Okounkov bodies and toric degenerations.* Math. Ann. 356, No. 3 (2013), 1183–1202.

[2] Bauer, T., Küronya, A., Szemberg, T., *Zariski chambers, volumes, and stable base loci.* J. Reine Angew. Math. 576 (2004), 209–233.

[3] Bauer, Th., Kovács, Küronya, A., Mistretta, E. C., Szemberg, T., Urbinati, S., *On positivity and base loci of vector bundles.* European Journal of Mathematics 1 (2015), 229–249.

[4] Sébastien Boucksom, Tommaso de Fernex, and Charles Favre, *The volume of an isolated singularity.* Duke Math. J. Volume 161, Number 8 (2012), 1455–1520.

[5] Cox, D. A., Little, J. B., Schenck, H., "Toric Varieties". Graduate Studies in Mathematics, American Mathematical Society, Volume 124, 2011.

[6] Ein, L. and Lazarsfeld, R. and Mustaţă, M. and Nakamaye, M. and Popa, M., *Asymptotic invariants of line bundles.* Pure and Applied Mathematics Quarterly, vol. 1 (2005), Special Issue: In memory of Armand Borel. Part 1, 379–403.

[7] Hu, Y, Keel, S., *Mori dream spaces and GIT.* Michigan Math. J. Volume 48, Issue 1 (2000), 331–348.

[8] Kaveh, K, Khovanskii, A. G., *Newton-Okounkov bodies, semigroups of integral points, graded algebras and intersection theory.* Ann. of Math. (2) 176 (2012), no. 2, 925–978.

[9] Lazarsfeld, R., Mustaţă, M., *Convex bodies associated to linear series.* Ann. Sci. Éc. Norm. Supér. (4) 42 (2009), no. 5, 783–835.

[10] Łuszcz-Świdecka, P., *On Minkowski decomposition of Okounkov bodies on a Del Pezzo surface.* Annales Universitatis Paedagogicae Cracoviensis Studia Mathematica, 10(1) (2012), 105–115.

[11] Łuszcz-Świdecka, P., Schmitz, D., *Minkowski decomposition of Okounkov bodies on surfaces.* ArXiv:1304.4246 (2013).

[12] Novelli, C., and Urbinati, S., *A note on the birational geometry of tropical line bundles.* ArXiv:1504.03827 (2015).

[13] Okounkov, A., *Brunn-Minkowski inequality for multiplicities.* Invent. Math. 125 (1996), 405–411.

[14] Pokora, P., Schmitz, D., Urbinati, S., *Minkowski decomposition and generators of the moving cone for toric varieties.* Arxiv e-prints, 1310.8505 (2013).

Computed Tomography: a real case example of inverse problem

Elena Morotti (*)

Abstract. X-ray computed tomography (CT) is a well known medical imaging technique, that seeks to reveal internal structures hidden by the skin and bones. Mathematically, the CT process can be modelled as a linear system and the image reconstruction is a challenging inverse problem. In this talk I will show both phisical and mathematical basic concepts, to explain the CT process, and the two possible approaches to solve the problem (leading to analitical or iterative numerical methods). Finally, I will shortly introduce the Digital Breast Tomosynthesis (DBT) technology, that is a 3D emerging technique for the diagnosis of breast tumors, together with numerical results for a simulated problem.

1 Historical overview

Computed tomography is a recent technique allowing us to see inside a human body, but such a visualization strictly invoques only one section at a time, of the object of interest. Tomography makes use of X-rays, discovered by Wilhelm Conrad Rontgen in 1895: as soon as scientists realized X-ray capability of crossing any body, medical imaging was born. Classical tomography, in particular, developed into a medical imaging technique in 1930s, thanks to the italian radiologist Alessandro Vallebona's studies about stratigraphy. Thanks to the arrival of computers, tomographic imaging had a turning point when it developped into *computed* tomography. This technique was designed by Mr. Allan Cormack (a phisicist) and Mr. Godfrey Hounsfield (an engineer) and it is based on a circular scan of the object: many projections of the same slice, in fact, are acquired from many angled views, in a round angle trajectory. First CT device was installed in London in 1971 and the two inventors won the Nobel Prize for Medicine in 1979, for such a revolutionary innovation.

Figure 1 shows the evolution of CT devices. First generation ones (on the left hand side) had a primitive X-ray source, hence it emitted only one ray at a time. That's why a translation motion of the source-detector couple was required to catch projections of the whole section. After that, according to the circular strategy, source and detector rotated

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: morotti@math.unipd.it . Seminar held on June 15th, 2016.

to the second scanning angle and a new scan was performed with translation steps. Repeting such a small step translation for every view made the resulting exam very slow. That' s why second generation devices (central figure) opened the X-ray beam into a small fan: catching a larger number of projections, the translation step size could increase, hence each angled view took less time. Of course, in this machineries the detector was wider, in order to receive all the rays.

