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

Scuole di Dottorato in Matematica Pura e Computazionale

Seminario Dottorato 2023/24

Preface

This document offers an overview of the activity of Seminario Dottorato 2023/24.

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

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

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

Padova, June 20th, 2024

Corrado Marastoni, Tiziano Vargiolu

Abstracts (from Seminario Dottorato's webpage)

Wednesday 4 October 2023

An Introduction to Non-Connected Linear Algebraic Groups and their partition into Jordan Classes and Lusztig Strata

Martina COSTA CESARI (Padova, Dip. Mat.)

Linear algebraic groups are a class of mathematical structures that combine concepts from algebra and geometry. This suggests that algebraic groups can be approached from different perspectives, such as Group Theory, Algebraic Geometry, and Combinatorics. They have applications in several directions (Invariant Theory, Physics). Linear algebraic groups are affine varieties with a compatible group structure. They were introduced in the late 1800s to study continuous symmetries of differential equations. An important class of algebraic groups consists of non connected algebraic groups. In the first part of the talk I will introduce basic notions and examples of linear algebraic groups, and in particular of non connected linear algebraic groups. The last part of the presentation is devoted to explore some partition of these objects, in particular Jordan classes and Lusztig strata, and investigating their geometric properties.

Wednesday 8 November 2023

A 2D Bin Packing Problem in the Sheet Metal Industry: models and solution approaches

Chiara TURBIAN (Padova, Dip. Mat. with Salvagnini Italia S.p.A., Sarego, Italy)

The Bin Packing Problem (BPP) is a well-studied problem in Operations Research, and, in its basic formulation, it aims at packing a set of items into a finite set of bins by minimizing the number of used bins. Due to its wide range of applications, several variants of the problem have been proposed during the last decades, which differ from each other by dimensionality, additional constraints, and characteristics of the items or the bins. We consider a Two-Dimensional Bin Packing Problem (2DBPP) arising in Salvagnini Italia, a multinational corporation working in the sheet metal industry. In our problem, the basic 2DBPP is enriched by the presence of technological constraints emerging from the context, such as precedence relations between groups of items and conditional safety distances between items. We present exact and heuristic approaches to solve the problem, both based on Mixed Integer Linear Programming (MILP), and we show related computational results.

Wednesday 22 November 2023

Groups and geometry: from algebraic varieties to Galois representations and vice versa

Khai Hoan NGUYEN DANG (Padova, Dip. Mat.)

Since a very long time ago, there has been an effective approach to study geometry via group theory. In this talk, we will focus on objects given by sets of solutions of a system of polynomial equations, called algebraic varieties. Galois theory makes a bridge between the geometry of algebraic varieties and group theory in terms of Galois representations. The talk will survey some basic but still interesting aspects of these connections and provide several examples. We will also provide a uniform way to investigate a certain class of algebraic varieties, named Abelian varieties.

Wednesday 13 December 2023

Wasserstein Generative Models

Amna MOHSIN (Padova, Dip. Mat.)

In this seminar firstly, I will present an introduction about optimal transport. Nowadays optimal transport importance extends to diverse domains, ranging from mathematics and computer science to economics and image processing. Subsequently, I will talk about the Wasserstein distance, particularly the W1 distance, which is a powerful metric for measuring the dissimilarity between probability distributions and it provides a more stable and meaningful measure than traditional metrics like the Kullback-Leibler divergence. This metric is used in particular within generative models, which are modern deep learning techniques that may be used to generate objects such as images, text, or any other structure. I will introduce these models and explain their application domain and discuss their properties, especially in relation to the limitation in their use of the W1 distance. Then, I will talk about the main objective of the thesis, which builds upon prior work which introduce a dynamics based method that allows us to obtain very accurate computations of the Wasserstein distance. The objective is to apply this method effectively within generative models to overcome the limitations in the traditional methods used to compute the W1 distance, and how we expect this method to improve the models performances.

Wednesday 20 December 2023

Double-negation in the Foundation of Constructive Mathematics

Pietro SABELLI (Padova, Dip. Mat.)

In this talk, we will first introduce the fundamental ideas of Constructive Mathematics through basic examples taken from ordinary mathematical practice and focus on its computational aspect. Secondly, we will review how the logical principle of the Excluded Middle, dating back to Aristotle, and, more generally, the concept of negation play a crucial role in distinguishing Constructive Mathematics from Classical Mathematics. Finally, we will give a non-technical overview of Gödel's double-negation interpretation of arithmetic and present our new result, which generalises it to the "Minimalist Foundation", a foundation for constructive mathematics designed in Padua by M.E. Maietti and G. Sambin.

Wednesday 24 January 2024

Digital twins: a general overview and the application to bread baking

Laura RINALDI (Padova, Dip. Mat.)

A digital twin is composed of two existing systems: the tangible system of physical reality and its virtual and numerical replica which is enabled by real data and underling models through the underlying use of digital technologies. The presence of digital twins is motivated by the necessity of obtaining some information about the real system questioning the virtual one by a non-intrusive manner. Such technology helps us to monitor the real system, to carry out maintenance tasks or optimize some process. In this talk, I will present an industrial application which consists in the building of an embedded digital twin of the bread baking process to the end of monitoring the energy consumption to avoid waste.

Wednesday 31 January 2024

PDE's and Conservation Laws: from the basics to current research

Luca TALAMINI (Padova, Dip. Mat.)

In this talk I will try to introduce you to the world of PDE's in general and conservation laws in particular. In the first part of the talk we will focus on rather classical topics. Via a lot of examples I will try to give you a feeling of what a PDE is really about and what it means "to solve it". In the last part we take a look at conservation laws (a particular class of PDE's). Besides being my current main research topic, conservation laws provide me with a great tool to illustrate modern challenges in the field of non-linear PDE's.

Thursday 15 February 2024

A differential game model for sponsored content

Chiara BRAMBILLA (Padova, Dip. Mat.)

Let us consider a communication platform distinguished for its high-quality content, where advertising can take two different forms: traditional and sponsored (also known as native advertising in the marketing literature). Native advertising is a widely used marketing tool that aims to mimic the regular topics of the platform on which it is placed. Due to this striking resemblance, native advertising may be very effective, but at the same time, it may negatively influence the perceived credibility of the media outlet. In our model, a firm allocates investments to both traditional and native advertising on such a platform. Meanwhile, the media outlet must grapple with the trade-off between the profit accrued from publishing native advertising and the ensuing decline in credibility. We formalise this problem as a hierarchical infinite-time horizon linear state differential game, played a‘ la Stackelberg, where the media outlet acts as the leader while the firm is the follower. Finally, we characterise a time-consistent open-loop equilibrium and obtain the conditions that make it optimal for the media outlet to accept native advertising.

Wednesday 28 February 2024

Classical Modular Forms and the k-square problem

Marco BARACCHINI (Padova, Dip. Mat.)

Modular forms are objects belonging to the world of complex analysis. They are holomorphic functions with some transformation properties. In this talk I will introduce two arithmetic problems that could be studied using the theory of modular forms. In the second part of the talk we will see the definition of modular forms and some classical results in this area.

Wednesday 13 March 2024

p-adic numbers and characteristic p

Pietro VANNI (Padova, Dip. Mat.)

For each prime p, p-adic numbers form an extension of the rational numbers that, being topologically complete, allows one to use analytic methods in arithmetic. In this talk I will introduce p-adic numbers outlining their basic properties and the role they play in number theory. Then I will give an idea on how one can employ p-adic numbers to study algebraic varieties (i.e. systems of polynomial equations) in characteristic p.

Wednesday 10 April 2024

A sphere rolling on a plane: a journey into nonholonomic mechanics

Mariana COSTA VILLEGAS (Padova, Dip. Mat.)

The problem of the sphere rolling on a plane is one of the most classical examples of nonholonomic mechanical systems. I will use this example to give a brief introduction to this kind of systems, their properties, and the main tools that are used to study them which go from geometry and dynamics to symmetries and Lie groups. We will then consider classical and new affine variations of this problem and see that the dynamics ranges from integrable to chaotic depending on the specifics of the system. Finally, I will discuss some curious and surprising phenomena occurring in specific examples.

Wednesday 24 April 2024

Collective periodic behaviors in large-volume interacting particle systems

Elisa MARINI (Padova, Dip. Mat.)

In the first part of this seminar, we will give an overview of collective periodic behaviors in large systems of interacting components. Loosely speaking, such phenomena consist in nearly-periodic oscillations which characterize the long-time dynamics of some macroscopic quantity of the system, and which cannot be ascribed to any external periodic force applied to the system, nor to any oscillatory behavior of its components, but rather arise from the interaction among these latter.

Although they are ubiquitous in real-world systems (they are observed for instance in neural networks, predator-prey dynamics, epidemiology), such behaviors are still poorly understood from a theoretical standpoint. In the second part of the seminar, we will present a toy model of interacting diffusions displaying collective oscillations. This will serve as an example of the mechanisms which may originate collective periodic behaviors and to give an idea of the mathematics involved in the rigorous study of such phenomena.

Thursday 9 May 2024

Differential games and large population limits beyond the classic Mean-Field setting

Davide Francesco REDAELLI (Padova, Dip. Mat.)

Differential game theory is a branch of mathematics that touches many fields such as control and game theories, probability, stochastic and partial differential equations. An interesting aspect of it is studying strategies of the players which are optimal in that they produce a situation of equilibrium, for example in the famous sense due to Nash, and also seeing what happens when the number of players grows and possibly becomes infinite. In this talk I will try to give a brief introduction to this theory aimed at a wide audience of mathematicians possibly unaware of the subject, with the final purpose of presenting the main topics which my doctoral research focused on.

Thursday 23 May 2024

Numerical Solution of Wave Propagation Phenomena in Viscoelastic Materials

Nicolò CRESCENZIO (Padova, Dip. Mat.)

Many materials, such as plastics, wood, concrete and metals at high temperatures, exhibit a mechanical behaviour that is intermediate between the elastic and the viscous one. Consequently, these materials cannot be adequately described using the well-known classical theories of elasticity and viscosity and it is therefore necessary to consider a more general theory that is capable of modelling the behaviour of these materials, also known as viscoelastic materials. In the first part of the talk, we will provide a brief overview of the theory of linear viscoelasticity, with a particular focus on the so-called Kelvin-Voigt rheology. Then, we will discuss the problem of viscoelastic wave propagation phenomena in a Kelvin-Voigt heterogeneous material and show numerical results obtained by means of a Galerkin spectral approach.

Wednesday 5 June 2024

Strichartz estimates for the Dirac equation in different settings

Elena DANESI (Padova, Dip. Mat.)

The Dirac equation is a first order partial differential equation. It was first derived by Paul Dirac in 1928 in order to describe the free motion of a spin 1/2 particle on $\mathbb{R}^3$, in according with the principles of quantum mechanics and special relativity. In the following years its definition has

been generalized in order to be adapted to curved backgrounds. From the mathematical side, it can be listed within the class of dispersive equations, together with the Schrödinger, wave and Klein-Gordon equations. In the years, because of the study of nonlinear systems, a lot of effort has been devoted to developing tools to quantify the dispersion of a system. Among these tools we find a priori estimates on the solutions, such as Strichartz or local smoothing estimates. In the first part of the talk I will focus on the Schrödinger and wave equations in order to present these kind of estimates as well as some classical tools to prove them. In the second part, I will first introduce the Dirac equation on $\mathbb{R} \times \mathbb{R}^3$ and describe its connection with the above mentioned equations. I will then present the equation in curved spacetimes. To conclude, I will survey some recent results concerning the validity of Strichartz estimates for the "curved" Dirac equation in specific settings, in particular compact or asymptotically flat manifolds.

Thursday 20 June 2024

Topological Data Analysis (TDA): basic concepts and applications

Cinzia BANDIZIOL (Padova, Dip. Mat.)

In the last two decades, with the ever higher increasing amount of data of many kinds and, usually, of high dimension, it has revealed meaningfull to be able to extract new and additional infomation from data, overall if it is related to intrinsic properties of themselfes. It has motivated the birth of a new field of research, the so called Topological Data Analysis. Thanks to the strong theoretical basis of algerabraic topology, it allows to extract qualitative information from dataset, as point clouds, images, graphs, time series, ecc., related to the "shape of data", and to use them into machine learning and deep learning frameworks. In this talk, first we will introduce the main tool of TDA called persistent homology with basic definitions and notions. Then, after the introduction of the classification problem, we will discuss how to use the new topological information in such a context.

An Introduction to Non-Connected Linear Algebraic Groups and their partition into Jordan Classes and Lusztig Strata

Martina Costa Cesari (*)

Abstract. Linear algebraic groups are a class of mathematical structures that combine concepts from algebra and geometry. This suggests that algebraic groups can be approached from different perspectives, such as Group Theory, Algebraic Geometry, and Combinatorics. They have applications in several directions (Invariant Theory, Physics). Linear algebraic groups are affine varieties with a compatible group structure. They were introduced in the late 1800s to study continuous symmetries of differential equations. An important class of algebraic groups consists of non connected algebraic groups. In the first part of the talk I will introduce basic notions and examples of linear algebraic groups, and in particular of non connected linear algebraic groups. The last part of the presentation is devoted to explore some partition of these objects, in particular Jordan classes and Lusztig strata, and investigating their geometric properties.

1 Preliminars on affine varieties

In this section we recall some classical facts about algebraic geometry, for a complete description one can see [8].

Let $\mathbb{K}$ be an algebraically closed field. We consider $\mathbb{K}[T_1, \dots, T_n]$ the ring of polynomials in n indeterminates with coefficients in $\mathbb{K}$.

Definition 1.1 Let $I \subseteq \mathbb{K}[T_1, \dots, T_n]$ be an ideal. The *algebraic set* associated with I is

$$V(I) = \{v \in \mathbb{K} \mid f(v) = 0 \text{ for every } f \in I\} \subseteq \mathbb{K}^n.$$

Proposition 1.2 *The collection of algebraic sets fulfill the axioms of the closed subsets of a topology on* $\mathbb{K}^n$.

Proof. We observe that $\mathbb{K}[T_1, \dots, T_n] = V((0))$ and $\emptyset = V(\mathbb{K}[T_1, \dots, T_n])$, then $\mathbb{K}^n$ and $\emptyset$ are closed.

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: martina.costacesari@math.unipd.it . Seminar held on 4 October 2023.

Moreover the collection of affine algebraic sets is closed by taking arbitrary intersections: if $\{X_j\}_{j\in J}$ is a family of algebraic sets with defining ideals $\{I_j\}_{j\in J}$, then $\bigcap_{j\in J} X_j = V\left(\sum_{j\in J} I_j\right)$, where $\sum_{j\in J} I_j := \{\sum_{k=1}^{m} f_k \mid m \in \mathbb{N}, f_k \in I_j \text{ for some } j \in J\}$ is an ideal in $\mathbb{K}[X_1, \ldots, X_n]$.
Furthermore the collection of algebraic sets is closed by taking finite unions: if X, Y are affine algebraic sets with defining ideals I_X and I_Y, then $X \cup Y = V(I_X I_Y)$, where $I_X I_Y := \{fg \mid f \in I_X, g \in I_Y\}$ is an ideal in $\mathbb{K}[X_1, \ldots, X_n]$. Therefore all the axioms are fulfilled. □

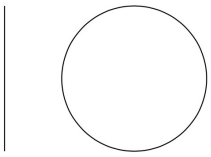
Definition 1.3 The topology on $\mathbb{K}^n$ given by taking $V(I)$, with I an ideal in $\mathbb{K}[T_1, \ldots, T_n]$, as closed subsets, is called *Zariski topology*. We denote by $\mathbb{A}^n_{\mathbb{K}}$ the vector space $\mathbb{K}^n$ endowed with the Zariski topology. The topological space $\mathbb{A}^n_{\mathbb{K}}$ is called the n-dimensional affine space over $\mathbb{K}$.

Example 1.1

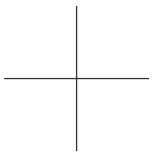
(a) Any point $P = (x_P) \in \mathbb{A}^1_{\mathbb{C}}$ on the affine complex line is completely determined by its coordinate x_P. It is an algebraic set, as $\{P\} = V((X - x_P))$. Similarly, any finite collection of points on the affine complex line is an algebraic set.

(b) The following are algebraic sets in the complex affine plane $\mathbb{A}^2_{\mathbb{C}}$ with coordinates T_1, T_2.

$V(T_2^2 - T_1^3 - T_1^2)$



$V((T_2^2 - T_1^2 - 1)(T_1 + 2)$



$V(T_1 T_2)$

Definition 1.4 An *affine variety* over $\mathbb{K}$ is a closed subset of $\mathbb{A}^n_{\mathbb{K}}$ for some $n \in \mathbb{N}$.

1.1 Morphism of affine varieties

Definition 1.5 Let $X \subseteq \mathbb{A}^m_{\mathbb{K}}$ and $Y \subseteq \mathbb{A}^n_{\mathbb{K}}$ be affine varieties. A map
$\phi : X \longrightarrow Y$ is a morphism of affine varieties, if it is the restriction of a map

$$\tilde{\phi} : \mathbb{A}^m_{\mathbb{K}} \longrightarrow \mathbb{A}^n_{\mathbb{K}}$$

$$(x_1, \ldots, x_m) \mapsto \begin{pmatrix} \phi_1(x_1, \ldots, x_m) \\ \vdots \\ \phi_n(x_1, \ldots, x_m) \end{pmatrix}$$

with $\phi_i \in \mathbb{K}[T_1, \ldots, T_m]$ for all $i = 1, \ldots, n$.

Example 1.2 We give two easy examples of algebraic morphisms.

- The maps $\alpha, \mu : \mathbb{A}^n_{\mathbb{K}} \to \mathbb{A}^1_{\mathbb{K}}$ defined by $\alpha(x_1, \dots, x_n) = \sum_{i=1}^n x_i$ and $\mu(x_1, \dots, x_n) = \prod_{i=1}^n x_i$ are morphisms.
- The projection on the i-th component $\pi_i : \mathbb{A}^n_{\mathbb{K}} \to \mathbb{A}^1_{\mathbb{K}}$ defined by $(x_1, \dots, x_n) \mapsto x_i$ is a morphism, for $i \in \{1, \dots, n\}$.

2 Linear algebraic groups

Definition 2.1 A *linear algebraic group* is an affine algebraic variety G with a group structure such that:

$$\mu : G \times G \to G \quad (x, y) \mapsto xy \quad \text{and} \quad \iota : G \to G \quad x \mapsto x^{-1}$$

are morphisms of algebraic varieties.

Example 2.1

- Consider the additive group $G = (\mathbb{C}, +)$. This is a linear algebraic group, indeed the maps
$$\begin{array}{rl} \mu : G \times G & \to G \\ (x_1, x_2) & \mapsto x_1 + x_2 \end{array} \qquad \begin{array}{rl} \iota : G & \to G \\ x & \mapsto -x \end{array}$$
are clearly morphisms of affine varieties.
- The multiplicative group $\mathrm{GL}_1(\mathbb{C})$ of invertible complex numbers $\mathbb{C}^*$ is a linear algebraic group.