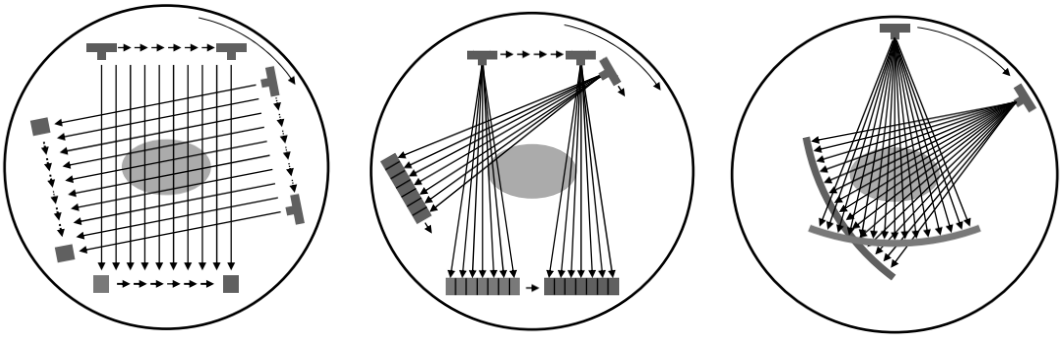


Figure 1. Tomographic device sketchs, in their evolution: from first generation on the left to third generation one on the right.

Due to technology developments, in 1990s a new CT device generation was designed: translations are no longer necessary because a 35-50 degree fan beam can be emitted by a powerfull source, and a long detector can be installed on the device (see figure on the right). The resulting acquisition phase may take only few seconds for the medical exam. Such a rapidity, pushed researchers to develop the so-called *helical* CT for multi-slice investigations, that can be seen as a volumetric tomographic imaging technique. In addition, reconstruction softwares may run real time and achieve high-quality images for new generation devices, while earlier ones were equipped with slow softwares, providing low-resolution outputs in many minutes.

2 Physical basis concepts

2.1 Lambert Beer law and Radon transform

To understand phisical aspects of the CT process, let's start with a simple example. A mono-energetic X-ray exits from the source with I_0 intensity and it passes through a ds thick object, made of a certain material. The detector reveals I photons, hence $dI = I_0 - I$ photons have been absorbed by the small volume. The law describing such attenuation is the $Lambert - Beer\ law$ and it states that

$$I = I_0 e^{-\mu ds} \tag{1}$$

where $\mu = \mu_\lambda \geq 0$ is the attenuation coefficient of the matter, according to the wave length λ of the emitted X-ray.

In a real case, we can identify every point of the slice of interest with a bi-dimensional (x, y) coordinate and the attenuation coefficient is a real function $\mu(x, y)$ over the spatial domain of the section.
In particular, μ is a map: for every spatial point (x, y), we get its attenuation coefficient, hence we can deduce what this volumetric element it is made of. For instance, $\mu(x, y) \approx 0$ means (x, y) is made of air, which doen not absorb any photons, and it occurs whenever (x, y) is outside the body.

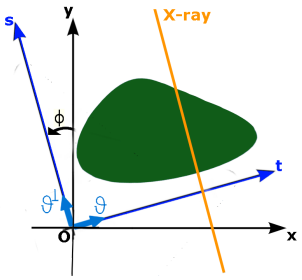


Figure 2. khbsd

Let us fix one ray through its path L and the emittion angle Φ (with respect to the Cartesian system xOy). The total absorption for the I_0-intensity ray, after crossing the object along L, is

$$I = I_0 \exp\left(-\int_L \mu(x, y)d\ell\right)$$

hence:

$$-\log\left(\frac{I}{I_0}\right) = +\int_L \mu(x, y)d\ell \geq 0 \tag{2}$$

Let us consider a parallel beam and the Cartesian system tOs with versors $\theta = (\cos\Phi, \sin\Phi)$ and $\theta^\perp = (-\sin\Phi, \cos\Phi)$, as shown in Figure 2. In this new system, any X-ray L is a function of Φ (or θ) and t. For this reason we can reformulate the line integral in (2) in these new coordinates, getting

$$\int_L \mu(x, y)d\ell = \int_{-\infty}^{+\infty} \mu(t\theta + s\theta^\perp)ds \;=: P_\theta(t) \tag{3}$$

that perfectly mirrows the projection P of the object along a θ-sloping ray in t. From now on, θ stands for the scanning angle too, since it determine Φ uniquely.

Given the current scanning angle θ, the *Radon transform of* μ is defied as the map $\mathcal{R}_\theta : \mu(x, y) \mapsto P_\theta$, such that

$$(\mathcal{R}_\theta\mu)(t) = P_\theta(t) = \int_{-\infty}^{+\infty} \mu(t\theta + s\theta^\perp)ds \qquad \forall\ t \in \mathbb{R} \tag{4}$$

The circular process, defining CT, is based on continuous radial acquisition, i.e. a Radon transform of the slice of interest is measured by the detector from all the angles in $[0, 2\pi]$. Of course devices can only performe small angular steps, hence a finite number of scans are performed from prefixed angles $\theta_k \in \{\theta_1, \ldots, \theta_{n_{angles}}\}$, and because of the finiteness of detector components also projections are recorded in a finite numeber of points t_i $\forall i = 1, \ldots, n_{pixel}$.
The display of all the collected data is called *sinogram* when organized in the Cartesian system (θ, t).

2.2 Back-projection process

Once we collect all the projection data, the mathematical inverse problel is defined: how can we come back to the original slice, starting from the sinogram information?

The basic idea is to project backward every data onto the original ray-path causing such absorbtion.

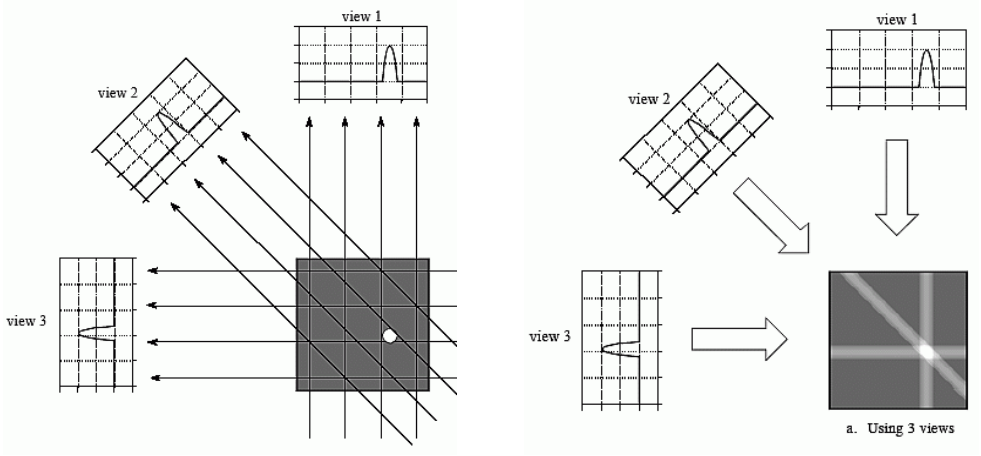


Figure 3. Three projections are acquired for a simple slice and back-projected onto an empty image, simulating first steps of a BP algorithm.

Figure 3 shows how the Back Projection algorithm works, taking only three views of a simple section. For each Radon transform, the small object in this slide provides non-zero values in particular absisses, accoding to the scanning angle: considering what we can learn from each back-projected image on a unique slice, we obtain the tomographic reconstruction. From the shown example, it becomes clear how taking *many* views (from a circular trajectory) is important to achieve high-quality results.

Several problems arise in practical implementations. First of all we need to know exacly which (x, y) points are involved for each data and tracing all the X-rays is an expensive task. Secondly, real data are corrupted by noise, whose propagation must be faced in the inverse problem resolution.
Earliest commercial softwares were based on an *analitical approach*: takint into account the Fourier Slice theorem (also known as Central Slide theorem), it is possible to skip the heavy step of ray-tracing, on behalf of 1-dimensional Fourier transform of every Radon transform. Moreover, in the frequency domain it is possible to apply suitable smoothing filters and reduce high-frequencies, that tipically enphasyse noise propagatoin. These features define the well-known *Filtered Back Projetion algorithm* that has been widely used and developped for many commercial softwares, in the past decades. See [1] for more details.

3 Solving the inverse problem

3.1 Tomographic linear system

Recently, scientific community is slowly dismissing the analitical approch for tomographic imaging, because it does not provide any longer satisfying reconstructions for many to-

mographic applications. On the other hand, recent studies confirm an increasing interest for *iterative methods* that may be succesfully adapted to any tomographic case. This approach can't avoid the ray-tracing step, because it is exactly based on the discretization of equation (2) and on the inversion of the Radon transform (4), but it offers many different advantages.