Proposition 2.2 *The general linear group $GL_n(\mathbb{K})$, consisting of $n \times n$ invertible matrices with coefficients in $\mathbb{K}$, is a linear algebraic group.*

Proof. Let $M_{n\times n}$ be the space of $n \times n$ matrices with coefficients in $\mathbb{K}$. We can identify $M_{n\times n}$ with the affine space $\mathbb{A}^{n^2}_{\mathbb{K}}$. The determinant of a matrix is the polynomial in the indeterminates T_{ij}, with $1 \le i, j \le n$:

$$\det = \sum_{\sigma \in \mathfrak{S}_n} \mathrm{sgn}(\sigma) \prod_{i=1}^n T_{i,\sigma(i)}.$$

We have

$$\mathrm{GL}_n(\mathbb{K}) = \left\{ x = \left((x_{ij})_{i,j}, t \right) \in \mathbb{A}^{n^2+1}(\mathbb{K}) \mid \det\left((x_{ij})_{i,j} \right) t - 1 = 0 \right\}.$$

Thus, given the ideal $I = (\det\left((x_{ij})_{i,j} \right) t - 1) \subseteq \mathbb{K}[T_{i,j}, t]$, the general linear group $\mathrm{GL}_n(\mathbb{K}) = V(I)$, so it is an affine variety. Moreover if $A, B \in \mathrm{GL}_n(\mathbb{K})$, then each entry of the product AB can be expressed as a polynomial functions in the entries of A and

B. Similarly, the entries of A^{-1} can be expressed by means of Laplace's rule as polynomial functions in the entries of A and of $(\det A)^{-1}$. So the multiplication map and the inverse map are morphism of affine varieties. □

Remark 2.3 The general linear group $\mathrm{GL}_n(\mathbb{K})$ is connected.

Example 2.2 Every closed subgroup of $\mathrm{GL}_n(\mathbb{K})$ is a linear algebraic group. For example the orthogonal group of order n over $\mathbb{K}$ that is

$$G = \mathrm{O}_n(\mathbb{C}) = \left\{A \in \mathrm{GL}_n(\mathbb{C}) \mid A^T A = 1\right\},$$

is a linear algebraic group. Moreover $\mathrm{O}_n(\mathbb{C})$ is non-connected, in fact it has two connected components, one containing the matrices with determinant equal to 1, and the other containing the matrices with determinant equal to -1, namely

$$G^\circ = \mathrm{SO}_n(\mathbb{C}) = \{A \in G \mid \det A = 1\} \text{ and } (-1)G^\circ$$

So far we have seen examples of linear algebraic groups, given as groups of matrices, this is not a case. In fact we have the following theorem.

Theorem 2.4 [8, Theorem 2.3.7] *Let G be a linear algebraic group over $\mathbb{K}$. Then, for some $n \in \mathbb{N}$ there exists an embedding*

$$\rho : G \hookrightarrow \mathrm{GL}_n(\mathbb{K}).$$

In particular G is isomorphic to a closed subgroup of $\mathrm{GL}_n(\mathbb{K})$.

2.1 Jordan decomposition

This section is devoted to one of the founding instruments to study linear algebraic groups. We recall the definition of semisimple and unipotent matrices.

Let $A \in \mathrm{GL}_n(\mathbb{K})$. We say that A is *semisimple* is it is diagonalizable, i.e. if there exists a basis $\{v_1, \ldots, v_n\}$ of $\mathbb{C}^n$ and $\lambda_i \in \mathbb{C}^\times$such that $Av_i = \lambda_i v_i$ for all $i = 1, \ldots, n$.

The matrix A is called *unipotent* if $A - 1$ is nilpotent, i.e. if there exists $k \in \mathbb{N}$ such that $(A-1)^k = 0$.

From the classical Jordan normal form, up to conjugation, we can see A as the sum of A_s, a semisimple matrix (namely the diagonal part of the Jordan normal form), and A_n, a nilpotent matrix (namely the upper diagonal part of the Jordan normal form). These matrices, A_s, A_n, commute.

From the additive decomposition $A = A_s + A_n$, we can obtain a multiplicative decomposition

$$A = A_s(\mathrm{Id} + A_s^{-1} A_n),$$

where A_s and $(\mathrm{Id} + A_s^{-1} A_n)$ commute. We have that $(\mathrm{Id} + A_s^{-1} A_n)$ is a unipotent element that we denote by A_u. So we have proven the following proposition.

Proposition 2.5 *For every $A \in \mathrm{GL}_n(\mathbb{K})$ there exist a decomposition $A = A_s A_u$ with A_s semisimple, A_u unipotent, and such that A_s, A_u commute.*

We can generalize this result to every linear algebraic group.

Definition 2.6 Let G be a linear algebraic group. An element $g \in G$ is called *semisimple* if for any embedding $\rho : G \hookrightarrow \mathrm{GL}_n(\mathbb{K})$, its image $\rho(g)$ is a semisimple matrix. Analogously $g \in G$ is called *unipotent* if $\rho(g)$ is a unipotent matrix.

Theorem 2.7 [8, Theorem 2.4.8] *Let G be a linear algebraic group. For every element $g \in G$ there exist a semisimple element $g_s \in G$ and a unipotent $g_u \in G$ such that*

$$g = g_s g_u = g_u g_s.$$

Remark 2.8 The identity is the only element which is both semisimple and unipotent.

Example 2.3 In $G = \mathrm{GL}_2(\mathbb{C})$ consider $A = \begin{pmatrix} \alpha & 1 \\ 0 & \alpha \end{pmatrix}$. Then $A = A_s A_u$, with $A_s = \begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix}$ and $A_u = \begin{pmatrix} 1 & \alpha^{-1} \\ 0 & 1 \end{pmatrix}$. Hence A is unipotent if and only if $\alpha = 1$, otherwise it is nor semisimple nor unipotent.

We are interested in study a particular class of linear algebraic group.

Definition 2.9 Let G be a connected linear algebraic group. Then G is called *reductive* if its maximal closed unipotent normal subgroup is trivial.

Example 2.4 The group SL_n, that is the group of invertible $n \times n$-matrices with determinant equal to 1, is reductive.

2.2 The Weyl group

In this section we introduce the notion of Weyl group associated with a connected reductive algebraic group.

Definition 2.10 A linear algebraic group is called a *torus* if it is isomorphic to D_n, the subgroup of GL_n consisting of diagonal matrices, for some $n \in \mathbb{N}$.

Let G be a connected reductive algebraic group. Let T be a subgroup of G such that T is a torus. We denote by $N_G T$ the normalizer in G of T, and by $C_G(T)$ the centralizer of T in G.

Definition 2.11 The *Weyl group* of G is the group $W := N_G T / C_G(T)$.

By [7, Theorem 3.10], the group W is finite.

Example 2.5 We consider the subgroup of diagonal matrices T in SL_n. The subgroup T is a torus. The Weyl group $W = N_G(T)/T$ is isomorphic to $\mathbb{S}_n$, the n-symmetric group.

2.3 Non-connected linear algebraic groups

In this section we deal with the structure of non connected algebraic groups. Let G be an algebraic group. We denote by G° the connected component of G containing the identity.

We recall the following result [8, Proposition 2.2.1].

Proposition 2.12 *The connected component G° is a closed normal subgroup of finite index. Moreover any closed subgroup of G of finite index contains G°.*

Let $x \in G$ and consider the multiplication morphism

$$\mu_x : G \longrightarrow G$$

$$g \mapsto gx.$$

Since μ_x is a homeomorphism, then the image of connected components are connected components, so $\mu_x(G^\circ) = G^\circ x$ is a connected component of G. Let D be the connected component of G containing x, then $x \in G^\circ x \cap D$, therefore $D = G^\circ x$.

By Proposition 2.12, we have that G/G° is a finite group of order m for some $m \in \mathbb{N}$. Then G has finitely many connected components.

For $x \in G$, we consider the automorphism of G

$$\begin{aligned} c_x : G &\longrightarrow G \\ g &\mapsto xgx^{-1} \end{aligned} \tag{1}$$

We refer to c_x as the *conjugation morphism* by x, and we denote it by τ. It restricts to an automorphism of G°.

Let $\widetilde{G} = G^\circ \rtimes \langle \tau \rangle$ be an algebraic group with the following operation

$$g_1\tau^{k_1} g_2 \tau^{k_2} = g_1 \tau(g_2)^{k_2} \tau^{k_1+k_2}.$$

We are interested in study the action of G° on a connected component D of G, given by the conjugation. For this reason it is not restrictive to study the group of the form as $\widetilde{G}$. Indeed we have the following result.

Lemma 2.13 *The varieties $G^\circ\tau$ and $G^\circ x$ are isomorphic as G°-variety, where G° acts by conjugation.*

Proof. The isomorphism

$$\rho : G^\circ\tau \longrightarrow G^\circ x\,, \qquad\qquad h\tau \mapsto hx$$

is G°-equivariant. Indeed, let $g_0 \in G^\circ$, then

$$\rho(g_0 h \tau g_0^{-1}) = \rho(g_0 h \tau(g_0^{-1})\tau) =$$

$$= g_0 h \tau(g_0^{-1})x = g_0 h c_x(g_0^{-1})x = g_0 h x g_0^{-1} x^{-1} x = g_0 h x g^{-1} = g_0 \rho(h\tau) g_0.$$

□

Example 2.6 Let $n \geq 3$ and τ be the involution of $SL(n)$, defined as follows

$$\tau : SL(n) \longrightarrow SL(n)$$

$$X \mapsto {}^tX^{-1},$$

with ${}^tX^{-1}$ the transposed of the inverse of X.
The group $SL(n) \rtimes \langle \tau \rangle$ is a non-connected algebraic group with two connected components

$$G^\circ = SL(n) \text{ and } G^\circ\tau.$$

3 Jordan classes on non-connected algebraic groups

In [3] G. Lusztig introduced a finite stratification of a non connected reductive algebraic group G. We refer to the strata as Jordan classes.

Jordan classes have good geometrical properties, they are locally closed, irreducible and smooth subvarieties of G.

Each Jordan class is a union of G°-conjugacy classes of the same dimension, where G° is the connected component of G containing the identity. Moreover the closure of a Jordan class is a union of Jordan classes.

In the following if H is a subgroup of G we denote by H° the connected component of H containing the identity. For $h \in H$, we denote tby $C_H(h)$ the centralizer of h in H. The center of H is denoted by $Z(H)$.

Let a be an element of G with Jordan decomposition $a = a_s a_u$. Following [3, 2.1], we set

$$T(a) = (Z(C_G(a_s)^\circ) \cap C_G(a_u))^\circ = (Z(C_{G^\circ}(a_s)^\circ) \cap C_{G^\circ}(a_u))^\circ \tag{2}$$

We consider the equivalence relation on G:

$$a \sim h \text{ if } \exists x \in G^\circ \text{ such that } T(xhx^{-1}) = T(a) \text{ and } xhx^{-1} \in T(a)a.$$

Definition 3.1 A *Jordan class* is an equivalence class for the relation in (2), and we denote the Jordan class containing a by $J(a)$.

Example 3.1 Let $\mathrm{SO}_8(\mathbb{K})$ be the group of 8×8 matrices of determinant equal to 1 that leave invariant the bilinear form whose matrix with respect to the canonical basis in $\mathbb{K}^8$ is $\begin{pmatrix} 0 & I_4 \\ I_4 & 0 \end{pmatrix}$. We consider the two matrices $s = \mathrm{diag}\,(-1, I_3, -1, I_3)$ and $r = \mathrm{diag}\,\big(t, I_3, t^{-1}, I_3\big)$ for any $t \neq 0, \pm 1$ are in the same Jordan class.

4 Lusztig strata on non-connected algebraic groups

In this section we introduce the notion of Lusztig strata of a non connected reductive algebraic group given in [5].

Let G be a reductive algebraic group. We denote by G° the connected component of G containing the identity, and we denote by W its Weyl group.

Using a variant of the Springer's correspondence with trivial local system, Lusztig defined a map $\mathcal{E}$ from a connected component D of G to the set of irreducible representations of a subgroup of W depending on D.

4.1 The Springer's correspondence

The classical Springer correspondence with trivial local system is a map from the set of unipotent elements in a connected reductive algebraic group to the set of isomorphism classes of irreducible representation of its Weyl group [9].

Example 4.1 We consider $G = \mathrm{S}L_n$. The Weyl group of G is the symmetric group $\mathbb{S}_n$.

The set of unipotent orbits in G is parameterized by the partitions of n. Moreover, the set of isomorphism classes of irreducible representation of $\mathbb{S}_n$ is parameterized also by the partitions of n. Therefore the Springer correspondence in this case is a bijection of the set of partitions of n.

The Springer's correspondence in general is not bijective.

4.2 The map $\mathcal{E}$

We recall here the definition of the map $\mathcal{E}$ given in [5].

In the following we retain the notation of [2].

We recall that the Weyl group W is in bijection with the set of G°-orbits on $\mathcal{B} \times \mathcal{B}$, where $\mathcal{B}$ is the flag manifold of G° and where G° acts diagonally by conjugation. Also the group G acts diagonally on $\mathcal{B} \times \mathcal{B}$: this induces an action of G/G° on W. Let D be a connected component of G. We can see D as an element of G/G°. So this defines an automorphism $[D] : W \to W$ whose fixed point set is denoted by W^D. By [4, Appendix], W^D is a Coxeter group.

The map $\mathcal{E}$ is defined as follows.

$$\mathcal{E} : D \longrightarrow \mathrm{Irr}(W^D)$$
$$a \mapsto \mathcal{E}(a)$$

where $\mathcal{E}(a)$ is constructed according to the following rules

- Let $a = a_u \in D_{un}$. Let π be the morphism defined as in [5, 1.2], we denote by $\pi_!$ the sheaf functor given by direct image with compact support. The shifted intersection cohomology complex $\pi_! (\mathbf{Q}_l) [\dim D]$ has a decomposition as follows

$$\pi_! (\mathbf{Q}_l) [\dim D] = \oplus_\rho \rho \otimes \pi_! (\mathbf{Q}_l) [\dim D]_\rho$$

  where ρ runs through $\mathrm{Irr}\left(W^D\right)$ (for further details see [5, 1.2]). Then $\mathcal{E}(a)$ is the unique irreducible representation of W^D such that $\pi_!(\overline{\mathbb{Q}_l}[\dim D]_{\mathcal{E}(a)})_{|D_{un}}$ is (up to shift) the intersection cohomology complex of $\overline{G^\circ \cdot a}$ with coefficients in $\overline{\mathbb{Q}_l}$.

- If a_s is central in G, let D_{a_u} be the connected component of G containing a_u. We observe that $W^{D_{a_u}} = W^D$, because $D_{a_u} = G^\circ a_u$ and $D = G^\circ a_s a_u = a_s G^\circ a_u = a_s D_{a_u}$ with a_s central. Then $\mathcal{E}(a) \coloneqq \mathcal{E}(a_u)$.

- Let $a_s \notin Z(G)$. Note that a_s is central in $C_G(a_s)$, and we let D' be the connected component of $C_G(a_s)$ containing a. Let $W(C_G(a_s)^\circ)$ be the Weyl group of $C_G(a_s)^\circ$. We denote by $\mathcal{E}_{a_s}(a) \in \mathrm{Irr}(W(C_G(a_s)^\circ)^{D'})$ the image of a through the map $\mathcal{E}$ referred to the group $C_G(a_s)$. Then we set $\mathcal{E}(a) = \mathbf{j}_{W(C_G(a_s)^\circ)^{D'}}^{W(G^\circ)^D} \mathcal{E}_{a_s}(a)$, where $\mathbf{j}$ is the truncated induction as defined in [6, Section 3]. The description of $W(C_G(a_s)^\circ)^{D'}$ as a subgroup of $W(G^\circ)^D$ is given in [5, 1.6 (a)].

We observe that $\mathcal{E}(a)$ depends only on the G°-class of a.

4.3 Lusztig strata

Definition 4.1 Let $\rho \in \mathrm{Irr}(W^D)$. The fiber $\mathcal{E}^{-1}(\rho)$ is called a Lusztig stratum.

By [5, 1.16 (e)] Lusztig strata are unions of G°-orbits of the same dimension, so if X is a Lusztig stratum contained in D, then there exists $d \in \mathbb{N}$ such that $X \subset D_{(d)}$.
As observed in [5, 0.1], a Lusztig stratum is union of Jordan classes, indeed we have the following proposition.

Proposition 4.2 *If J is a Jordan class in D, and $a = a_s a_u, h = h_s h_u \in J$, then $\mathcal{E}(a) = \mathcal{E}(h)$, so strata are unions of Jordan classes.*

The fibers of the classical Springer correspondence are the unipotent conjugacy classes in the ambient group. It is expected that also the fibers of $\mathcal{E}$ have a good geometric behavior. Indeed, G. Lusztig in [5, 1.12 (b)] stated that the strata of D should be locally closed, as in the connected case [1].

Using Proposition 4.2, we describe Lusztig strata as unions of regular closures of Jordan classes. From this description, we are able to prove the following theorem.

Theorem 4.3 [2, Theorem 2.3] *Let X be a Lusztig stratum. Then X is a locally closed subset of D.*

References

[1] G. Carnovale, *Lusztig's strata are locally closed*. Archiv der Mathematik 115 (Mar. 2020). DOI: 10.1007/s00013-020-01448-1.

[2] Martina Costa Cesari, *Jordan classes and Lusztig strata in disconnected reductive groups*. Journal of Algebra 634 (2023), pp. 626–649. ISSN: 0021-8693. DOI: https://doi.org/10.1016/j.jalgebra.2023.06.046. URL: https://www.sciencedirect.com/science/article/pii/S002186932300368X.

[3] G. Lusztig, *Character sheaves on disconnected groups I*. Represent. Th. 7 (2003), pp. 374–403.

[4] G. Lusztig, "Hecke Algebras with Unequal Parameters". CRM monograph series. AMS, 2003. ISBN: 9780821833568. URL: https://books.google.de/books?id=7r4jDwAAQBAJ.

[5] G. Lusztig, *Strata of a disconnected reductive group*. Indagationes Mathematicae 32.5 (2021). Special issue to the memory of T.A. Springer, pp. 968–986. ISSN: 0019-3577. DOI: https://doi.org/10.1016/j.indag.2020.09.011. URL: https://www.sciencedirect.com/science/article/pii/S0019357720301038.

[6] G. Lusztig and N. Spaltenstein, *Induced Unipotent Classes*. Journal of the London Mathematical Society s2-19.1 (1979), pp. 41–52. DOI: https://doi.org/10.1112/jlms/s2-19.1.41. EPRINT: https://londmathsoc.onlinelibrary.wiley.com/doi/pdf/10.1112/jlms/s2-19.1.41. URL: https://londmathsoc.onlinelibrary.wiley.com/doi/abs/10.1112/jlms/s2-19.1.41.

[7] G. Malle and D. Testerman, "Linear Algebraic Groups and Finite Groups of Lie Type". Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2011. ISBN: 9781139499538. URL: https://books.google.de/books?id=4l52zICEq3EC.