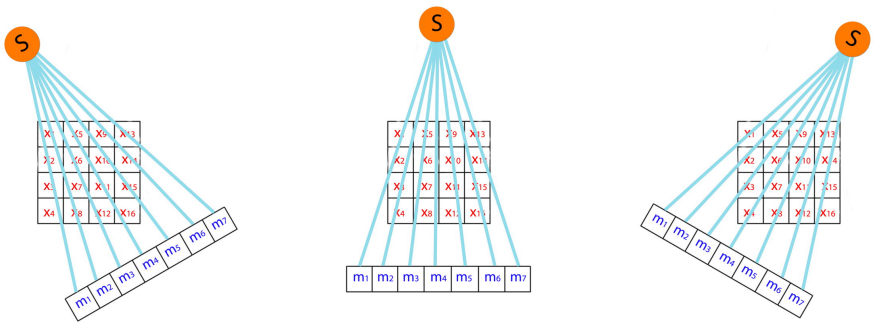


Figure 4. These images show all the elements we need to create the CT linear system for a simulation test with $N_x = N_y = 4$, $N = 16$, $n = 7$ and $\theta_1 = -30°$, $\theta_2 = 0°$, $\theta_3 = +30°$.

Let me fix the scanning angle θ. The i^{th} detector element reveals intensity I_i for the I_0-emitted X-ray, hence we define

$$m_i = -\log\left(\frac{I_i}{I_0}\right)$$

as the projection measurement.
The integral equation (2) (over a single ray L) should be discretized into a finite sum over the crossed voxels of the discretized object $x = (x_1, x_2, \ldots, x_N)$, that we can reorder in a vector shape taking $N = N_x \times N_y$ as desired image resolution. In a more general formulation, we compute

$$m_i = \sum_{j=1}^{N} a_{i,j} x_j \tag{5}$$

taking suitable weights $a_{i,j}$. Such coefficients, in fact, mirrow links between each i^{th} data m_i and each j^{th} voxel x_j. For examples, fixing the first scan in Figure 4 (the negative-angled scan on the left), we should have

$$a_{1,j} = \begin{cases} 1, & j \in \{1,2,3,4\} \\ 0, & \text{otherwise} \end{cases}$$

because the first ray does not hit any voxels x_j, for $j = 5, \ldots, 16$. How to choose and efficiently compute the non-negative coefficients $a_{i,j}$ has not been discussed in this seminario.

According to the detector resolution (let us call it n), every scan consists in a set of n measurements filling a right-hand side term and in n rows of a matrix made of $a_{i,j}$

coefficients, for $i = 1, \dots, n$ (for every scanning angle $\theta_k \forall k = 1, \dots, n_{angles}$).
As the source wheels around the object, from any scan we get further n data and further corrispective n rows, linking the data with the object, according to the geometry of any single acquisition. The resulting linear system is:

$$\begin{bmatrix} \_\_\_\_\_A^{\theta_1}\_\_\_\_\_ \\ \_\_\_\_\_A^{\theta_2}\_\_\_\_\_ \\ \_\_\_\_\_A^{\theta_3}\_\_\_\_\_ \\ \_\_\_\_\_A^{\theta_4}\_\_\_\_\_ \\ \dots \\ \dots \\ \dots \\ \dots \\ \_\_\_\_\_A^{\theta_{n_{angles}}}\_\_\_\_\_ \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \dots \\ \dots \\ \dots \\ \dots \\ \dots \\ \dots \\ \dots \\ x_N \end{bmatrix} = \begin{bmatrix} m^{\theta_1} \\ m^{\theta_2} \\ m^{\theta_3} \\ m^{\theta_4} \\ \dots \\ \dots \\ \dots \\ \dots \\ m^{\theta_{n_{angles}}} \end{bmatrix}$$

It has a large sparse matrix A with N columns and $n \times n_{angles}$ rows (divided in n_{angles} blocks A^{θ_k}), and the known vector contains all the collected data. In addition, A may be both over and under-determined: according to the specific medical invastigation, both structures are feasible.
Let us call b the known vector, to fit the classical notation. Actually speaking, we have

$$Ax = b = b_{exact} + \eta$$

where b_{exact} is the analitical result of Radon integration, while η stands for the noise on the recorded data. Such noise is due both to phisical aspects like photon scattering, and to technical limits like detector sensibility. For this reason, in actual medical imaging both poisson and gaussian noise must be faced.
By the way, the previous linear system is solved through a *minimization problem* of form:

$$\min_{x \in \mathbb{R}^N} \quad \mathcal{F}(A, x, b)$$

where $\mathcal{F}$ is a data-fitting function invoking the three elements A, x and b. A typical choise is the well-known Least Squares function defined as

$$\mathcal{F}(A, x, b) = \|Ax - b\|_2^2$$

Above all in the under-determinated case, a regularization term $\mathcal{R}(x)$ is added to the fitting quantity, in order to make the solution unique. This leads to define the *optimization problem* such that the solution is the minimizer of an objective function given by

$$\min_{x \in \mathbb{R}^N} \quad \mathcal{F}(A, x, b) + \alpha \mathcal{R}(x) \tag{6}$$

In general cases, adding regularization improves the output quality because it forces the solution to have some desiderable features. In fact, $\mathcal{R}(x)$ may include a-priori information that we have and such a possibility is a great advantage on the analitical approach. For instance, priors can be related to the image itself, like asking $x_j \geq 0 \quad \forall j = 1, ..., N$ that is coherence with the phisical assumptions we made on our elements, or to some of its derivative, like asking ∇x to be sparse "enough".

To sum up, we can conclude that analitical methods are fast and simple and they need few algorithmic parameters, but they also provide low-quality reconstructions and don't let the user add significant changes to the algorithms. On the other hand, iterative algorithms compute sequences $\{x^k\}_{k=1,2....}$ of solutions converging to the exact one x^*, so they are computationally expensive and they may take long time to converge to x^*. Moreover they can ask for large storage, too. However, iterative approach lets the user choose data-fitting function and incorporate priors, hence it provide high-quality images.

3.2 Digital Breast Tomosynthesis

Let us see some numerical results focusing on a specific tomographic system: the Digital Breast Tomosynthesis (DBT) that is an emerging 3d- technique, which is slowly replacing the mammographic exam for diagnostic aims. In fact, bi-dimensional imaging suffers for unaccuracy due to the tissue overlapping in the flat representation of a volumetric object: in many cases, mammographic diagnosis can't detect small tumoral masses because of the interposition of glandular and adipose tissues.
To get a final volumetric object instead of a single slice, underlying phisical laws do not change of course: the only difference is that all the CT device components increase one dimension. The emitted beam becomes conical, for example, and the detector will be bidemsional, hence every projection becomes a bidimentional image made of $n_x \times n_y = n$ pixels. In addition, the breast is numerically partitioned into $N_x \times N_y \times N_z = N$ voxels and it will be treated like a stack of slices, but every computation is performed according to the 3d essence of the object of interest.

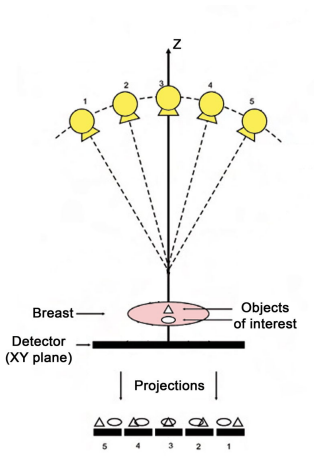


Figure 5. DBT machinery

In particular, DBT is characterised by a limited angular range (up to 30/40 degrees) and by sparse views (i.e. a small n_{angles} parameter, that is around 11 or 15), because it is a diagnostic exams hence the total ammount of radiations per patient must be small. Figure 5 shows a DBT device draft.
Because of the lack of many angled views information, DBT is an example of underdetermined CT system. Moreover, matrix A is very sparse and huge (A is $10^5 \times 10^6$ tipically), hence the regualrization item is necessary to direct the iterative algorithm to an effective solution and, in the meanwhile, it faces up the noise propagation.

The model we fix is

$$\min_{x\in\mathbb{R}^N} \|Ax-b\|_2^2 + \alpha TV(x) \tag{7}$$

where

$$TV(f) := \sum_{j_x=1}^{N_x}\sum_{j_y=1}^{N_y}\sum_{j_z=1}^{N_z} \sqrt{(x_{j_x+1,j_y,j_z}-x_{j_x,j_y,j_z})^2 + \dots}$$

$$\overline{\dots + (x_{j_x,j_y+1,j_z}-x_{j_x,j_y,j_z})^2 + (x_{j_x,j_y,j_z+1}-x_{j_x,j_y,j_z})^2 + \beta^2}$$

is a sort of differentiable 2-norm of the image gradient, thanks to the $\beta > 0$ parameter. The regularization parameter is heuristically set as $\alpha = 0.01$.