[8] T.A. Springer, "Linear Algebraic Groups". Progress in Mathematics. Springer, 1998. ISBN: 9780817640217. URL: https://books.google.de/books?id=2XcCJRH4p8YC.

[9] T.A. Springer, *Trigonometric sums, Green functions of finite groups and representations of Weyl groups*. Inventiones Math. 36 (1976), pp. 173–209. DOI: https://doi.org/10.1007/BF01390009.

A 2D Bin Packing Problem in the Sheet Metal Industry: models and solution approaches

Chiara Turbian (*)

Abstract. The Bin Packing Problem (BPP) is a well-studied problem in Operations Research, and, in its basic formulation, it aims at packing a set of items into a finite set of bins by minimizing the number of used bins. Due to its wide range of applications, several variants of the problem have been proposed during the last decades, which differ from each other by dimensionality, additional constraints, and characteristics of the items or the bins. We consider a Two-Dimensional Bin Packing Problem (2DBPP) arising in Salvagnini Italia, a multinational corporation working in the sheet metal industry. In our problem, the basic 2DBPP is enriched by the presence of technological constraints emerging from the context, such as precedence relations between groups of items and conditional safety distances between items. We present exact and heuristic approaches to solve the problem, both based on Mixed Integer Linear Programming (MILP), and we show related computational results.

1 Introduction

The Bin Packing Problem (BPP) is classical problem in the Operations Research literature that, in its general form, aims at packing a set of items into a finite set of bins by minimizing some efficiency index, e.g., the number of used bins. The BPP has a wide range of industrial applications (e.g., in wood, glass, paper or sheet metal industries, and in freight transportation), thus many variants of this problem have been proposed and studied in the last decades. These variants differ from each other by the following aspects: dimensionality (e.g., 1D-BPP [9], 2D-BPP [5], or 3D-BPP [6]); characteristics of the items (e.g., rectangular [5], or irregular shaped [4] items); characteristics of the bins (e.g., identical [3], or variable sized [2] bins); additional constraints (e.g., items rotations [7], weighted items [1], constrained barycenters [8]). In this report, we consider a real-world application of a Two-Dimensional Bin Packing Problem (2DBPP) with rectangular items, variable bin sizes, and practical constraints arising in Salvagnini Italia, a multinational corporation that produces, in particular, computer numerical control (CNC) machines designed to cut the sheet metal. We describe in details the problem, with a particular focus on its practical

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: turbian@math.unipd.it . With Salvagnini Italia S.p.A., via Ingegnere Guido Salvagnini 51, 36040 Sarego, Italy. E-mail: chiara.turbian@salvagninigroup.com . Seminar held on 8 November 2023.

constraints (Section 2), and, after a brief introduction on Mathematical Programming, we provide a formulation of the problem (Section 3). In Section 4, we provide some generalities on the two main categories of solution methods that are used in Operations Research (i.e., exact and heuristic methods), followed by the solution approaches that we propose to solve our specific problem. To conclude, in Section 5, we comment computational results related to our procedures, highlighting advantages and disadvantages of using exact or heuristic methods.

2 Problem Definition

In our problem we have to cut two sets of rectangular items, compulsory and optional ones, from a limited amount of material sheets of different sizes. The goal is to cut all the compulsory items and, possibly, optional items by minimizing the material waste of the used sheets. We have several constraints to consider in our problem, some classical ones of the 2DBPP, and some others peculiar to our framework. As classical limitations we have that items cannot overlap each other and cannot be split into smaller pieces to better fit in the sheets. Furthermore, to each item we can assign mandatory or forbidden placement areas, and the freedom to rotate by 90 degrees. On the other hand, specific to the context, we have additional constraints on the reciprocal position of some items within the sheet and in the solution. To preserve the quality of the cut, we can assign a *margin* to an item, representing the minimum distance from any other item in the sheet. In particular, if two items have no margins, they can either share a side with each other (in this case, we say that they *share a common cut*), or they must respect a safety distance that depends on the sheet type. At last, different levels of precedence can be assigned to each compulsory item, indicating that items of higher precedence should be accommodated in sheets scheduled for cutting before sheets containing items of lower precedence. This characteristic drives our goal of generating a sequence of sheets, thus establishing a production order among them. Within our framework, the precedence constraint can be classified as either *hard*, as previously explained, or *soft*, where fulfillment is required only if it does not increase material waste.

3 Mathematical Modeling

3.1 General

Mathematical Programming is one of the possible approaches to tackle an optimization problem. It involves mathematical programming models, that are used to describe the characteristics of the optimal solution by means of mathematical relations. These models consist of the following elements: sets (which group the elements of the system), parameters (i.e., the data of the problem, which represent the known quantities depending on the elements of the system), decision variables (these are the unknown quantities, on which we can act in order to find different possible solutions to the problem), constraints (i.e., the mathematical relations that describe conditions imposing the feasibility of the solutions), objective function (this is the quantity to maximize or minimize, written as a function of

the decision variables). Solving an optimization problem formulated as a mathematical programming model means deciding the values of the variables that satisfy all the constraints and maximize or minimize the objective function. These values are the solution to the problem. A Linear Programming model is a mathematical programming model in which the objective function is a linear expression of the decision variables, and the constraints are given by a system of linear equations and/or inequalities. Depending on the nature of the domain of the decision variables, we have: Linear Programming models (LP) if all the variables can take real values, Integer Linear Programming models (ILP) if all the variables are allowed to take integer values only; Mixed Integer Linear Programming models (MILP) if some variables can take real values and others are allowed to take integer values only. In the following section we will see that our problem is described by two MILP formulations. In all generality, a MILP model is of the form

$$\begin{aligned} max/min \quad & c^T x \\ s.t. \quad & Ax \leq b \\ & x \geq 0 \\ & x_i \in \mathbb{Z}, \quad i \in I \end{aligned}$$

where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$, and $I \subseteq \{1, \ldots, n\}$ is the index set of the integer variables (i.e., x_i with $i \in I$ is an integer variable, x_i with $i \notin I$ is a continuous variable).

3.2 Application

The problem outlined in Section 2 presents two hierarchical goals: reducing material waste and, if waste quantities are identical, minimizing the number of soft precedence constraints that are violated. This problem is addressed through two successive steps, each relying on Mathematical Programming. The initial model (Model 1) focuses on capturing the primary features of the problem, emphasizing the minimization of material waste. Meanwhile, the subsequent model (Model 2) is designed to minimize the number of soft precedence violations under maximum waste constraint.

Notation

We report in the following the adopted notation. Consider the collection of compulsory items as $I_c = \{1, \ldots, n_c\}$, the one of optional items as $I_o = \{n_c + 1, \ldots, n_c + n_o = n\}$, and their union as $I = I_c \cup I_o$. Each item $i \in I$ is defined by its width ω_i and height η_i. Additionally, each item $i \in I$ can have a designated margin from the sheet's border (σ_i) or from other items (μ_i). We denote $\overline{M}$ as the maximum margin among all items. Parameters ϵ_i and ϕ_i confine the allowable placement range of item i to a maximum and minimum distance from the sheet's border, respectively. The level of hard precedence for an item $i \in I_c$ is represented by an integer ρ_i. Let $P = \{(i, j) \mid i, j \in I_c, \rho_i < \rho_j\}$, comprising all pairs of items in a hard precedence relationship. Assume there are n_s available sheets, with $S = \{1, \ldots, n_s\}$ denoting the sheet positions in the production order, and $T = \{1, \ldots, n_t\}$ representing sheet types. Each sheet type $t \in T$ is specified by its width Ω_t and height H_t. Let $\overline{\Omega} = \max_{t \in T} \Omega_t$, $\overline{H} = \max_{t \in T} H_t$, and $\overline{\Delta} = \max\{\overline{\Omega}, \overline{H}\}$. Every sheet type t has

a maximum available quantity of sheets denoted by β_t and a safety distance τ_t. For each $i, j \in I$, $s \in S$, and $t \in T$, the following variables are introduced: x_i (or y_i) representing the distance of item i's lower-left corner from the left (or bottom) side of the sheet; r_i is 1 if item i is placed vertically, otherwise 0; l_{ij} (or b_{ij}) is 1 if item i is positioned to the left (or below) j, otherwise 0; c_{ij} is 1 if items i and j share a common cut, otherwise 0; m_{ij} represents the margin between items i and j; f_{is} is 1 if item i is cut from the sheet in position s, otherwise 0; g_{st} is 1 if the sheet in position s belongs to type t, otherwise 0; a_{ijs} is 1 if both items i and j are placed on the sheet in position s, otherwise 0.

Model 1

We consider the following MILP model as Model 1.

$$\text{(1)} \qquad \min \sum_{s \in S} \left[\sum_{t \in T} (\Omega_t H_t) g_{st} - \sum_{i \in I} (\omega_i \eta_i) f_{is} \right]$$

s.t.

$$\text{(2)} \qquad l_{ij} + l_{ji} + b_{ij} + b_{ji} + (1 - f_{is}) + (1 - f_{js}) \geq 1 \qquad \forall i, j \in I, s \in S$$

$$\text{(3)} \qquad x_i + (1 - r_i)\omega_i + r_i \eta_i + m_{ij} \leq x_j + \overline{\Omega}(1 - l_{ij}) \qquad \forall i, j \in I$$

$$\text{(4)} \qquad x_i + (1 - r_i)\omega_i + r_i \eta_i + m_{ij} \geq x_j - \overline{\Omega}(1 - l_{ij}) \qquad \forall i, j \in I$$

$$\text{(5)} \qquad y_i + (1 - r_i)\eta_i + r_i \omega_i + m_{ij} \leq y_j + \overline{H}(1 - b_{ij}) \qquad \forall i, j \in I$$

$$\text{(6)} \qquad y_i + (1 - r_i)\eta_i + r_i \omega_i + m_{ij} \geq y_j - \overline{H}(1 - b_{ij}) \qquad \forall i, j \in I$$

$$\text{(7)} \qquad m_{ij} \geq max\{\mu_i, \mu_j, \tau_t\}(1 - c_{ij}) - \overline{\Delta}\,(2 - a_{ijs} - g_{st}) \qquad \forall i, j \in I, \; s \in S, t \in T$$

$$\text{(8)} \qquad m_{ij} \leq \overline{\Delta}(1 - c_{ij}) \qquad \forall i, j \in I$$

$$\text{(9)} \qquad a_{ijs} \geq f_{is} + f_{js} - 1 \qquad \forall i, j \in I, s \in S$$

$$\text{(10)} \qquad max\{\mu_i, \mu_j\}(l_{ij} + l_{ji}) \leq \overline{M}(1 - c_{ij}) \qquad \forall i, j \in I$$

$$\text{(11)} \qquad max\{\mu_i, \mu_j\}(b_{ij} + b_{ji}) \leq \overline{M}(1 - c_{ij}) \qquad \forall i, j \in I$$

$$\text{(12)} \qquad \sum_{t \in T} g_{st} \leq 1 \qquad \forall s \in S$$

$$\text{(13)} \qquad \sum_{s \in S} g_{st} \leq \beta_t \qquad \forall t \in T$$

$$\text{(14)} \qquad \sum_{s \in S} f_{is} = 1 \qquad \forall i \in I_c$$

$$\text{(15)} \qquad \sum_{s \in S} f_{is} \leq 1 \qquad \forall i \in I_o$$

$$\text{(16)} \qquad \sum_{i \in I} f_{is} \leq n \sum_{t \in T} g_{st} \qquad \forall s \in S$$

$$\text{(17)} \qquad f_{js} \leq \sum_{r=1}^{s} f_{ir} \qquad \forall s \in S, \; (i, j) \in P$$

$$x_i + (1 - r_i)\omega_i + r_i\eta_i \leq \sum_{t \in T} (\Omega_t - \phi_i - \sigma_i) g_{st} + \overline{\Omega}(1 - f_{is}) \qquad \forall i \in I, s \in S \tag{18}$$

$$y_i + (1 - r_i)\eta_i + r_i\omega_i \leq \sum_{t \in T} (H_t - \phi_i - \sigma_i) g_{st} + \overline{H}(1 - f_{is}) \qquad \forall i \in I, s \in S \tag{19}$$

$$x_i \geq \phi_i + \sigma_i - \overline{\Omega}(1 - f_{is}) \qquad \forall i \in I, s \in S \tag{20}$$

$$y_i \geq \phi_i + \sigma_i - \overline{H}(1 - f_{is}) \qquad \forall i \in I, s \in S \tag{21}$$

$$x_i + (1 - r_i)\omega_i + r_i\eta_i \leq \epsilon_i + \sigma_i + \overline{\Omega}(1 - f_{is}) \qquad \forall i \in I, s \in S \tag{22}$$

$$y_i + (1 - r_i)\eta_i + r_i\omega_i \leq \epsilon_i + \sigma_i + \overline{H}(1 - f_{is}) \qquad \forall i \in I, s \in S \tag{23}$$

$$x_i \geq \sum_{t \in T} (\Omega_t - \epsilon_i - \sigma_i) g_{st} - \overline{\Omega}(1 - f_{is}) \qquad \forall i \in I, s \in S \tag{24}$$

$$y_i \geq \sum_{t \in T} (H_t - \epsilon_i - \sigma_i) g_{st} - \overline{H}(1 - f_{is}) \qquad \forall i \in I, s \in S \tag{25}$$

$$x_i,\ y_i \geq 0 \qquad \forall i \in I \tag{26}$$

$$l_{ij},\ b_{ij},\ c_{ij} \in \{0, 1\} \qquad \forall i, j \in I \tag{27}$$

$$r_i \in \{0, 1\} \qquad \forall i \in I \tag{28}$$

$$m_{ij} \geq 0 \qquad \forall i, j \in I \tag{29}$$

$$f_{is} \in \{0, 1\} \qquad \forall i \in I, s \in S \tag{30}$$

$$g_{st} \in \{0, 1\} \qquad \forall s \in S, t \in T \tag{31}$$

$$a_{ijs} \in \{0, 1\} \qquad \forall i, j \in I,\ s \in S \tag{32}$$

Objective (1) aims to minimize the total material waste across the utilized sheets. In equation (2), we establish correlations between items: when two items share a sheet, they are positioned adjacently or one below the other. Constraints (3-6) for non-overlapping ensure that item margins are considered in their definitions. Constraints (7-8) delineate the margin between two items: if they share a cut, the margin is zero; otherwise, it's at least the larger value between the items' margin and the safety distance. To maintain linearity, variables a in constraints (9) indicate whether two items share a sheet. Constraints (10-11) regulate the potential for shared cuts between items, accounting for their relative positions; an item with a specified margin cannot share a cut with another item. Equations (12-13) ensure that at most one type is assigned to each sheet, within the available quantities. Constraints (14-15) mandate that compulsory items are placed in exactly one sheet, while optional items can be placed in at most one sheet. Equation (16) impose that if an item is placed on a sheet, a type must be assigned to it. Constraints (17) establish hard precedence between items, ensuring that a sheet accommodating an item i with higher precedence than item j is not cut after the sheet accommodating j. Equations (18-21) ensure that each item remains completely within a sheet and maintains the specified margin from the sheet's border. They also prevent items from being placed in their forbidden zones. Similarly, constraints (22-25) guarantee item placement within their mandatory zones. Constraints (26-32) specify the domains of the variables.

Model 2

To incorporate soft precedence relationships, we denote the soft precedence level, specifically for compulsory items, as $\tilde{\rho}_i$ for each $i \in I_c$. Similarly to hard precedence, we define the set $\tilde{P} = (i,j) \mid i,j \in I_c, \tilde{\rho}_i < \tilde{\rho} j$ to represent these soft precedence relationships. Subsequently, we introduce decision variables q_{ijs} for all $(i,j) \in \tilde{P}$ and for all $s \in S$. These variables take value 1 if the placement of item j at position s violates the soft precedence concerning item i, and 0 otherwise. We consider the following MILP model as Model 2:

$$\min \sum_{(i,j)\in\tilde{P}} \sum_{s\in S} q_{ijs} \tag{33}$$

s.t.

$$(2) - (32)$$

$$q_{ijs} \geq f_{js} - \sum_{r=1}^{s} f_{ir} \qquad \forall s \in S, (i,j) \in \tilde{P} \tag{34}$$

$$\sum_{s\in S} \left[\sum_{t\in T} (\Omega_t H_t) g_{st} - \sum_{i\in I} (\omega_i \eta_i) f_{is} \right] \leq v^* \tag{35}$$

$$q_{ijs} \in \{0,1\} \qquad \forall s \in S, (i,j) \in \tilde{P}. \tag{36}$$

In the objective (33), we minimize the number of soft precedence violations, that are counted by constraints (34). Moreover, we impose that the total amount of material waste has to be at most a certain quantity v^* in constraint (35).

4 Algorithms

4.1 General

Solution approaches for optimization problems can be categorized into two classes: *exact methods* and *heuristic methods*. Exact methods, theoretically, have the capability to furnish an optimal solution, i.e., they can produce a feasible solution that optimizes either the minimization or maximization of the objective function. Conversely, heuristic methods yield feasible solutions without any assurance of optimality. The selection of a solution approach heavily relies on the problem's structure and the limitations imposed by the context. In some cases, we might discover efficient exact algorithms suited for solving an optimization problem. Alternatively, formulating the problem as a Mixed Integer Linear Programming (MILP) model allows for its resolution via a MILP solver (e.g. Cplex, Gurobi, Xpress, AMPL, OPL, etc.). These solvers employ general-purpose exact algorithms that theoretically ensure the attainment of the optimal solution. However, their downside lies in their exponential computational complexity, potentially resulting in an exponential increase in solving time relative to the problem's size. Consequently, employing exact solution methods is not always feasible or appropriate basically due to two concurrent issues: the inner complexity of the problem (e.g. NP-hard nature) and the time constraint for providing a solution. In such scenarios, it becomes necessary to resort to methods that offer satisfactory solutions within acceptable computational time frames, even if they cannot guarantee

optimality. These circumstances call for the utilization of heuristic methods. In recent times, considerable attention has been directed towards *matheuristic methods*, which combine mathematical programming with general heuristic algorithms. The distinctive feature of matheuristics lies in the central role occupied by the mathematical programming model, forming the foundation of the overall heuristic design. Consequently, matheuristics do not represent a rigid paradigm but rather a conceptual framework for devising heuristics that maintain a mathematical core.