Fixed the numerical model, now it is important to choose a suitable iterative method to solve it. In this seminario we focus on numerical results obtained with two different methods: a Scaled Gradient Projection (SGP) algorithm is compared to a Fixed Point based one (FP). The former one is a first order method (see [2]): each iteration is not too expensive and it converges in little time thanks to accelerating strategies. Furthermore, it is possible to add a non-negative constraint to the basic model (7), without losing convergence. The latter one makes use of second order information and it needs on a further inner inverse system resolution for each iteration. Actually, a smart approximation of the Hessian of the objective function is computed as suggested in [3], hence the FP algorithm only requires little more storage than the SGP.

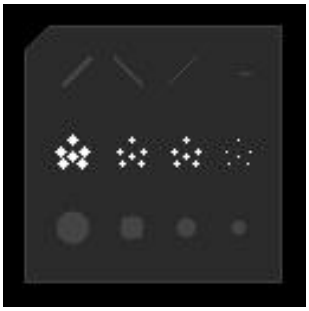


Figure 6. Central slice of the CIRS phantom. It is discretized into 15 horizontal slices made of 128×128 pixels.

Figure 6 shows the central slice of a digital version of a mammographic phantom called CIRS.
In this section fibers, microcalcifications and masses are put in an adipose uniform background, restrained by a voxel-thin skin layer. Such phantom is made of $128 \times 128 \times 15$ voxels and simulations are performed from 13 128×128 projections, acquired in an angular range of [-17, +17] degrees. Gaussian noise has been added to the exact projections, to make simulations more realistic.

The converging results in Figure 7 are the outputs we get imposing the convergence rule

$$\|x_k - x_{k-1}\| < 10^{-4}$$

as a stopping criterion. These slices are shown in reverse gray-scale to better appreciate their quality.

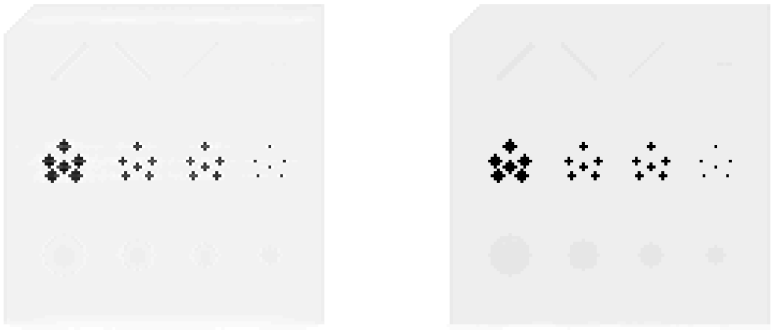


Figure 7. CIRS central slices, obtained by the SGP algorithm (on the left) and the FP one (on the right).

The SGP takes 103 seconds to provide its output and such speed is very appreciable. The FP, instead, takes 505 seconds but its reconstruction is much more accurate: for instance, all the masses are well detected and the phantom edges are more accurate. Both methods mannage in detecting the smallest microcalcifications and face noise very well: these are important confirmations of their capabilities and of the prefixed model.
To deep this comparison, it is convenient also to see how the two algorithms behave in earlier reconstructions. In particular, Figure 8 shows what we are mainly intereressed in: two profiles are taken on the central CIRS layer to analyse the horizontal resolution for microcalcifications and for masses, while two vertical profiles are taken to establish if the algorithms are able to detect objects in their actual depth.

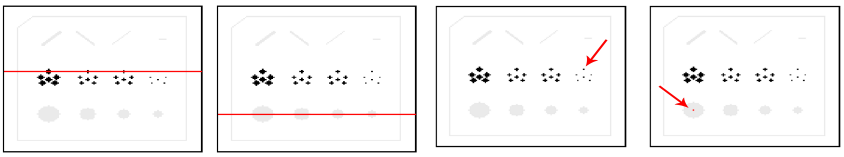


Figure 8. Elements of interest on the central slice: horizontal profiles over microcalcifications and over masses on the eighth layer, to invastigate the xy-resolution, and vertical profiles passing through a voxel-thin microcalcification and through the biggest mass, to investigate the z-resolution.

Figures 9 and 10 show these profiles on the outputs we get, running the reconstructing softwares for 5, 20 and 60 seconds and up to the convergence.
The SGP speed is noticeable on the earliest outputs (5 second ones are in the first row), where the FP lines are still too close to their initial zeros values. This feature makes this algorithm a valid candidate for commercial systems, which perform only few iterating steps for each patient. On the other hand, as time goes by, the FP reconstruction becomes more and more accurate and it recovers all the objects with their actual intensities, while the SGP provides smaller values and inexact edges.