4.2 Application

We propose two procedures to solve the problem: an exact approach and a matheuristic. The first approach, called the *Lexicographic procedure*, is based on the following logic. To find a solution that minimizes material waste and, in the case of equal waste, minimizes the number of violated soft precedence constraints, we first solve Model 1 and, then, we solve Model 2 with v^* held constant at the optimal value derived from Model 1. Both Model 1 and Model 2 are solved using off-the-shelf MILP solvers. The second approach, referred to as the *Iterative procedure*, operates on the concept of iteratively determining the maximum position of items in the sheets' sequence based on their soft precedence. We group compulsory items according to their soft precedence into subsets $I_k = \{i \in I_c \mid \tilde{\rho}_i = k\} \subseteq I_c$, where k ranges from 1 to $\tilde{\rho}_{max}$ being the maximum soft precedence level. Additionally, a new integer parameter δ_k is introduced for each k, denoting that items in I_k must be placed in a sheet to be cut at a position less than or equal to δ_k. This parameter is determined iteratively across the soft precedence values $k = 1, \ldots, \tilde{\rho}_{max}$. During each iteration k, we solve Model 1 considering only compulsory items up to soft precedence k and disregarding optional items. Starting from the second iteration, constraints are imposed to define the maximum position in the sheets' sequence for items up to soft precedence $k-1$ through the equations:

$$\sum_{s \in \{1,\ldots,\delta_{\bar{k}}\}} f_{is} = 1 \quad \forall i \in I_{\bar{k}}, \ \ \forall \bar{k} = 1, \ldots, k-1. \tag{37}$$

The parameter δ_k at each iteration is set to the number of used sheets in the solution derived from iteration k. These constraints ensure that higher soft precedence items are positioned in initial sheets, allowing lower soft precedence items to be placed in subsequent sheets or used to minimize waste in initial sheets, this time at the expense of precedence violations. After fixing all δ-parameters, we solve Model 1 one last time, incorporating optional items, by adding the following constraints:

$$\sum_{s \in \{1,\ldots,\delta_k\}} f_{is} = 1 \quad \forall i \in I_k, \ \ \forall k = 1, \ldots, \tilde{\rho}_{max}. \tag{38}$$

5 Computational Results

The proposed algorithms have been implemented using Python, and computational experiments have been conducted to evaluate their performance in comparison to the optimization procedure currently employed by the company. We conducted tests on a total

of 64 instance classes, which differ from each other by the number of: sheet types (1 or 3), compulsory items (5, 10, 15 or 20), optional items (none or one third of the number of compulsory items), items with required margin (none or all), or items with soft precedence (none or all). These configurations align with the production requirements of certain clients associated with the reference company, imparting practical relevance to the randomized test instances. The results presented are based on a benchmark of 20 instances for each class, summing up to a total of 1280 instances. These instances were solved on a machine equipped with an Intel Core i7-8700 processor running at 3.2GHz and possessing 32GB of RAM. CPLEX 12.8 was utilized as the MILP solver for these experiments. All procedures were executed within a time limit of 600 seconds, balancing operational needs with the time requirements of our methods.

Group	Company	Lexicographic				Iterative					
	Err	Time	F	S	B	Time	F	W	S	B	Err
A	12.1	0.7	0.0	77.2	22.8	0.6	0.0	0.3	78.4	21.3	0.6
B	33.3	142.9	15.3	54.7	30.0	52.6	3.7	4.4	58.8	33.1	10.8
C	21.3	390.6	58.7	25.6	15.7	284.9	40.9	1.9	38.4	18.8	14.1
D	10.3	500.1	77.2	16.2	6.6	409.2	61.6	0.0	25.6	12.8	7.9
Total	*19.3*	*258.6*	*37.8*	*43.4*	*18.8*	*186.9*	*26.6*	*1.6*	*50.3*	*21.5*	*8.4*

Table 1: Computational results - instances grouped by size.

Table 1 reports a comparative analysis of the Company's procedure, the Lexicographic method, and the Iterative approach, grouped by the number of compulsory items (5, 10, 15, 20) within Group A, B, C, and D instances, respectively. The focus remains on waste minimization, the primary objective of optimization. Columns within the table represent investigated metrics: percentage error relative to the best waste obtained by any of the three procedures (*Err*), runtime in seconds (*Time*), and, for our solving procedures, the percentage of instances categorized by whether the procedure fails, i.e., it runs out of the time limit before solving to optimality all the involved models (F), provides worse (W), same (S), or better (B) waste in comparison to the Company's procedure. Comparing with the Company's approach, the exact Lexicographic procedure exhibits better results in 18.8% of instances. However, its extended execution times severely limit its applicability to larger instances. Indeed, only around 23% of the largest instances (Group D) were solvable within the time limit (resulting in a failure rate of 77.2%). Conversely, our proposed Iterative matheuristic demonstrates lower failure rates, at the cost of an error margin with respect to the the optimal or best-known waste. Nonetheless, the error remains limited (14.1% in the most challenging group) and consistently lower than that of the Company's procedure. Furthermore, the Iterative procedure is still beneficial, with respect to the Company's one, in a comparable or even larger number of instances than the Lexicographic procedure (see columns B). In conclusion, we have developed two algorithms tailored for a real 2DBPP problem, integrating several practical attributes relevant to the sheet metal industry. The first procedure yields exact solutions by solving a sequence of two MILP formulations, valuable for benchmarking heuristic approaches, including the one currently

adopted by the reference company. On the other hand, the second procedure employs soft-precedence-guided decomposition to reduce runtime by iteratively solving smaller MILP models. This matheuristic often yields improved solutions compared to the Company's approaches in waste minimization and precedence satisfaction. Additionally, it scales better with instance size than the exact approach, presenting a promising integration into the current fast heuristic algorithm adopted by the company. Future research will focus on further reducing execution times and incorporating additional practical attributes.

References

[1] Mauro Maria Baldi, Teodor Gabriel Crainic, Guido Perboli, and Roberto Tadei, *The generalized bin packing problem.* Transportation Research Part E: Logistics and Transportation Review, 48(6): 1205–1220, 2012.

[2] Donald K. Friesen and Michael A. Langston., *Variable sized bin packing.* SIAM Journal on Computing, 15(1): 222–230, 1986.

[3] Manuel Iori, Vinícius L. de Lima, Silvano Martello, Flávio K. Miyazawa, and Michele Monaci, *Exact solution techniques for two-dimensional cutting and packing.* European Journal of Operational Research, 289(2): 399–415, 2021.

[4] Aline A.S. Leao, Franklina M.B. Toledo, José Fernando Oliveira, Maria Antónia Carravilla, and Ramón Alvarez-Valdés, *Irregular packing problems: A review of mathematical models.* European Journal of Operational Research, 282(3): 803–822, 2020.

[5] Andrea Lodi, Silvano Martello, and Michele Monaci, *Two-dimensional packing problems: A survey.* European Journal of Operational Research, 141(2): 241–252, 2002.

[6] Silvano Martello, David Pisinger, and Daniele Vigo, *The three-dimensional bin packing problem.* Operations Research, 48(2): 256–267, 2000.

[7] Hongtao Tang, Xixing Li, Shunsheng Guo, Shuwei Liu, Li Li, and Lang Huang, *An optimizing model to solve the nesting problem of rectangle pieces based on genetic algorithm.* Journal of Intelligent Manufacturing, 28, 03 2015.

[8] Alessio Trivella and David Pisinger, *The load-balanced multi-dimensional bin-packing problem.* Computers & Operations Research, 74: 152–164, 2016.

[9] José Manuel Valério de Carvalho, *LP models for bin packing and cutting stock problems.* European Journal of Operational Research, 141(2): 253–273, 2002.

Groups and geometry: from algebraic varieties to Galois representations and vice versa

Khai Hoan Nguyen Dang (*)

Abstract. Since a very long time ago, there has been an effective approach to study geometry via group theory. In this talk, we will focus on objects given by sets of solutions of a system of polynomial equations, called algebraic varieties. Galois theory makes a bridge between the geometry of algebraic varieties and group theory in terms of Galois representations. The talk will survey some basic but still interesting aspects of these connections and provide several examples. We will also provide a uniform way to investigate a certain class of algebraic varieties, named Abelian varieties.

1 History: Erlangen Program


Vergleichende Betrachtungen

über

neuere geometrische Forschungen

von

Dr. Felix Klein,
o. ö. Professor der Mathematik an der Universität Erlangen.

Programm

zum Eintritt in die philosophische Facultät und den Senat
der k. Friedrich-Alexanders-Universität
zu Erlangen.

Our journey begins with the Erlangen Program, a conceptual method in mathematics presented by Felix Klein in the 1870s. The program is a framework for studying geometries based on their underlying groups of transformations, which are now known as the symmetries of the geometries. The mathematical representation of the hierarchy of geometries took the form of a hierarchy of these groups and their invariants. The program is named after the University Erlangen-Nürnberg, where Klein was affiliated. For instance, the Euclidean group of symmetries preserves lengths, angles, and areas, while the most general projective transformations preserve only the incidence structure and the cross-ratio. By abstracting the underlying groups of symmetries from the geometries, the relationships between them could be reestablished at the group level. Felix Klein published it under the title 'Vergleichende Betrachtungen über neuere geometrische Forschungen' [5].

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: dkhn@math.unipd.it . Seminar held on 22 November 2023.

One of the program's key insights is the move towards abstraction and generalization. By focusing on the symmetries and invariants of geometric structures, Klein aimed to create a more abstract and general understanding of mathematics. The program influenced numerous mathematicians, including Bernhard Riemann, Henri Poincare, Herman Weyl, Elie Cartan, and many others, and played a crucial role in shaping the development of modern mathematics. The ideas of the Erlangen Program have had a significant impact beyond geometry, influencing various branches of mathematics and even physics. For example, Noether's theorem, which establishes a fundamental connection between symmetries and conservation laws in physics, was deeply influenced by Klein's thought.

2 Algebraic Varieties

An algebraic variety is a geometric object defined by polynomial equations.

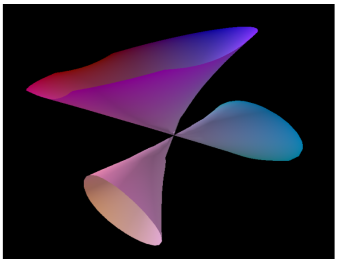


Figure 1: $x^2y - y^3 - z^2 = 0$

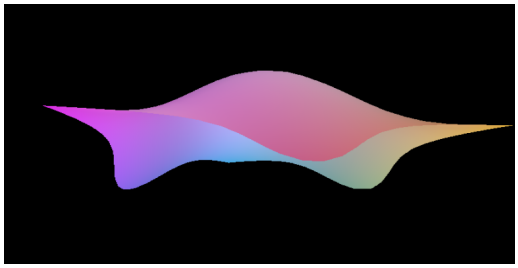


Figure 2: $x^3 + y^3 + z^3 + w^3 = 0$

Fundamental objects in the study of algebraic varieties are elliptic curves, given by the equation $y^2 = x^3 + ax + b$ with $a, b \in \mathbb{Z}$.

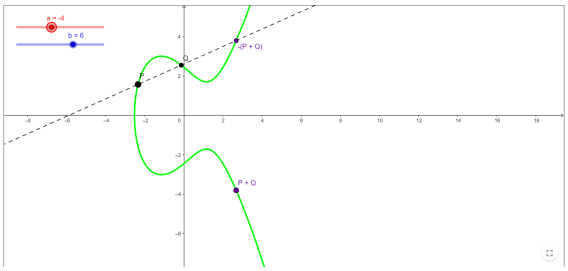


Figure 3: $xy^2 = x^3 - 4x + 6$

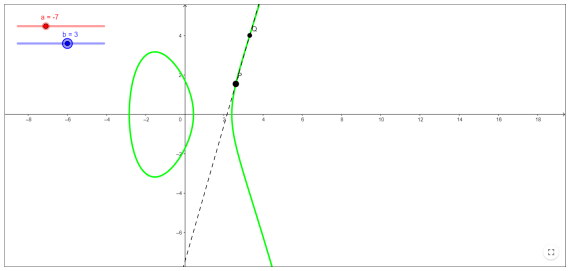
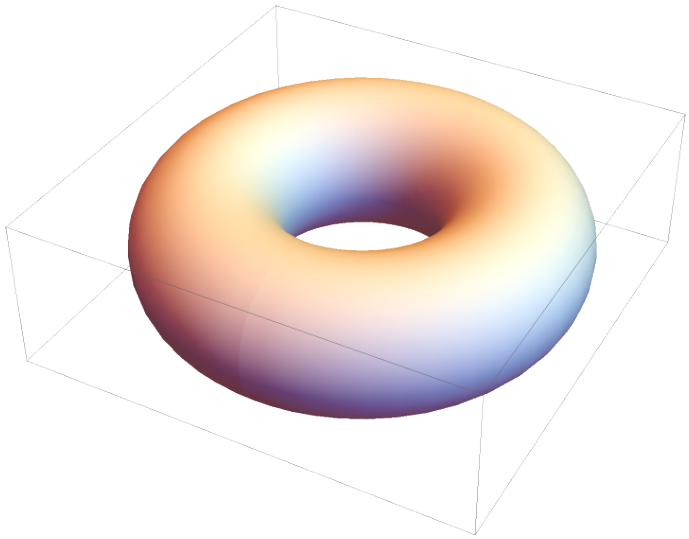


Figure 4: $y^2 = x^3 - 7x + 3$

Abelian varieties are higher dimensional generalizations of elliptic curves. The following property tells us the uniformization of Abelian varieties [12].

Proposition 1 *Complex points of an abelian variety over* $\mathbb{C}$ *make the shape of a torus, i.e., for any abelian variety* A *over* $\mathbb{C}$ *we have*

$$\mathrm{A}(\mathbb{C}) \cong \mathbb{C}^g/\mathbb{Z}^{2g}$$



Number theorists investigate the geometry of integral or rational points of algebraic varieties. For a good reference, see [15].

Theorem 1 [Mordell-Weil Theorem] *The group of rational points of an Abelian variety A over $\mathbb{Q}$ is finitely generated, i.e.,*

$$A(\mathbb{Q}) = \mathbb{Z}^r \times \{torsion\}$$

Example 1

(a) For $A : y^2 = x^3 - 4x + 6$, we have $r = 1$ and the torsion is trivial.

(b) For $A : y^2 = x^3 - 7x + 3$, we have $r = 2$ and the torsion is trivial.

The rank r of E is a mysterious object. It appears in the Birch and Swinnerton-Dyer conjecture, one of the seven Millennium Prize Problems. However, we know well the torsion part [11].

Theorem 2 [Mazur 1977] *Let $E/\mathbb{Q}$ be an elliptic curve over rational numbers. The torsion subgroup of $E(\mathbb{Q})$ is isomorphic to one of the following*

$$\mathbb{Z}/n\mathbb{Z} \text{ or } \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2m\mathbb{Z},$$

where $n \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12\}$ and $m \in \{1, 2, 3, 4\}$. In particular, $|E(\mathbb{Q})_{tor}| \leq 16$.

Now, let A be an abelian variety over a finite field $\mathbb{F}_p$, the group $A(\mathbb{F}_p)$ is necessarily finite [17].

Theorem 3 [Hasse-Weil Theorem] *Let A be an abelian variety of dimension n over a finite field $\mathbb{F}_p$, then*

$$(p + 1 - 2\sqrt{p})^n \leq |A(\mathbb{F}_p)| \leq (p + 1 + 2\sqrt{p})^n$$

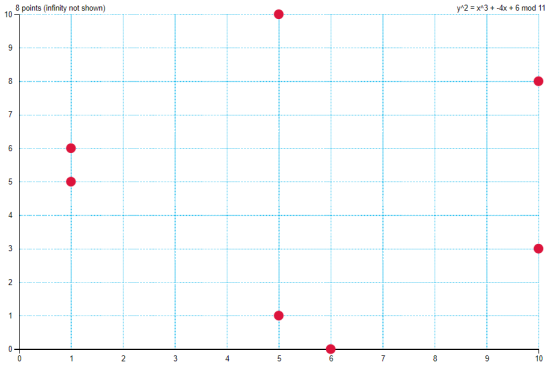


Figure 5: $y^2 = x^3 - 4x + 6 \mod 11$

An abelian variety is called good at a prime number p if it is still smooth after reducing modulo p. Otherwise, we call it bad.

Proposition 2 [15] *Let $E/\mathbb{Q}$ be an elliptic curve defined by $y^2 = x^3 + ax + b$, and let p be a prime that does not divide the discriminant $\Delta(E) = -16(4a^3 + 27b^2)$. Then E has good reduction at p.*

Example 2

(a) The elliptic curve $A : y^2 = x^3 - 4x + 6$ has bad reduction at 197.

(b) The elliptic curve $A : y^2 = x^3 - 7x + 3$ has good reduction at 197.

3 Galois Representations

Given a group G, we want to consider a linear space associated to G.

Definition 1 Let G be a topological group and V be a vector space over a topological field k of dimension n. An n-dimensional representation of G is a continuous homomorphism of groups

$$\rho : G \to \mathrm{Aut}(V) \cong \mathrm{GL}_n(k)$$

Example 3 Trivial representation: $\rho(g) = 1$ for all $g \in G$.

The goal of modern number theory is to understand $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$.

Remark 1

(a) $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ is a profinite group.

(b) There is no explicit description (in terms of generators and relations) known for $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$.

To study $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$, people look at its representations.

Example 4 There is an 1-dimensional representation called cyclotomic representation

$$\chi_p : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathbb{Z}_p^*$$

We construct χ_p as follows. For each $n \geq 1$, let ζ_{p^n} be a primitive p^n-th root of unity and let $K_n = \mathbb{Q}(\zeta_{p^n})$ be the corresponding cylotomic extension of $\mathbb{Q}$. We have

$$\mathrm{Gal}(K_n/\mathbb{Q}) \cong (\mathbb{Z}/p^n\mathbb{Z})^*$$

Hence, for each n we can construct a representation

$$\chi_{p,n} : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{Gal}(K_n/\mathbb{Q}) \cong (\mathbb{Z}/p^n\mathbb{Z})^*$$

which is compatible when n varies. Taking the inverse limit we obtain a representation

$$\chi_p : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \varprojlim_n (\mathbb{Z}/p^n\mathbb{Z})^* \cong \mathbb{Z}_p^*$$

People look at $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ with specific additional structures associated to the prime numbers, i.e., for each p prime we consider p-adic numbers $\mathbb{Q}_p$ and the embedding

$$\mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p) \hookrightarrow \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$$

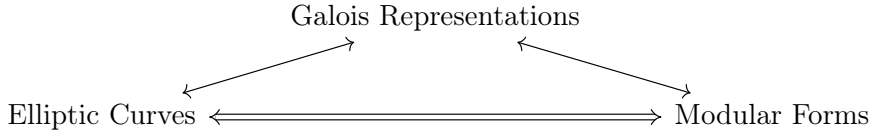
Proposition 3 *Reduction modulo p gives rise to a surjective morphism* $\mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p) \twoheadrightarrow \mathrm{Gal}(\overline{\mathbb{F}}_p/\mathbb{F}_p)$ *Denote by I_p its kernel.*

Definition 2 We call a representation unramified at p if $\rho(I_p) = \{1\}$.

4 Bridges Between Two Islands

4.1 Modularity Theorem

Perhaps the most famous breakthrough in the 20th century is the proof of Fermat's last theorem by Andrew Wiles [18]. It is based on the correspondence between elliptic curves and modular forms by means of Galois representations.



The correspondence implies Fermat's last theorem, one of the most notable theorems in the history of mathematics.