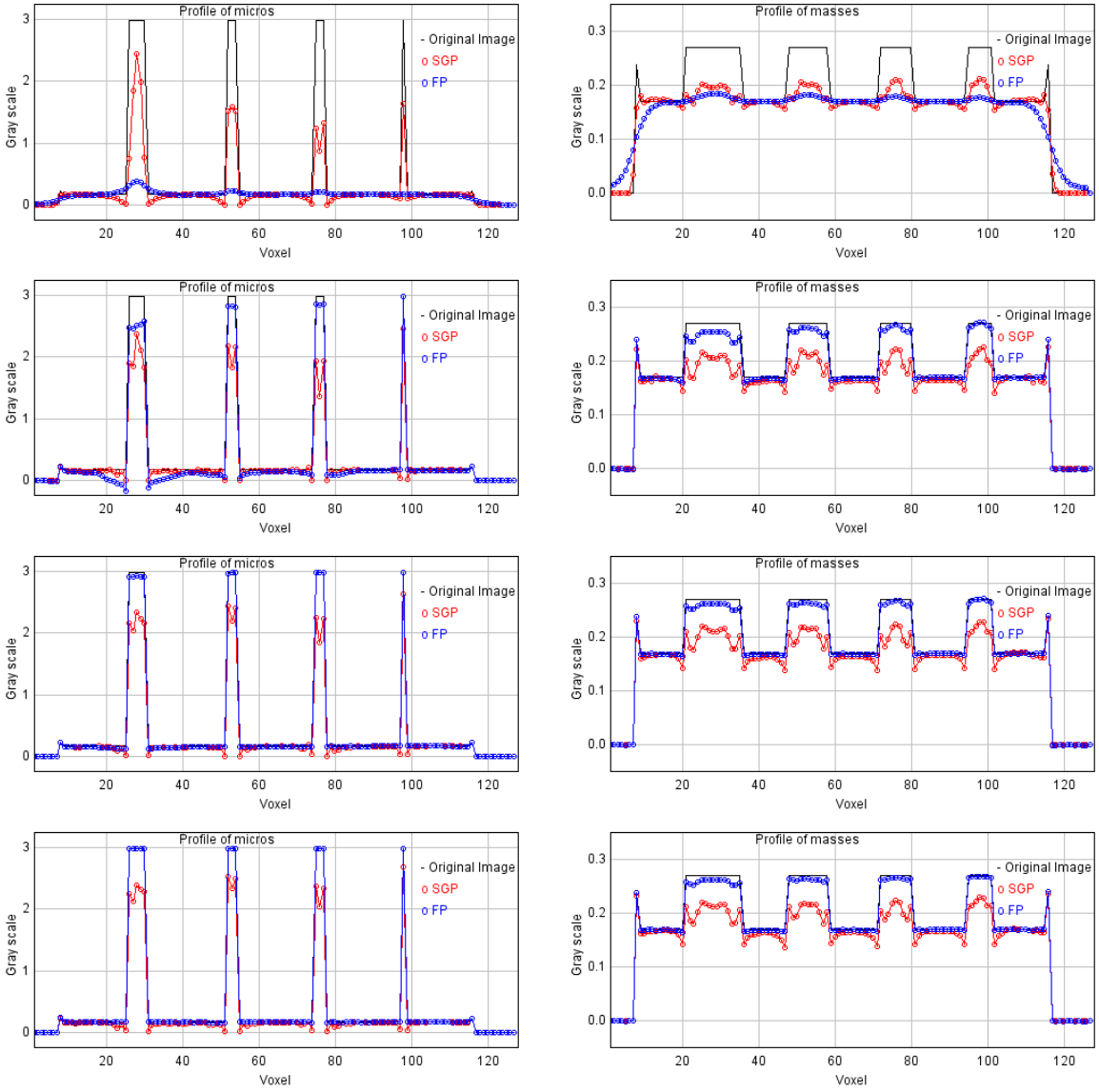


Figure 9. Plots of the profiles over microcalcifications on the left, over masses on the right.

Focusing on the third dimension, Figure 10 reflects the most important divergence between the two methods: while a small but high-value object is always exactly placed in its real position, the three layer thik mass is spread in all the breast highness with softer values by the SGP algorithm.