4.2 Faltings' Theorem

Let A be an Abelian variety over a field K of characteristic 0 (e.g $\mathbb{Q}$ or $\mathbb{Q}_p$). Tate module T_pA characterizes p-power torsion points of the abelian variety, i.e., T_pA is a set of sequences $(x_n)_{n\in\mathbb{N}}$ such that $x_n \in \ker(p^n : A(\overline{K}) \to A(\overline{K}))$ and satisfy a compatible relation.

(a) Points of an Abelian variety A are inherited from the action of the absolute Galois group $\mathrm{Gal}(\overline{K}/K)$.

(b) Tate module is a Galois representation of $\mathrm{Gal}(\overline{K}/K)$ of dimension 2g, where g is the dimension of A.

A celebrated theorem due to Faltings roughly says that a morphism between two given Abelian varieties can be determined by a morphism of corresponding Tate modules (up to an isomorphism) [2].

Theorem 4 *Let A and B be Abelian varieties over $\mathbb{Q}$. Then for all prime p, we have an isomorphism*

$$\mathrm{Hom}_{\mathbb{Q}}(A, B) \otimes \mathbb{Z}_p \cong \mathrm{Hom}_{\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})}(T_pA, T_pB)$$

While the left-hand side captures the geometric properties of Abelian varieties (isogenies), the right-hand side concerns Galois representations (Tate modules).

Furthermore, the Tate module can determine the type of reduction of an Abelian variety [14].

Theorem 5 [Neron-Ogg-Shafarevich-Serre-Tate] *The abelian variety is good at p if and only if the Tate module $T_\ell A$ is unramified representation at ℓ for some $\ell \neq p$.*

When $\ell = p$ we replace the word unramified by crystalline, defined by Fontaine in the 80's in p-adic Hodge theory [1].

Theorem 6 [Fontaine, Coleman-Iovita, Breuil] *The abelian variety is good at p if and only if the Tate module is crystalline.*

5 Uniform Understanding of Geometry of Abelian Varieties

Mathematicians have been seeking a common property among objects.

(a) Uniformization theorem of Riemann surfaces: Every simply connected Riemann surface is conformally equivalent to one of three Riemann surfaces: the open unit disk, the complex plane, or the Riemann sphere.

(b) Belyi's theorem: Every algebraic curve C defined over $\overline{\mathbb{Q}}$ is uniformized by a finite index subgroup of the modular group, i.e., there is a finite index subgroup Γ of $\mathrm{PSL}_2(\mathbb{Z})$ such that $\Gamma\backslash\overline{\mathcal{H}} \cong X(\mathbb{C})$, where $\mathcal{H}$ is the complex upper half plane.

Every abelian variety of dimension n over complex numbers is uniformized by a lattice of rank $2n$. The proof comes from the integration of a form $\omega \in \mathrm{H}^0(\mathrm{A}, \Omega^1_{\mathrm{A}})$ along a class $\gamma \in \mathrm{H}_1(\mathrm{A}(\mathbb{C}), \mathbb{Z})$ which gives rise to a morphism

$$\begin{aligned} \phi : \mathrm{H}_1(\mathrm{A}(\mathbb{C}), \mathbb{Z}) &\to \mathrm{LieA} \otimes_{\mathbb{Q}} \mathbb{C} \\ \gamma &\mapsto \left(\omega \mapsto \int_\gamma \omega \right) \end{aligned}$$

It turns out that the map ϕ is injective, and its corkernel is precisely $\mathrm{A}(\mathbb{C})$, i.e.,

$$\mathrm{A}(\mathbb{C}) \cong \frac{\mathrm{LieA} \otimes_{\mathbb{Q}} \mathbb{C}}{\phi(\mathrm{H}_1(\mathrm{A}(\mathbb{C}), \mathbb{Z}))} \cong \mathbb{C}^g / \Lambda$$

Question 1 *Do we have a similar uniformization for p-adic world?*

The first answer to the uniformization question of abelian varieties over p-adic fields was due to Tate [16], who invented a new theory over non-archimedean fields analogous to complex analytic spaces called rigid analytic spaces. In his foundational work, he showed that an elliptic curve with multiplicative reduction is uniformized as a rigid analytic space $\mathbb{C}_p^*/\mathbb{Z}$ where $\mathbb{C}_p$ is an analogue of $\mathbb{C}$ in p-adic context. Notably, in the International Congress of Mathematicians 1970, Raynaud [13] presented a program handling the case of abelian varieties: the uniformizing space $\mathbb{C}^g$ has to be replaced by an extension G of an abelian variety B (with good reduction) by a split affine torus T. The abelian variety A itself is a rigid geometric quotient G/Γ where now Γ is a lattice in G of rank $\dim T$.

Theorem 7 [Raynaud Uniformization] *Abelian variety A over $\mathbb{Q}_p$ = Quotient of semi-abelian variety G (i.e. an extension of a torus by an abelian variety) by a lattice Γ*

$$\begin{array}{ccccc} & & \Gamma & & \\ & & \downarrow & & \\ T & \longrightarrow & G & \longrightarrow & B \\ & & \downarrow & & \\ & & A & & \end{array}$$

In 2003, Fontaine [4] showed that we can recover the $\mathbb{C}_p$ points of A from just torsion points of $A(\overline{K})$. Following Fontain's program, Iovita, Morrow and Zaharescu [6] recently investigated Fontaine's integration [3] to provide a new type of p-adic uniformization of abelian varieties with good reduction (i.e. when reducing mod p we still obtain a smooth abelian variety), which resembles the classical complex uniformization.

Let A be an abelian variety over $\mathbb{Q}_p$ with good reduction. Denote by $\mathbb{C}_p := \widehat{\overline{\mathbb{Q}_p}}$ an analogue of $\mathbb{C}$ in p-adic world. Let $\mathbb{C}_p(1)$ be $\mathbb{C}_p$ as a set, together with cyclotomic action of $\mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p)$. Fontaine defined a p-adic integration map

$$\phi : T_p\mathrm{A} \to \mathrm{LieA} \otimes_{\mathbb{Q}_p} \mathbb{C}_p(1)$$

Proposition 4 *Under a mild condition, the Fontaine integration is an injection and its cokernel captures p-torsion and non-torsion points of the Abelian variety.*

Denote by

$$A^{(p)}(\overline{\mathbb{Q}}_p) = A(\overline{\mathbb{Q}}_p)/A_{p'-tor}(\overline{\mathbb{Q}}_p)$$

where $A_{p'-tor}(\overline{\mathbb{Q}}_p)$ is the prime-to-p torsion point of $A(\overline{\mathbb{Q}}_p)$.

Theorem 8 *Under a mild condition, there is an injective $\mathbb{Z}_p$-linear homomorphism which is $\mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p)$-equivariant*

$$\iota_A : A^{(p)}(\overline{\mathbb{Q}}_p) \hookrightarrow \frac{\mathrm{Lie}A \otimes_{\mathbb{Q}_p} \mathbb{C}_p(1)}{T_pA}$$

We can describe the image of ι_A as the class of crystalline elements of

$$\frac{\mathrm{Lie}A \otimes_{\mathbb{Q}_p} \mathbb{C}_p(1)}{T_pA}$$

Moreover, one can recover $\mathrm{A}^{(p)}(\overline{\mathbb{Q}}_p)$ from the triple

$$\big(T_pA,\ \mathrm{Lie(A)} \otimes_{\mathbb{Q}_p} \mathbb{C}_p(1), \phi : T_pA \hookrightarrow \mathrm{Lie(A)} \otimes_K \mathbb{C}_p(1)\big)$$

My doctoral thesis has investigated the p-adic uniformization of abelian varieties à la Iovita-Morrow-Zaharescu in semi-stable reduction cases (i.e. when reducing modulo p we get singularities). The main goal of my PhD work aims to develop a framework so that the method in good reduction cases can work effectively to a larger extent.

To deal with the semi-stable case, I have made use of the logarithmic geometry, which was initiated by Fontaine and Illusie in the 90s, and developed by Kato and others. Roughly speaking, the bad variety becomes good after adding a suitable logarithmic structure. The main contribution of my thesis is the notion of logarithmic Fontaine integration, which can be seen as an extending Fontaine integration in the sense of Iovita et al in semi-stable reduction cases.

Theorem 9 *Let A be a semi-stable abelian variety over K. There exists a G_K-equivariant morphism called logarithmic Fontaine integration (also denoted by ϕ)*

$$\phi : T_pA \to \mathrm{Lie}(A) \otimes_K \mathbb{C}_p(1)$$

Under the hypothesis $T_pA^{I_K} = 0$, the kernel of the logarithmic Fontaine integration is the prime-to-p torsion points of $\mathrm{A}(\overline{K})$, denoted by $\mathrm{A}_{p'\text{-}torsion}(\overline{K})$.

Assume that $T_pA^{I_K} = 0$, we obtain an inclusion called p-adic uniformization map

$$\iota_A : \mathrm{A}^{(p)}(\overline{K}) := \mathrm{A}(\overline{K})/\mathrm{A}_{p'\text{-}torsion}(\overline{K}) \hookrightarrow \frac{\mathrm{Lie(A)} \otimes_K \mathbb{C}_p(1)}{\phi(T_p\mathrm{A})} \tag{1}$$

The image of ι_A corresponds to semi-stable elements in $\frac{\mathrm{Lie(A)}\otimes\mathbb{C}_p(1)}{\phi(T_p\mathrm{A})}$. Moreover, one can recover $\mathrm{A}^{(p)}(\overline{K})$ from the triple

$$(T_pG,\ \mathrm{Lie(A)}\otimes_K \mathbb{C}_p(1),\ \phi : T_pG \to \mathrm{Lie(A)}\otimes_K \mathbb{C}_p(1))$$

The main ingredient in our construction is the theory of logarithmic Abelian varieties (à la Kato et al [7], [8], [9], [10]). The proof of the above theorem follows the strategy in [6].

Via the uniformization map, we can identify the p-divisible points on both sides, i.e., $\iota_A(\mathrm{A}(\overline{K})[p^\infty]) = \dfrac{\mathrm{Lie(A)}\otimes_K \mathbb{C}_p(1)}{\phi(T_p\mathrm{A})}[p^\infty]$. This will shed a new light on the study of non-torsion points of $A(\overline{K})$ which are mapped to certain classes of semi-stable elements in $\dfrac{\mathrm{Lie(A)}\otimes_K \mathbb{C}_p(1)}{\phi(T_p\mathrm{A})}$.

References

[1] Coleman, Robert and Iovita, Adrian, *The Frobenius and monodromy operators for curves and abelian varieties*. Duke Math. J. 97/1 (1999), 171–215.

[2] Faltings, Gerd, “Finiteness theorems for abelian varieties over number fields”. In "Arithmetic Geometry", Springer, 1986.

[3] Fontaine, Jean-Marc, *Formes différentielles et modules de Tate des variétés abéliennes sur les corps locaux*. Invent. Math. 65/3 (1981/82), 379–409.

[4] Fontaine, Jean-Marc, *Presque Cp-représentations*. Documenta Mathematica, Extra Volume Kato (2003), 283–383.

[5] Klein, Felix, *Vergleichende Betrachtungen über neuere geometrische Forschungen*. Mathematische Annalen 43/1 (1893), 63–100.

[6] Iovita, Adrian and Morrow, Jackson S. and Zaharescu, Alexandru, *On p-adic uniformization of abelian varieties with good reduction*. Compositio Math. 158/7 (2022), 1449–1476.

[7] Kajiwara, Takeshi and Kato, Kazuya and Nakayama, Chikara, *Logarithmic abelian varieties*. Nagoya Math. J. 189 (2008), 63–138.

[8] Kajiwara, Takeshi and Kato, Kazuya and Nakayama, Chikara, *Logarithmic abelian varieties, Part IV: Proper models*. Nagoya Math. J. 219 (2015), 9–63.

[9] Kajiwara, Takeshi and Kato, Kazuya and Nakayama, Chikara, *Logarithmic abelian varieties, part V: Projective models*. Yokohama Math. J. 64 (2018), 21–82.

[10] Kajiwara, Takeshi and Kato, Kazuya and Nakayama, Chikara, *Logarithmic abelian varieties, Part VI: Local moduli and GAGF*. Yokohama Math. J. 65 (2019), 53–75.

[11] Mazur, Barry, *Modular curves and the Eisenstein ideal*. Publ. IHES 47/1 (1977), 33–186.

[12] Mumford, David, "Abelian varieties". Tata Institute of Fundamental Research Studies in Mathematics 5, Corrected reprint of the 2nd (1974) edition.

[13] Raynaud, Michel, *Variétés abéliennes et géométrie rigide.* Actes du Congres International des Mathematiciens 1970 (1971), 473–477.

[14] Serre, Jean-Pierre and Tate, John, *Good Reduction of Abelian Varieties.* Annals of Mathematics 88/3 (1968), 492–517.

[15] Silverman, Joseph H., "The arithmetic of elliptic curves". Springer Graduate Texts in Math. 106 (2009).

[16] Tate, John, *Rigid analytic spaces.* Invent. Math. 12 (1971), 257–289.

[17] Weil, André, *Numbers of solutions of equations in finite fields.* Bull. AMS (1949), 497–508.

[18] Wiles, Andrew, *Modular elliptic curves and Fermat's last theorem.* Annals of Mathematics 141/3 (1995), 443–551.

Wasserstein Generative Models

Amna Mohsin (*)

Abstract. In these notes firstly, I will present an introduction about *optimal transport*. Nowadays optimal transport importance extends to diverse domains, ranging from mathematics and computer science to economics and image processing. Subsequently, I will talk about the Wasserstein distance, particularly the W_1 distance, which is a powerful metric for measuring the dissimilarity between probability distributions and it provides a more stable and meaningful measure than traditional metrics like the Kullback-Leibler divergence. This metric is used in particular within generative models, which are modern deep learning techniques that may be used to generate objects such as images, text, or any other structure. I will introduce these models and explain their application domain and discuss their properties, especially in relation to the limitation in their use of the W_1 distance. Then, I will talk about the main objective of the thesis, which builds upon prior work which introduce a dynamics-based method that allows us to obtain very accurate computations of the Wasserstein distance. The objective is to apply this method effectively within generative models to overcome the limitations in the traditional methods used to compute the W_1 distance, and how we expect this method to improve the models performances.

1 Optimal Transport

The optimal Transport problem was first introduced by Gaspard Monge in 1781, who considered the problem of transporting some material from an initial to a final configuration while minimizing the total transportation cost.

1.1 Monge Formulation

In the Monge formulation [16], the goal is to determine a transport map that minimizes the total cost or distance of moving mass from the source distribution to the target distribution. The transport map specifies how each point in the source distribution is mapped to a point in the target distribution.

Mathematically, given two non-negative measures measures f^+ and f^- with equal volume (we will denote with $\mathcal{M}_+(X)$ the set of the non-negative measures defined on a measure space X), the ambient spaces for the measures f^+, f^- are two complete and separable spaces X and Y , but in most of the cases we will assume $X = Y = \Omega$ where Ω

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: amna.mohsin@math.unipd.it. Seminar held on 13 December 2023.

is an open, bounded, convex, and connected domain in $\mathbb{R}^d$ and with smooth boundary. A map **transport map** $T : X \to Y$ is called a ***transport map*** from f^+ to f^- if:

$$f^- = T_{\#} f^+$$

The Monge formulation seeks to find a transport map T that minimizes the total transportation cost, it is as follows:

Minimize:

$$\mathbb{M}(T) = \int_X c(x, T(x))\, df^+(x) \tag{1}$$

among all transport maps T from f^+ to f^-. Here $c(x, y)$ is the cost of moving a unit of mass from point x to point y. In the original Monge problem the cost function was the distance $|x - y|$.

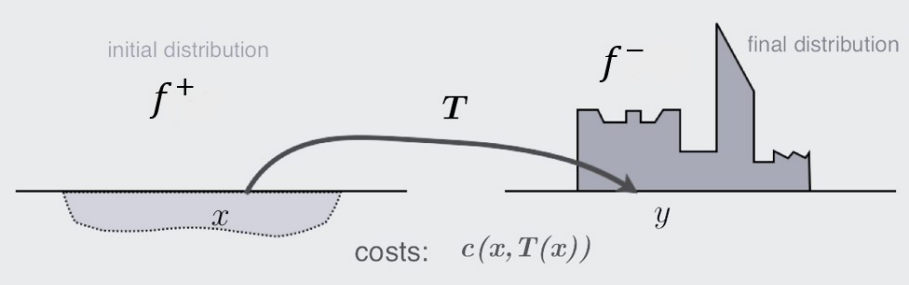


Figure 1: Monge's Optimal Transport Problem

Despite its historical significance, the Monge formulation can be ill-posed and the optimal map may not exist. For example when f^+ is a Dirac measure and f^- is not, the set of transport maps is empty since the image measure $T_{\#} f^+$ is atomic [16]. This has led to the development of alternative formulations, such as Kantorovich formulation, which relaxes some of the constraints and allows for multiple optimal transport plans.

1.2 Kantorovich formulation

After 150 years, Leonid Kantorovich revisited Monge's problem from a different view point [16]. He consider transport plans rather than transport maps. Given that π_X and π_Y are the projection maps on X and Y, respectively, we define the set of transport plans as follow: A probability measure $\gamma \in \mathcal{M}_+(X \times Y)$ is called a ***transport plan*** from f^+ to f^- if:

$$(\pi_X)_{\#}\gamma = f^+, \quad (\pi_Y)_{\#}\gamma = f^-$$

i.e., the marginals of γ are f^+ and f^-.

Denoting by $\mathcal{A}dm(f^+, f^-)$ the set of all transport plans from f^+ to f^-, the Kantorovich formulation is the following:

Minimize:

$$\mathbb{K}(\gamma) = \int_{X\times Y} c(x,y)\, d\gamma(x,y) \tag{2}$$

among all transport plans $\gamma \in \mathcal{A}dm(f^+, f^-) \subset \mathcal{M}_+(X \times Y)$ from f^+ to f^-.

Kantorovich Formulation is now known as the *Monge-Kantorovich* problem.

One of the key advantages of the Kantorovich relaxed formulation is that, under mild assumptions on the cost function c, it becomes relatively straightforward to establish the existence of the optimal transport plan. As stated in the following theorem

Theorem 1 *For any $c : X \times Y \to \mathbb{R}$ lower semi-continuous, then Kantorovich problem admits a solution $\gamma^* \in \mathcal{A}dm(f^+, f^-)$.*

The proof utilizes the direct method of the calculus of variations [14].

1.3 Kantorovich Dual Problem

The Kantorovich dual problem [14, 16] arises from the primal optimal transport problem, and it can be stated as follows. Given any two non-negative finite measures f^+ and f^- on X and Y satisfying $f^+(X) = f^-(Y)$, and given a cost function $c : X \times Y \to \mathbb{R}$. We can define the set $\mathcal{L}_c$ to be the set of all measurable functions $(u, v) \in L^1(\, df^+) \times L^1(\, df^-)$ satisfying

$$u(x) + v(y) \leq c(x,y) \tag{3}$$

The problem is then to find $(u^*, v^*) \in \mathcal{L}_c$ that solves the maximization problem

$$\sup_{(u,v)\in\mathcal{L}_c} \mathbb{I}[u,v] := \int_X u(x)\, df^+(x) + \int_Y v(y)\, df^-(y) \tag{4}$$

And we have the following duality theorem [16]:

Theorem 2 (Kantorovich Duality) *Given two non-negative finite measures f^+ and f^- on X and Y satisfying $f^+(X) = f^-(Y)$, and a lower semi-continuous cost function $c : X \times Y \to \mathbb{R}$, the following equality holds*

$$\min_{\gamma\in\mathcal{A}dm(f^+,f^-)} \mathbb{K}(\gamma) = \max_{(u,v)\in\mathcal{L}_c} \mathbb{I}[u,v]$$

We give the following definition that is necessary for the existence result below

Definition 1 Given a function $\chi : X \to \overline{\mathbb{R}}$ we define its $c-$transform $\chi^c : Y \to \overline{\mathbb{R}}$ by

$$\chi^c(y) = \inf_{x\in X} c(x,y) - \chi(x)$$

We also define the $\overline{c}-$transform of $\zeta^{\overline{c}} : Y \to \overline{\mathbb{R}}$ by

$$\zeta^{\overline{c}}(x) = \inf_{y \in Y} c(x,y) - \zeta(y)$$

We say a function v defined on Y is $\overline{c}-$concave if there exists χ such that $v = \chi^c$, similarly, a function u on X is said to be $c-$concave if there is ζ such that $u = \zeta^{\overline{c}}$. And we denote by $c-conc(X)$ and $\overline{c}-conc(Y)$ the sets of $c-$ and $\overline{c}-$concave functions, respectively [14].

Remark 1 As we are interested in the case where $c = d$, it is worth to mention this property about the c-transform of 1-Lipschitz functions, if f is 1-Lipschitz, then $f^c = -f$ ([15], sec. 5).

A consequence of these considerations is the following existence result.

Proposition 1 *Assume that X, Y are compact and the function c is continuous then the dual problem admits a solution pair $(u^*, v^*) \in \mathcal{L}_c$ with $v^* = (u^*)^c \in \overline{c} - conc(Y)$. This means that the dual problem can be rewritten as:*

$$\sup_{u \in c-conc(X)} \int_X u(x)\, df^+(x) + \int_Y (u^c)\, df^-(y)$$

The function u^* is called Kantorovich Potential.

1.4 Wasserstein Distance

The Wasserstein distance, also called Earth Mover's Distance (EM) or Kantorovich-Rubinstein distance, is a metric used to measure the difference between two probability distributions. Unlike traditional metrics, such as the Euclidean distance or the Kullback-Leibler divergence, it considers both distribution values and spatial relationships between elements. This is valuable for comparing distributions representing mass, probability, etc. In the context of optimal transport theory, it finds the most efficient way to transform one distribution into another while minimizing transportation cost. We mainly consider costs of the form $c(x,y) = |x-y|^p$ in $\Omega \subset \mathbb{R}^d$.

Here we give the basic definitions of Wasserstein distances and spaces in $\Omega \subset \mathbb{R}^d$ [14]. When Ω is unbounded we need to restrict to the following set of probabilities, let $p \in [1, +\infty)$ we denote by $\mathcal{P}(\Omega)$ the set of probability measures on Ω

$$\mathcal{P}_p(\Omega) := \left\{ \mu \in \mathcal{P}(\Omega) : \int_\Omega |x|^p \, d\mu < +\infty \right\}.$$

For $f^+, f^- \in \mathcal{P}_p(\Omega)$, the Wasserstein metric is defined as follows:

$$W_p(f^+, f^-) := \min \left\{ \int_{\Omega \times \Omega} |x-y|^p \, d\gamma : \gamma \in \mathcal{A}dm(f^+, f^-) \right\}^{\frac{1}{p}}, \tag{5}$$

In this work we are interested in the case when $c(x, y) = |x - y|$, hence we have

$$W_1(f^+, f^-) := \min\left\{ \int_{\Omega\times\Omega} |x - y|\, d\gamma : \gamma \in \mathcal{A}dm(f^+, f^-) \right\} \tag{6}$$

1.5 Monge-Kantorovich equations

Before we can dive into Monge Kantorovich equations, we need to define a quantity called *Optimal Transport Density*. We give the definition as stated in [5].

Definition 2 (Optimal Transport Density) Consider $\Omega \subset \mathbb{R}^d$ an open, bounded, connected, and convex domain with smooth boundary. Take two nonnegative measures f^+ and f^- on Ω such that $df^+(\Omega) = df^-(\Omega)$. Given $\gamma^* \in \mathcal{A}dm(f^+, f^-)$ a minimizer for the Kantorovich Problem in (2) with cost $c(x, y) = |x - y|$, the *Optimal Transport Density* (OT density) $\mu^* \in \mathcal{M}_+(\Omega)$ associated to f^+, f^- is defined as

$$\langle \mu^*, \phi \rangle := \int_{\Omega\times\Omega} \int_0^1 |\omega'_{x,y}(t)| \phi(\omega_{x,y}(t))\, dt\, d\gamma(x, y) \qquad \forall \phi \in \mathcal{C}(\Omega) \tag{7}$$

where

$$\omega_{x,y}(t) = (1 - t)x + ty.$$

In the following proposition we can see the importance of the OT density in the theory of L^1-OTP [4] (Optimal Transport proplem in the case $c(x, y) = |x - y|$).

Proposition 2 *Consider $\Omega \subset \mathbb{R}^d$ an open, bounded, connected, and convex domain with smooth boundary. Take two nonnegative measures f^+ and f^- on Ω such that $df^+(\Omega) = df^-(\Omega)$. If f^+ and f^- admit L^p-densities with $1 \le p \le +\infty$, a solution v^* of the Beckmann Problem belongs to $[L^p(\Omega)]^d$ and it can be written as*

$$v^* = -\mu^* \nabla u^* \tag{8}$$

where μ^* is the OT density $\mu^*(f^+, f^-)$ and u^* is a solution of the Dual Kantorovich Problem.

We now come to the most interesting finding of L^1-OTP, which asserts that the optimal transport density can be characterized by the a system of equations, known as the Monge-Kantorovich equations (MK equations), see [4]. The authors illustrate that by examining the p-Laplacian equation below, it is possible to construct a solution for the classical Monge-Kantorovich problem

$$-\mathrm{div}(|Du_p|^{p-2} Du_p) = f^+ - f^- \tag{9}$$

as p approaches infinity, the objective is to demonstrate the convergence of u_p towards u, where u is a solution that satisfies

$$|Du| \le 1, \quad -\mathrm{div}(aDu) = f^+ - f^- \tag{10}$$

with a non-negative density a, the goal is to construct a flow by solving an ordinary differential equation that incorporates a, Du, f^+, and f^-. So here comes the following proposition of Monge-Kantorovich Equations

Proposition 3 (Monge-Kantorovich Equations) *Consider $\Omega \subset \mathbb{R}^d$ an open, bounded, connected, and convex domain with smooth boundary. Take two nonnegative measures f^+ and f^- on Ω such that $df^+(\Omega) = df^-(\Omega)$. Assume that f^+ and f^- admit L^p-densities. The OT density $\mu^*(f^+, f^-)$ and Kantorovich potential u^* solve the following equations*

$$-div(\mu^* \nabla u^*) = f \quad in\ \Omega \tag{11}$$

$$|\nabla u^*| \leq 1 \quad in\ \Omega \tag{12}$$

$$|\nabla u^*| = 1 \quad a.e.\ in\ \mu^* > 0 \tag{13}$$

with $f = f^+ - f^-$.

1.6 Dynamic Monge-Kantorovich

In [5, 6, 7], the authors developed a dynamic formulation of the Monge-Kantorovich equations which enables an efficient and accurate numerical solution of Optimal transport problems. In particular, the DMK (Dynamic Monge-Kantorovich) equations are given as follows:

Given a domain $\Omega \subset \mathbb{R}^d$, two positive functions f^+ and f^- in $L^1(\Omega)$ such that $\int_\Omega f^+ dx = \int_\Omega f^- dx$, find the pair $(\mu, u) : [0, +\infty) \times \Omega \to \mathbb{R}^+ \times \mathbb{R}$ that satisfies:

$$\begin{aligned} &- \operatorname{div}(\mu(t,x)\nabla u(t,x)) = f(x) = f^+(x) - f^-(x) \\ &\partial_t \mu(t,x) = \mu(t,x)\left(|\nabla_x u(t,x)| - 1\right) \\ &\mu(0,x) = \mu_0(x) > 0. \end{aligned}$$

The infinite-time solution of this problem is conjectured to be exactly the solution pair (μ^*, u^*) of the Monge-Kantorovich equations. In support of this conjecture, the authors were able to show local existence and uniqueness of the solution of the related ODE in Banach spaces [6].

Lyapunov-candidate An efficient numerical solver for the DMK is proposed in [7]. In addition, the authors introduce a *Lyapunov-candidate* functional for the system above and they were able to show that this functional admits the optimal transport density as a unique minimizer and more important for us, this proposed Lyapunov-candidate functional can be effectively used to calculate the Wasserstein-1 distance between two measures.

The Lyapunov-candidate functional $\mathcal{L}$ is formed by the sum of an energy functional $\mathcal{E}_f$ and a mass functional $\mathcal{M}$ and for general $\mu \in L^1(\Omega)$ and given by:

$$\mathcal{L}(\mu) := \mathcal{E}_f(\mu) + \mathcal{M}(\mu),$$

where

$$\mathcal{E}_f(\mu) := \sup_{\phi \in \mathcal{C}^1(\overline{\Omega})} \int_\Omega \left(f\phi - \mu \frac{|\nabla \phi|^2}{2} \right) dx \qquad \mathcal{M}(\mu) := \frac{1}{2} \int_\Omega \mu \, dx.$$

Now for a given μ we are interested in the following proposition proven in [7].

Proposition 4 *Given $f = f^+ - f^- \in L^1(\Omega)$ with zero mean, then the optimal transport density $\mu^*(f)$ is a minimizer for $\mathcal{L}$ with value equal to the Wasserstein-1 distance between f^+ and f^- (namely, $\mathcal{L}(\mu^*) = W_1(f^+, f^-)$).*

2 Generative Adversarial Networks

Generative Adversarial Networks (GANs) represent a revolutionary class of artificial intelligence algorithms that have taken the field of machine learning by storm. Introduced by Ian Goodfellow and his colleagues in 2014 [8], GANs have demonstrated impressive capabilities in generating realistic and diverse data, such as images, audio, and even text.

The GAN framework involves a dynamic interaction between the generator and discriminator networks. The generator's objective is to produce data instances that are indistinguishable from real data, whereas the discriminator is responsible for identifying genuine data from generated data. This process resembles a game between the two networks, where the generator constantly improves its ability to produce realistic data in response to the feedback provided by the discriminator. Over time, this iterative process ideally leads to the generation of data that is so realistic that even the discriminator struggles to tell it apart from real data. However, achieving this balance can be challenging and requires careful tuning of hyperparameters, architecture, and training techniques.

GANs have been applied to a wide array of tasks with impressive results. As in many different AI frameworks such as, photo realistic image generation [3], image-to-image translation [18], text-to-image translation [17], photo editing [13], super resolution [10], and image inpainting [12].

Despite their remarkable capabilities, GANs do come with challenges such as training instability and mode collapse, which we will be discussing in this section.

2.1 First introduction to GANs

In the paper [8], the authors investigate a specific scenario where the generative model produces samples through a multilayer perceptron by introducing random noise. Likewise, the discriminative model is constructed as a multilayer perceptron. In this particular instance, both models can be trained exclusively using the well-established backpropagation technique. Additionally, we can extract samples from the generative model using forward propagation.

The adversarial game described above can be formulated *mathematically* by minimax of a target function between the *discriminator function* $D(x;\theta_d)$ where $D : \mathbb{R}^d \to [0,1]$ is a differentiable function depends on parameters θ_d, D yields a single scalar output, and the *generator function* $G(z;\theta_d)$ where $G : \mathbb{R}^d \to \mathbb{R}^d$ is also a differentiable function. $D(x)$ signifies the likelihood that x originated from the data rather than p_g. The training involves maximizing D's accuracy in labeling both training examples and samples from G. Simultaneously, G is trained to minimize $\log(1 - D(G(z)))$. In order to grasp the distribution p_g of data x generated by the generator, we define a prior input noise variables

$p_z(z)$. In [8] the target *loss function* is proposed to be:

$$V(D,G) = \mathbb{E}_{x\sim p_{data}} \log D(x) + \mathbb{E}_{z\sim p(z)} \log(1 - D(G(z))), \tag{14}$$

where $\mathbb{E}_\mu(f) = \int f d\mu$ is the expectation functional of f with respect to the probability measure μ. Hence, GANs solve the minimax problem:

$$\min_G \max_D V(D,G) = \min_G \max_D \left\{ \mathbb{E}_{x\sim p_{data}} \log D(x) + \mathbb{E}_{z\sim p(z)} \log(1 - D(G(z))) \right\}. \tag{15}$$

Now we address some theoretical results stated in [8], and, for more details for the proofs, refer to their paper.

Proposition 5 *For G fixed, the optimal discriminator D is*

$$D^*_G(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)}.$$

Setting $C(G) = \max_D V(G,D)$, we have the following theorem.

Theorem 3 *The global minimum of the virtual training criterion $C(G)$ is achieved if and only if $p_g = p_{data}$. At that moment, $C(G)$ attains a value of $-\log 4$.*

As in the proof of the above theorem, with simple calculations we get:

$$C(G) = -\log(4) + 2JSD(p_{data}|p_g), \tag{16}$$

where JSD is the Jensen-Shannon divergence. Hence, GANs are actually minimizing the Jensen-Shannon divergence between p_{data} and p_g.

Once the optimization of G and D is completed, the generator G can be used to sample objects by first sampling z according to the distribution of the input, and then computing $G(z)$.

2.2 Understanding GANs training dynamics

As we said earlier, GANs are still very difficult to train. In [1] authors focus on rigorous examination of issues like instability, saturation during GANs training, and mode collapse (it occurs when the generator of the GAN starts producing a limited set of very similar or identical samples). As refer in the paper [8], when the discriminator gets better, the updates to the generator get consistently worse. This situation can only occur when the distributions are not continuous or possess separate supports. A potential reason for the distributions lacking continuity is if their supports are lying on low-dimensional manifolds.

In theorem 2.1, 2.2 in [1], they prove that in both cases where the two distributions have disjoint support and where their support lie on low-dimensional manifolds, then there is a **perfect discriminator** which are smooth and constant almost everywhere (by perfect discriminator, we mean $p_{data}[D(x) = 1] = 1$ and $p_g[D(x) = 0] = 1$.

The first issue that arise is that, as the discriminator gets better, the gradient of the generator vanishes (as stated and proved in Theorem 2.4 in [1]. So, to prevent gradient vanishing in situations where the discriminator is highly confident, a different gradient step is used for the generator, as follow

$$\nabla_{\theta_g} \mathbb{E}_{z\sim p(z)} \left[-\log D(G(z)) \right].$$

Running experiments with this $-\log D$ cost function, authors observed that the gradient norms grow quickly (for details and proof see Theorem 2.6 in [1]).

The authors in [2] suggest to use and alternative divergence or distance for the GANs, they chose to use Wasserstein distance (see (5)), and it was the first time to introduce *Wasserstein GANs* (WGANs).

2.3 First introduction to WGANs

WGANs address the primary training challenges encountered in traditional GANs. Specifically, when training WGANs, there is no need to carefully manage the balance between training the discriminator and generator. Additionally, WGANs reduce the common issue of mode collapse that is frequently observed in GANs. The provided example in [2] (Example 1) demonstrates how seemingly straightforward sequences of probability distributions converge when considering the Wasserstein distance, but fail to converge when using different distances and divergences (the comparison was between Wasserstein distance, Total Variation distance, Kullback-Leibler divergence, and Jensen-Shannon divergence).

With mild conditions we can show that Wasserstein distance $W(p_{data}, p_{\theta_g})$ is a continuous loss function on θ_g, which makes it much more sensible cost function for the GANs problem than at least the Jensen-Shannon divergence, as we see in the following theorem, but first we need to set the following assumption [2]

Assumption 1. Let $g : \mathcal{Z} \times \mathbb{R} \to \mathcal{X}$ be locally Lipschitz between finite dimensional vector spaces. We will denote $g_\theta(z)$ it is evaluation on coordinates (z, θ). We say that g satisfies Assumption 1 for a certain probability distribution p over $\mathcal{Z}$ if there are local Lipschitz constants $L(\theta, z)$ such that

$$\mathbb{E}_{z\sim p} \left[L(\theta, z) \right] < +\infty.$$

Theorem 4 *Consider a fixed distribution p_r defined over the space $\mathcal{X}$. Additionally, let Z be a random variable, for example, following a Gaussian distribution, defined over a separate space denoted as $\mathcal{Z}$. We have a function $g : \mathcal{Z} \times \mathbb{R}^d \to \mathcal{X}$, which we will denote as $g_\theta(z)$, where z represents the first coordinate and θ the second. Let p_θ represent the distribution of the random variable $g_\theta(Z)$. Then:*

- *If g is continuous in θ, so is $W(p_r, p_\theta)$.*
- *If g is locally Lipschitz and satisfies regularity Assumption 1, then $W(p_r, p_\theta)$ is continuous everywhere, and differentiable almost everywhere.*

- *The statements above are false for the Jensen-Shannon divergence $JS(p_r, p_\theta)$ and all the KLs*

For more details see [2].

This makes Wasserstein distance to have nicer properties when optimized. Nonetheless, finding the infimum in equation (6) poses a significant computational challenge. So let us recall the Kantorovich-Rubinstein duality [15] (also see (4)).

$$W_1(p_{data}, p_{\theta_g}) = \sup_{||f||_L \leq 1} \mathbb{E}_{x \sim p_{data}} [f(x)] - \mathbb{E}_{x \sim p_{\theta_g}} [f(x)] . \tag{17}$$

Moreover, the topology induced by the Wasserstein distance is the weakest among the other divergences and distances, giving it the edge for practical real-world applications. We state the following theorem (detailed proof can be found in [2]).

Theorem 5 *Let P denote a distribution defined on a compact space $\mathcal{X}$, and let $(P_n)_{n \in \mathbb{N}}$ be a sequence of distributions on the same space $\mathcal{X}$. Considering all limits as n approaches infinity, the following statements hold:*

(a) *The following statements are equivalent:*

- *$\delta(P_n, P) \to 0$ with δ the total variation distance.*
- *$JS(P_n, P) \to 0$ with JS the Jensen-Shannon divergence.*

(b) *The following statements are equivalent:*

- *$W(P_n, P) \to 0$.*
- *$P_n \xrightarrow{\mathcal{D}} P$ where $\xrightarrow{\mathcal{D}}$ represent convergence in distribution for random variables*

(c) *$KL(P_n||P) \to 0$ or $KL(P||P_n) \to 0$ imply the statement in* (a).

(d) *The statements in* (a) *imply the statements in* (b).

We now address the following theorem which consider differentiating $W_1(p_r, p_\theta)$ [2].