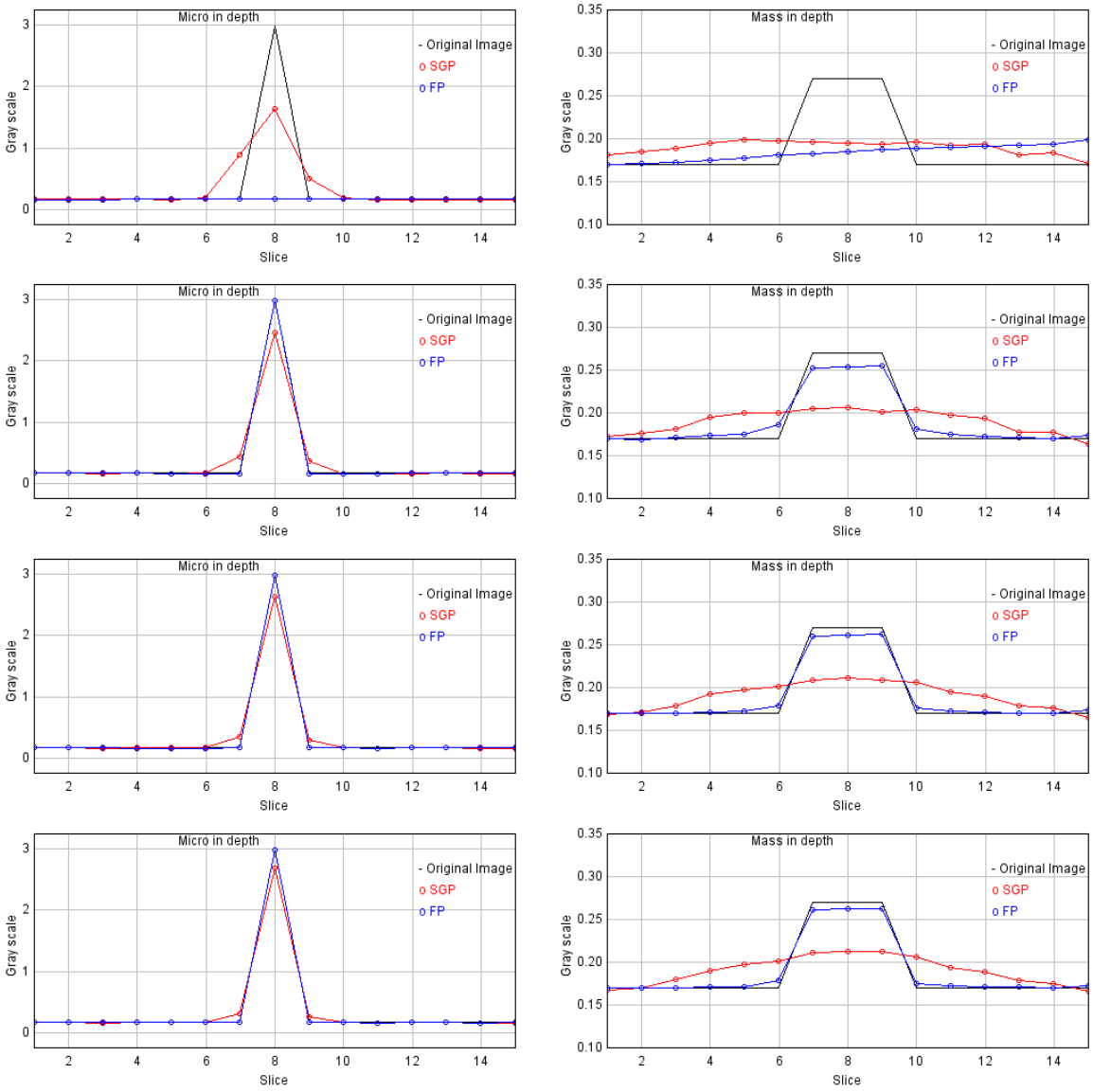


Figure 10. Plots of the vertical profiles including a voxel-thin microcalcification on the left, or a masse on the right.

4 Conclusions

To conclude, in this seminario dottorato we have seen basic notions about the tomographic proceeding underlying the well-known medical exam, and a mathematical description of such process. After that, we have understood the intrinsic meaning of the resulting linear system and set a suitable model for the optimization problem we need to solve. At the end, the DBT technique has been introduced and two different methods have been compared on a DBT simulation, showing how difficult the tomographic imaging problem could be.

References

[1] Kak, A. C., "Principles of Computed Tomographic Imaging". IEEE PRESS, 1988.

[2] Bonettini, S., Zanella, R. and Zanni, L., *A scaled gradient projection method for constrained image deblurring*. Inverse Problems, 25 (2009).

[3] Vogel, C. R., "Computational methods for Inverse Problems". SIAM, 2002.