Theorem 6 *Let p_r be any distribution. Consider p_θ as the distribution of $g_\theta(Z)$, where Z is a random variable with density p and g_θ a function satisfying assumption 1. Then, there is a solution $f : \mathcal{X} \to \mathbb{R}$ to the problem*

$$\max_{||f||_L \leq 1} \mathbb{E}_{x \sim p_r} [f(x)] - \mathbb{E}_{x \sim p_\theta} [f(x)] ,$$

and we have

$$\nabla_\theta W_1(p_r, p_\theta) = \mathbb{E}_{z \sim p(z)} [\nabla_\theta f(g_\theta(z))]$$

when both terms are well-defined.

Observe that when we substitute the condition $||f||_L \leq 1$ with $||f||_L \leq K$ (K-Lipschitz with a constant K), we obtain $W_1(p_r, p_\theta)$ up to a multiplicative constant. Consequently, if we work with a parametric set of functions $\{f_w\}_{w\in\mathcal{W}}$ that all satisfy the K-Lipschitz condition for some K, we can work with the following problem:

$$\max_{w\in\mathcal{W}} \mathbb{E}_{x\sim p_r}\left[f_w(x)\right] - \mathbb{E}_{z\sim p(z)}\left[f_w(g_\theta(z))\right]. \tag{18}$$

We call f_w the 'critic', it is an analogue to the discriminator in GANs. In order to maintain parameters w lie in a compact space $\mathcal{W}$ and hence enforce the Lipschitz constraint, in [2] they use *weight clipping* after each gradient update.

The very first benefit of WGANs is that it provides an estimate of EM that correlates well with the quality of the generated samples. And the better the critic gets, the higher quality the gradients we use to train the generator. So we no longer need to balance generator and discriminator's training [2].

2.4 Improving WGANs Training

The proposed WGAN represents a step towards achieving more stable GAN training. However, it can encounter issues where it generates low-quality samples or struggles with convergence. In [9] they have identified that these problems frequently arise from the employment of weight clipping in WGAN, which is employed to enforce a Lipschitz constraint on the critic. The authors propose an alternative way to enforce the Lipschitz constraint, which is through *gradient penalty*. A function is considered 1-Lipschitz if and only if its gradients have a magnitude no greater than 1 everywhere. Therefore, they consider the direct restriction of the gradient norm of the critic's output with respect to its input. The new objective functional as in [9] is

$$L = \mathbb{E}_{x\sim p_g}\left[f(x)\right] - \mathbb{E}_{x\sim p_r}\left[f(x)\right] + \lambda\, \mathbb{E}_{z\sim p(z)}\left[(||\nabla_z f(z)||_2 - 1)^2\right] \tag{19}$$

where the first two terms is the original critic loss, and the last term is the gradient penalty.

The proposed approach offers an edge over weight clipping by enhancing both the training speed and the quality of generated samples. And it preserves the property of correlating the sample quality and the convergence toward a minimum.

2.5 How Accurately Do WGANs Approximate the Wasserstein Distance?

As we have seen, Wasserstein distance is expressed using Kantorovich duality as the difference between expected values of a potential function under real data and model distributions. As mentioned in [11], this introduces at least three sources of errors in the approximation of the Wasserstein distance: the approximated discriminator and constraints, the estimation of the expectation value, and the optimization. The authors consider the c-transform formulation, which allows for more accurate estimation of the Wasserstein distance. But the surprising finding in their research is that the approach that most accurately approximates the Wasserstein distance does not yield the best looking images in the generative setting.

To describe the objective function in the $c-$transform technique of approximating the Wasserstein distance, recall Definition 1, then we can compute the $c-$transform over the minibatches as

$$\phi_\omega^c(y_i) \approx \widehat{\phi_\omega^c}(y_i) = \min_j \{c(x_j, y_i) - \phi_\omega(x_j)\} \tag{20}$$

The objective is written as

$$\max_\omega \left\{ \frac{1}{N} \sum_{i=1}^{N} \phi_\omega(x_i) + \frac{1}{N} \sum_{i=1}^{N} \widehat{\phi_\omega^c}(y_i) \right\} \tag{21}$$

Authors in [11] runs many experiments, and the outcomes clearly demonstrate that, at each iteration, the $c-$transform provides more accurate estimates of the Wasserstein distances compared to gradient penalty or weight clipping methods. However, the resulting images are blurry, whereas the highest-quality images are generated using the gradient penalty method, while weight clipping has not yet reached a convergence.

The authors raises many questions upon the results they got. First, the question of whether the precise Wasserstein-1 distance between batches is indeed the quantity that should be considered in the context of generative modeling?

Furthermore, it is intriguing to observe how the gradient penalty method excels in the generative scenario, despite its somewhat less precise approximation of the Wasserstein-1 distance, as indicated by their experiments. This raises the question of what attributes make it such an effective objective in the generative context?

In our work we want to use the formula of calculating the Wasserstein-1 distance as in Proposition 4, and see how it will impact the performance of WGANs.

References

[1] M. Arjovsky and L. Bottou, *Towards principled methods for training generative adversarial networks*. arXiv preprint arXiv:1701.04862 (2017).

[2] M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein generative adversarial networks*. In International conference on machine learning, pp. 214–223. PMLR (2017).

[3] A. Brock, J. Donahue, and K. Simonyan, *Large scale gan training for high fidelity natural image synthesis* (2018).

[4] L.C. Evans and W. Gangbo, "Differential equations methods for the Monge-Kantorovich mass transfer problem". American Mathematical Soc., 1999.

[5] E. Facca, F. Cardin and M. Putti, *Biologically inspired formulation of optimal transport problems*.

[6] E. Facca, F. Cardin, and M. Putti, *Towards a stationary Monge-Kantorovich dynamics: The physarum polycephalum experience.* SIAM Journal on Appl. Math. 78/2 (2018), 651–676.

[7] E. Facca, S. Daneri, F. Cardin, and M. Putti, *Numerical solution of Monge-Kantorovich equations via a dynamic formulation.* Journal of Scientific Computing 82/3 (2020), 1–26.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D.Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets in advances in neural information processing systems* (nips) (2014).

[9] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, *Improved training of Wasserstein gans* (2017).

[10] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, *Photo-realistic single image super-resolution using a generative adversarial network* (2016).

[11] A. Mallasto, G. Montúfar, and A. Gerolin, *How well do wgans estimate the Wasserstein metric?.* arXiv preprint arXiv:1910.03875 (2019).

[12] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A.A. Efros, *Context encoders: Feature learning by inpainting* (2016).

[13] G. Perarnau, J. van de Weijer, B. Raducanu, and J.M. Álvarez, *Invertible conditional gans for image editing* (2016).

[14] F. Santambrogio, "Optimal transport for applied mathematicians". Progress in Nonlinear Differential Equations and Their Applications 87, Birkhäuser, NY, 2015.

[15] C. Villani, "Optimal Transport". Volume 338 of Old and New. Springer Science & Business Media, Berlin, Heidelberg, 2008.

[16] C. Villani, "Topics in optimal transportation". Volume 58. American Mathematical Soc., 2021.

[17] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, *Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks* (2016).

[18] J.-Y. Zhu, T. Park, P. Isola, and A.A. Efros, *Unpaired image-to-image translation using cycle-consistent adversarial networks* (2017).

Double-negation in the foundation of Constructive Mathematics

Pietro Sabelli [(*)]

1 An introduction to Constructive Mathematics

An *existential statement* is a proposition asserting the existence of a mathematical object satisfying some property. Consider for example the following.

Proposition 1 *There exist two irrational numbers a and b, such that a^b is rational.*

We now give two different proofs of the above statement, one will be called *classical*, and the other one *constructive*.

Proof (Classical). Consider $\sqrt{2}^{\sqrt{2}}$, if it is rational then we are done by taking $a = b = \sqrt{2}$; otherwise, take $a = \sqrt{2}^{\sqrt{2}}, b = \sqrt{2}$ and observe that $a^b = 2$. □

Proof (Constructive). Take $a = \sqrt{2}, b = \log_2(9)$ and observe $a^b = 3$. □

Notice how, following the classical proof, we did not come to know an explicit pair of rational numbers a, b with the desired property; we would if we were able to determine if $\sqrt{2}^{\sqrt{2}}$ is rational or not, but the proof is not informative about that. The constructive proof, on the other hand, leave us without uncertainties.

As a first approximation, one could say that *constructive mathematics*, as opposed to *classical mathematics*, is the mathematical discipline which regards as legitimate only the following kind of proofs of existence.

Definition 1 [Constructive proof - first draft] A proof of an existential statement is *constructive* if it explicitly exhibit the object with the desired properties specified by the statement.

Already from this simple definition we could see how the field of constructive mathematics focuses more on how we *prove* a statement, rather than its absolute *truth*. Proving

[(*)]Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: sabelli@math.unipd.it . Seminar held on 8 November 2023.

a theorem constructively rather then classically can be radically harder. Consider the following well-known example of a statement having a simple classical proof while being constructively still an open problem.

Proposition 2 *There exists a digit which occurs infinitely often in the decimal expansion of π.*

Proof (Classical). By contradiction, suppose that each digit occurs finitely many times, then the decimal expansion would be finite, but it is not since π is irrational. □

Notice again how, from the above proof, we still do not know *which is* the digit occuring infinitely many times.

The next step consists of generalising Definition 1 to proofs of arbitrary statements. We do so by introducing a new property, which is sometimes called *effectiveness*, they need to satisfy in order to be considered constructive.

Definition 2 [Constructive proof - second draft] A proof is *constructive* if it enjoys a *computational interpretation*, that is if we can interpret it as an algorithm to obtain the data specified by the statement it proves.

The above definition is rather vague and philosophical. To make it more precise and apt to (meta-)mathematical investigations, we refer to the semi-formal definition of constructive proof known as the *Brouwer-Heyting-Kolmogorov interpretation* which, by treating proofs themselves as mathematical objects (a common thing to do in logic), specifies what is a constructive proof in predicate logic.

Definition 3 [Constructive proof - BHK Interpretation]

(a) a constructive proof of a conjunctive statement $\varphi \wedge \psi$ consists of a pair of a proof of φ and a proof of ψ.

(b) a constructive proof of a disjunctive statement $\varphi \vee \psi$ consists of either a proof of φ, or one of ψ, and the choice must be explicit.

(c) a constructive proof of an implicative statement $\varphi \Rightarrow \psi$ consists of *an algorithm* that, given as input a proof of φ, returns a proof of ψ as output.

(d) a constructive proof of a negated statement $\neg\varphi$ consists of *an algorithm* that, given as input a proof of φ, returns a proof of $0 = 1$ as output.

(e) a constructive proof of an existential statement $\exists x \in A.P(x)$ consists of *a witness* $a \in A$ and a proof of $P(a)$;

(f) a constructive proof of a universal statement $\forall x \in A.P(x)$ consists of *an algorithm* that, given $a \in A$ as input, returns a proof of $P(a)$ as output.

Notice how point 5 actually generalises Definition 1 by requiring to exhibit a concrete witness to an existential statement.

Of course the specification above is still somewhat informal, mainly because it relies on the philosophical notion of *algorithm*, and treats proofs as possible data that can be in some way manipulated by them. We refer the interested reader to the *Kleene interpretation of constructive arithmetic* [3] as a way of fully formally implement the above specifications by encoding proofs as Turing machines. As far as we are concerned, it will be sufficient to regard an *algorithm* as a series of instructions which can be implemented as a computer program. The rest of this section is devoted to present examples of the BHK interpretation which will hopefully help to develop intuition.

Example 1 Consider the following formulation of Euclid's theorem on the infinity of prime numbers.

$$\forall n \in \mathbb{N}\, \exists p \in \mathbb{N}\, (\, p \text{ prime} \wedge p > n\,)$$

According to the BHK interpretation, a constructive proof of the above statement consists of an algorithm that, given a natural number n, outputs another natural number p, which is prime and greater than n.

We can thus phrase the original Euclid's proof to show that is indeed constructively valid.

Proof (Constructive). Given n, take as p an arbitrary prime factor (e.g. the smallest) of $(n! + 1)$. □

In the next examples we will instantiate the following pattern. Given a set A and a property $P(x)$ on the elements of A, we will consider the statement *for each element $a \in A$, either $P(a)$ holds or it does not.*

$$\forall a \in A\, (\, P(a) \vee \neg P(a)\,)$$

Notice that in classical mathematics, the above kind of statements are trivially true. However, to constructively prove them, we must provide an algorithm that, for each element $a \in A$, *decides* whether or not $P(a)$ holds. As the following two examples show this is not always possible.

Example 2

$$\forall n \in \mathbb{N}\, (\, n \text{ is even} \vee n \text{ is odd}\,)$$

A possible algorithm - that is a possible constructive proof - is the following: divide n by 2 and check if there is a remainder.

The next example shows that the gap between classically proving a theorem and constructively proving the same theorem can be infinite, in the sense that there exist statements which have a trivial classical proof, while being *unprovable* with constructive means.

Example 3

$$\forall f : \mathbb{N} \to \mathbb{N}\, (\exists n \in \mathbb{N}\,.\, f(n) = 0 \vee \neg \exists n \in \mathbb{N}\,.\, f(n) = 0)$$

A constructive proof of the statement asks, according to the BHK interpretation, to determine with an algorithm whether a given function $f : \mathbb{N} \to \mathbb{N}$ has a root. But this is

impossible, because there cannot exist an algorithm (think of it as a computer program) which checks the infinitely many values which f may assume.

Be aware that the fact that a theorem is *constructively unprovable* does *not* mean that it is *constructively false* (whatever it may mean).

The following example shows that an appropiate weakening of the previous example's statement is constructively provable.

Example 4

$$\forall f : \mathbb{N} \to \mathbb{N}\, \forall n \in \mathbb{N}\, (\exists m < n\,.\, f(m) = 0 \vee \neg \exists m < n\,.\, f(m) = 0)$$

It is constructive provable because this time we only need to check a fixed number, namely n, of values of the function f, and this is something which can be easily implemented.

The last two examples are of particular importance, because their statements are classically equivalent, but constructively very different. The trick of changing a statement into a classically equivalent one, but constructively weaker will be the guiding idea of the double-negation translation.

We end this section by giving an example involving point 3 of Definition 3, which can be a bit puzzling at first.

Example 5 Consider the ordinary proof of irrationality of $\sqrt{2}$.

$$\neg\, \exists a, b \in \mathbb{N}\,.\, \sqrt{2} = \frac{a}{b}$$

Proof (Constructive). Suppose there exist natural numbers a and b such that $\sqrt{2} = \frac{a}{b}$. Then, $a^2 = 2b^2$. The prime factor 2 appears an even number of times on the left and an odd number of times on the right. Then we can conclude $0 = 1 \pmod 2$, a contradiction. □

It is a constructively valid *proof of a negation* which, from some data (namely the numbers a, b with the fact that $a/b = \sqrt{2}$) performs an algorithmic manipulation of those data in order to get a contradiction.

2 Foundations of Constructive Mathematics

A foundational system for mathematics (whether classical or constructive) should take care of the formalisation of two broad areas: *logic*, that is how one reasons about mathematical objects, and the nature of the mathematical objects themselves, which we should call for historical reasons *set theory*.

Consider, for example, two famous foundational systems: Peano Arithmetic and Zermelo-Fraenkel set theory. The first was designed as a foundational system for a circumscribed mathematical field, namely arithmetic. The second is instead regarded as the current standard for the foundation of all classical mathematics. Their respective set theories are very different: Peano arithmetic limits to introduce the natural numbers, while Zermelo-Fraenkel relies on the ductile notion of Cantorian set to encode every possible mathematical

object. However, being both foundations for classical mathematics, although different in scope, they have the same underlying logic, namely *classical*, *predicate* (also known as *first-order*) *logic*.

We will see that, compared to the situation in classical mathematics, while in constructive mathematics the task of formalising its logic was settled in the 30s, the search for a general set-theoretic standard akin to Zermelo-Fraenkel for classical mathematics is still ongoing.

In the following, we first introduce the reader to the particular version of logic, called *intuitionistic*, which suits constructive mathematics and which became a solid standard, so much so that constructive mathematics is often made to coincide just with the use of intuitionistic logic. Later, we will overview various set-theoretic proposals for constructive mathematics - particularly the Minimalist Foundation, researched by the Logic Group at the University of Padova.

2.1 Intuitionistic logic

The *principle of excluded middle* goes back to Aristotle (4th century BC) and states that either a proposition is true or its negation is true. In symbols

$$\varphi \vee \neg\varphi$$

In the mathematical practice, the excluded middle allows to bifurcate a proof in two paths *without necessarily knowing* which is the true one. Consider for example the classical proof of Proposition 1, where we used the fact that either $\sqrt{2}^{\sqrt{2}} \in \mathbb{Q}$ or $\sqrt{2}^{\sqrt{2}} \notin \mathbb{Q}$, which was indeed the source of its non-constructivity.

Intuitionistic logic is the logic that does not assume the principle of excluded middle among its axioms, so that a simple equation holds

$$\text{Classical logic} = \text{Intuitionistic logic} + \text{Excluded middle}$$

or, equivalently,

$$\text{Intuitionistic logic} = \text{Classical logic} - \text{Excluded middle}$$

We will see that intuitionistic logic allows more nuanced distinctions in reasoning, and in this sense we hope it could be of interest also for a classical mathematician.

Remark 1 A common pitfall when first reading about constructive mathematics is thinking that the excluded middle *is false* in intuitionistic logic. This is far from being true. To see it, consider for example the following trivial statement

$$0 = 1 \vee 0 \neq 1$$

It is of course provable in constructive mathematics, because we can explicitly say which of the two sides we are going to prove (the right hand side in this particular case).

What intuitionistic logic forces us to do is to prove the excluded middle each time we need it, because we can never give it for granted.

Simply removing one axiom from classical logic has an enormous impact on the flavour of proofs we can write down. We analyze in detail one of the major consequences of working without the excluded middle, namely the impossibility of performing proofs by contradictions. First, we need to introduce another reasoning principle.

The law of *double negation elimination* asserts that if the negation of a proposition is contradictory, then the proposition itself is true. In symbols

$$\neg\neg\varphi \Rightarrow \varphi$$

In intuitionistic logic, the double negation elimination does not hold. In fact, intuitionistic logic can prove that it is equivalent to the excluded middle.

$$(\neg\neg\varphi \Rightarrow \varphi) \Leftrightarrow (\varphi \vee \neg\varphi)$$

Double negation elimination is always (tacitly) used in proofs by contradiction. In order to see this, we need to carefully distinguish between two proof techniques, which are often confused: the *proof of a negation*, which is constructively sound, and the *proof by contradiction*, which is instead classical.

Proof of a negation. Suppose you want to prove a negated statement $\neg\varphi$. Then you assume φ holds, you derive a contradiction from this fact, and you conclude $\neg\varphi$. This is a perfectly valid constructive proof technique according to the BHK interpretation, as reviewed also in Example 5.

Proof by contradiction. Suppose you want to prove an arbitrary statement φ. Then you assume $\neg\varphi$ holds, you derive a contradiction from this fact and, using the proof of negation, you conclude $\neg\neg\varphi$. Then you use the classical law of double negation elimination and obtain φ. This is the kind of classical proof used in Example 2.

2.2 Constructive set-theories

We mentioned that classical mathematics enjoys a standard foundational theory, that is Zermelo-Fraenkel axiomatic set theory and that the situation in constructive mathematics is very different. Indeed, in the literature of constructive mathematics there are many foundational theories with different scopes, philosophies, and techniques, and no one of them has already reached the privileged status of "standard". We list some of the most famous:

- Martin-Löf's type theory [6]
- The constructive version of Zermelo-Fraenkel set theory [1]
- The internal language of a Topos [4]
- The Calculus of Constructions [2]
- Homotopy Type Theory [7]

What is worst is that most of the time, the above foundations are mutually incompatible, in the sense that principles or axioms used by one theory cannot be exported in another theory without generating contradictions. This is reflected by the fact that if you want to prove a theorem in "constructive mathematics", it is not enough to pick your favourite foundational system in the list above and prove it in there, because no one guarantees you that the same proof can be adapted to work also for the other systems.

To rectify this situation Maria Emilia Maietti and Giovanni Sambin designed the Minimalist Foundation [5], a foundational system to be regarded as a common core between all the most relevant foundations for constructive mathematics. The main property of the Minimalist Foundation is its *compatibility*, which can be intuitively expressed by saying that theorems and proofs written in the Minimalist Foundation can be exported soundly in any other foundations. My research focuses in particular in investigating the metamathematical properties of the Minimalist Foundation, such as the one discussed in the next section.

3 Double-Negation Interpretation

Recall that in Example 4 we reformulated a constructively unprovable statement into a classically equivalent one, and then showed it was constructively provable. In this section we will present a simple way of mechanically extending this process to an arbitrary formula. This translation is called *double-negation* and sends each formula φ into a formula $\varphi^{\mathcal{N}}$ such that:

(a) $\varphi^{\mathcal{N}}$ is classically equivalent to φ;

(b) φ is classically provable if and only if $\varphi^{\mathcal{N}}$ is intuitionistically provable.

The idea of the translation stems from the following observation. In the previous section we saw that intuitionistic logic does not assume neither the double negation elimination nor the excluded middle. However, intuitionistically, we can *refute* the negation of the excluded middle, that is we can prove

$$\neg\neg(\varphi \vee \neg\varphi)$$

which, by double negation elimination, is of course classically equivalent to the excluded middle. The idea is then to insert a double negation $\neg\neg$ in front of any *disjunction*, and any *existential quantifier* appearing in φ. Formally, restricting to the case of Peano Arithmetic, the translation is defined recursively on the formula complexity in the following way.

$$\begin{aligned}(a = b)^{\mathcal{N}} &:= a = b\\ (\varphi \wedge \psi)^{\mathcal{N}} &:= \varphi^{\mathcal{N}} \wedge \psi^{\mathcal{N}}\\ (\varphi \Rightarrow \psi)^{\mathcal{N}} &:= \varphi^{\mathcal{N}} \Rightarrow \psi^{\mathcal{N}}\\ (\neg\varphi)^{\mathcal{N}} &:= \neg\varphi^{\mathcal{N}}\end{aligned}$$

$$
\begin{aligned}
(\varphi \vee \psi)^{\mathcal{N}} &:= \neg\neg(\varphi^{\mathcal{N}} \vee \psi^{\mathcal{N}}) \\
((\exists x \in A)\varphi)^{\mathcal{N}} &:= \neg\neg(\exists x \in A)\varphi^{\mathcal{N}} \\
((\forall x \in A)\varphi)^{\mathcal{N}} &:= (\forall x \in A)\varphi^{\mathcal{N}}
\end{aligned}
$$

Example 6 The proposition

$$\varphi \equiv \forall x \in \mathbb{N}\,(x = n \vee \exists y \in \mathbb{N}\,.\,x = y + 1)$$

is translated to

$$\varphi^{\mathcal{N}} \equiv \forall x \in \mathbb{N}\,.\,\neg\neg(x = n \vee \neg\neg\exists y \in \mathbb{N}\,.\,x = y + 1)$$

In particular, the double-negation interpretation has been succesfully exploited by Gödel to show the equiconsistency of Peano Arithmetic with its constructive counterpart, the so-called *Heyting Arithmetic*. In a sense, this result shows that, from a classical perspective, every theorem of arithmetic can be proved constructively, using an equivalent formulation of its statement.

3.1 Double negation translation for set theory

The double-negation interpretation in the case of Arithmetics (where the only set is $\mathbb{N}$), only accounts for the translation of logical connectives. Our work extended the translation to a whole set theory, namely that of the Minimalist Foundation. In order to take care of this additional complexity, we must perform the translation also inside sets, when these involve some propositions. Consider the following example.

Example 7 The set defined by comprehension

$$A = \{f \in \mathbb{N} \to \mathbb{N} \,|\, \exists x \in \mathbb{N}\,.\,f(x) = 0\}$$

is translated to

$$A^{\mathcal{N}} = \{f \in \mathbb{N} \to \mathbb{N} \,|\, \neg\neg\exists x \in \mathbb{N}\,.\,f(x) = 0\}$$

Our fundamental result, which extend the one by Gödel on Peano Arithmetic, is then the following.

Theorem 1 *If p is a classical proofs of a theorem φ in the Minimalist Foundation, then there exists a constructive proof $p^{\mathcal{N}}$ of the (classically equivalent) theorem $\varphi^{\mathcal{N}}$.*

References

[1] P. Aczel and M. Rathjen, *Notes on constructive set theory*. Mittag-Leffler Technical Report No. 40 (2001).

[2] T. Coquand, *Metamathematical investigations of a calculus of constructions*. Tech. rep. RR-1088. INRIA (Sept. 1989). URL: https://inria.hal.science/inria-00075471.

[3] S.C. Kleene, *On the interpretation of intuitionistic number theory*. Journal of Symbolic Logic 10/4 (1945), pp. 109–124. DOI: 10.2307/2269016.

[4] Maria Emilia Maietti, *Modular correspondence between dependent type theories and categories including pretopoi and topoi*. Math. Structures Comput. Sci. 15/6 (2005), pp. 1089–1149. ISSN: 0960-1295,1469-8072. DOI: 10.1017/S0960129505004962. URL: https://doi.org/10.1017/S0960129505004962.

[5] Maria Emilia Maietti and Giovanni Sambin, *Toward a minimalist foundation for constructive mathematics*. In: From sets and types to topology and analysis. Vol. 48. Oxford Logic Guides. Oxford Univ. Press, Oxford (2005), pp. 91–114. DOI: 10.1093/acprof:oso/9780198566519.003.0006. URL: https://doi.org/10.1093/acprof:oso/9780198566519.003.0006.

[6] P. Martin-Löf, "Intuitionistic Type Theory". Notes by G. Sambin of a series of lectures given in Padua, June 1980. Bibliopolis, Naples, 1984.

[7] The Univalent Foundations Program, "Homotopy type theory - Univalent foundations of mathematics". The Univalent Foundations Program, Princeton, NJ; Institute for Advanced Study (IAS), Princeton, NJ, 2013, pp. xiv+589.

Digital twins: a general overview and the application to bread baking

Laura Rinaldi (*)

Abstract. A digital twin is composed of two existing systems: the tangible system of physical reality and its virtual and numerical replica which is enabled by real data and underlying models through the use of digital technologies. The presence of digital twins is motivated by the necessity of obtaining some information about the real system questioning the virtual one by a non-intrusive manner. Such technology helps us to monitor the real system, to carry out maintenance tasks or optimize some process. Here I will present an industrial application which consists in the building of an embedded digital twin of the bread baking process to the end of monitoring the energy consumption to avoid waste.

1 General overview

Industry 4.0 is the term used to describe the fourth industrial revolution, which involves the integration of digital technologies such as IoT, cloud computing, analytics, AI, and machine learning into the manufacturing sector. A main target of Industry 4.0 is to increase productivity and efficiency across the value chain and enable the customization and optimization of products and services. One of the main concept of the fourth industrial revolution is the **digital twin** (DT) that helps to build a bridge between the physical and the digital world and becomes a fundamental component to make decisions, check the status, modify the behavior and perform predictions.

A DT is classically defined as the virtual replica of a real-world product, system, being, communities, even cities that are continuously updated with data from its physical counterpart [5]. The first known definition, that follows, was presented by NASA for reflecting the life of an air vehicle. The actual term digital twin was coined by Grieves [3] and Tuegel [7] in 2011 and further developed in Grieves' consecutive works [4].

Definition 1 The **digital twin** is an integrated multi-physics, multi-scale, probabilistic simulation of a complex product and uses the best available physical models, sensor updates, etc., to mirror the life of its corresponding twin.

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: lrinaldi@math.unipd.it . Seminar held on 24 January 2024.

Since DT is an inter-disciplinary subject there is a no-unique definition, because it depends on the scientific area and on the purpose. In fact, digital twin integrates all data models and other information of the physical object generated along its life cycle for a dedicated goal.

Definition 2 The **digital twin** is a digital representation of a physical item or assembly using integrated simulations and service data, holding information from multiple sources across the product life cycle.

Nevertheless all the definitions are representative and useful to develop a general idea. DT can be applied in different situations and the aim is to describe all the product life cycle phases, from its design and prototyping to its disposal. The previous definitions highlight that the main objective is reflecting the life cycle of any element, product, or system that works as its physical twin. Grieves and Vickers [4] described the difference between the following concepts:

- Digital Twin Prototype: describes the prototypical physical artifact. It contains the informational sets necessary to describe and produce a physical version that duplicates or twins the virtual version.
- Digital Twin Instance: describes a specific corresponding physical product that an individual digital twin remains linked to throughout the life of that physical product.
- Digital Twin Environment: a multi-domain physics application space for operating on digital twins.

Moreover, a digital twin is made-up of two existing systems:

(a) The tangible system of real physics;

(b) Its virtual replica which is enabled by real data and underlying models through the use of digital technologies.

The physical object in the physical world and the digital object in the digital world have to be connected between each other. The connection is given by data from the real system to virtual one and information that you can achieve, with indirect measurements, from the digital replica. Especially the physical world is composed of two main elements:

(a) Devices: the physical twins from which DTs are intended to be created;

(b) Sensors: elements physically connected to devices from which one could get data and information.

The digital world is instead composed of two main elements:

(a) The virtual environmental platform to construct a 3-dimensional digital model;

(b) DT which mirrors the reality and allows multiple operations.

Finally the connections between the two depend on the choice of each author and on the scope to which the DT is created for.

Many times in real-world application we often deal with limited knowledge of a system, due to the practical inaccessibility of the physical domain, of some variables or parameters. Such data inaccessibility could be due to a costly or impractical physical measurements instruments, or because the measurement procedure is destructive with respect the real system, therefore the need of having virtual alternatives. Thus the presence of DTs is motivated by the necessity of obtaining some information about the real system by questioning the virtual one with a non-intrusive manner.

A mathematical model give us a whole and detailed perspective of the system to monitor the reality through soft sensors and indirect measures by estimating the quantity of our interest. Such technology helps us to control the real system, to carry out maintenance tasks or to optimize some process but also, to avoid some failure or even to preform predictions. The objective is to have a digital representation suited to the purpose in terms of level of detail, completeness, accuracy and execution speed [11].

It is possible to divided the application of DTs in the three main types [6]:

(a) Plain gadget models which includes two sets of data: data-information obtained by sensors and the set of expectation values that want to be obtained by the gadget;

(b) Embedded DTs, where the interaction among the real and digital world happens in a bidirectional way;

(c) Networked twins, where different embedded DTs are connected between each other and communicate.

In the following sections we will work with **embedded digital twin** [11] i.e. the virtual representation of physical systems that run in embedded systems which have sensors and actuators to interact with the real world.

In general, sensors are technical devices that monitor their environment and continuously produce signals at a regular frequency. A physical sensor is a sensor that reacts to a physical stimulus and transmits a resulting impulse – typically through electrical signals that can be captured and stored in digital form. In contrast to physical sensors, a so-called virtual sensor is a pure software sensor which autonomously produces signals by combining and aggregating signals that it receives (synchronously or asynchronously) from physical or other virtual sensors [8]. Virtual sensors are useful to enrich available information about physical variables and parameters that cannot be provided by direct physical measurements.

Virtual sensing proves to be highly relevant when on-site measurement of the variable of interest is not feasible due to non-accessible locations, cost or the fact that introducing sensors would distort the system under test.

Definition 3 (Soft-sensor) Soft sensors are algorithms that process the available data to estimate these lacking measurements.

In the perspective of using these algorithms within the DTs one could think to take advantage of the mathematical description of the model that is the foundation of a DT, thus:

Definition 4 (Physics-aware soft-sensor [1]) Physics-aware soft-sensors are algorithms that performs an indirect measurement by exploiting a **physical-mathematical model** plus a possible **data-driven extension**, used within an estimation algorithm.

Following the analysis in [6] nowadays the DTs are involved in may fields, in particular the aeronautics and space area (69%) used to monitor the operation of a system and the reliability of the created model, in manufacturing area (19%) used to monitor the life cycle of a system and to optimize its design, and in the informatics area (4%) for IoT life cycle management, thus to create control architectures based on this technology. Using and improving today the DTs could have a great impact since this technology helps us to save resources, money and processing time.

Even if there is not a generic way to build an embedded digital twin we select and identify these main tasks to perform:

1 Construct a **numerical model** which simulates the physical process and governed by physical laws;

2 Define **inverse problems**, based on the previous numerical model, to set parameters;

3 Define an **algorithm to capture the information** which you are interested in.

The inverse problem allows to continuously calibrate the parameters which ensures that the digital twin is always updated with respect to its real counterpart. The simulations obtained by the numerical model could be computationally expensive or take a lot of time to give a feedback to the user. To overcame such possible issue it is necessary a surrogate model of reduced order which is based on the numerical one and that can bring the simulations to run online with the real process in embedded system. Producing reduced order models, reduces the complexity and makes affordable the execution performance.

2 Classical approximation strategies

For definiteness, we will employ a common geometrical setting throughout the whole section. We consider a general domain Ω in $\mathbb{R}^3$ and we identify its boundary as $\partial\Omega$, moreover, we introduce a final time $\tau > 0$ and we consider evolution of equations on the time interval $[0, \tau]$ written in the following way

$$\frac{\partial u}{\partial t} - \Delta u = f \quad \text{in } \Omega, \tag{1}$$

where $f = f(x)$ is a given function and the symbol Δ denotes the Laplacian operator. To approximate the weak formulation of the previous diffusion problem for u using the finite

element method FEM we follow the well known theory [10], thus for each $t > 0$ we seek $u(t) \in V$such that

$$\int_\Omega \frac{\partial u(t)}{\partial t}\ v d\Omega + a(u(t), v) = \int_\Omega f(t)\ v d\Omega \quad \forall v \in V \tag{2}$$

where V being an appropriate Hilbert space, subspace of $H^1(\Omega)$, $u(0) = u_0$ and $a(\,\cdot\,,\,\cdot\,)$ is the bilinear form associated to the elliptic operator. A sufficient condition for the existence and uniqueness of the solution to problem (2) is that the following hypotheses hold: the bilinear form $a(\cdot, \cdot)$ is continuous and weakly coercive, that is

$$\exists \lambda \geq 0, \exists \alpha > 0 : a(v, v) + \lambda ||v||^2_{L^2(\Omega)} \geq \alpha ||v||^2_V \quad \forall v \in V, \tag{3}$$

yielding for $\lambda = 0$ the standard definition of coercivity. Moreover, we require $u_0 \in L^2(\Omega)$ and $f \in L^2(Q)$. We now consider the Galerkin approximation of problem (2): for each $t > 0$, find $u_h(t) \in V_h$ such that

$$\int_\Omega \frac{\partial u_h(t)}{\partial t}\ v_h d\Omega + a(u_h(t), v_h) = \int_\Omega f(t)\ v_h d\Omega \quad \forall v_h \in V_h \tag{4}$$

with $u_h(0) = u_{0_h}$, where $V_h \subset V$ is a suitable space of finite dimension and u_{0_h} is a convenient approximation of u_0 in the space V_h. Such problem is called semi-discretization of (2), as the temporal variable has not yet been discretized. To provide an algebraic interpretation of the discretization we introduce a basis ϕ_j for V_h and we observe that it suffices that the dicretization is verified for the basis functions in order to be satisfied by all the functions of the subspace. Moreover, since for each $t > 0$ the solution to the Galerkin problem belongs to the subspace as well, we will have

$$u_h(x, t) = \sum_{j=1}^{N_h} u_j(t)\phi_j(x),$$

where the coefficients $u_j(t)$ represent the unknowns of problem. Denoting by $\dot{u_j}(t)$ the derivatives of the function $u_j(t)$ with respect to time, thus

$$\int_\Omega \sum_{j=1}^{N_h} \dot{u_j}(t)\phi_j(x)\ \phi_i(x) d\Omega + a(\sum_{j=1}^{N_h} u_j(t)\phi_j(x)\, \phi_i(x)) = \int_\Omega f(t)\ \phi_i(x) d\Omega\ \forall i = 1, 2, \ldots, N_h, \tag{5}$$

that is

$$\sum_{j=1}^{N_h} \dot{u_j}(t) \int_\Omega \phi_j(x)\ \phi_i(x) d\Omega + \sum_{j=1}^{N_h} u_j(t) a(\phi_j(x)\, \phi_i(x)) = \int_\Omega f(t)\ \phi_i(x) d\Omega\ \forall i = 1, 2, \ldots, N_h, \tag{6}$$

that in matrix form reads as

$$M\dot{\mathbf{u}}(t) + A\mathbf{u}(t) = f(t), \tag{7}$$

where $\mathbf{u}(t) = (u_1(t), u_2(t), ..., u_{N_h}(t))^T$ is the vector of unknowns.

Let take the real positive parameter $\Delta t = t_{k+1} - t_k, k = 0, 1, \ldots,$ denotes the discretization step, while the superscript k indicates that the quantity under consideration refers to the time t_k. Therefore in a general setting as the previous in (7) the forward Euler (or explicit Euler) method works as

$$M\frac{\mathbf{u}^{k+1} - \mathbf{u}^k}{\Delta t} + A\mathbf{u}^k = f^k, \tag{8}$$

the backward Euler (or implicit Euler) method works as

$$M\frac{\mathbf{u}^{k+1} - \mathbf{u}^k}{\Delta t} + A\mathbf{u}^{k+1} = f^{k+1}, \tag{9}$$

instead the semi-implicit Euler works as

$$M\frac{\mathbf{u}^{k+1} - \mathbf{u}^k}{\Delta t} + A\mathbf{u}^{k+1} = f^k. \tag{10}$$

3 Industrial application

For definiteness, we will employ a common geometrical setting throughout the whole section. We consider two domains Ω_b and Ω_t in $\mathbb{R}^3$ which represent the region of the bread dough and of the tray respectively. With these provisions, we can identify their two boundaries as $\partial\Omega_\mathrm{b}$ and $\partial\Omega_\mathrm{t}$, with the following properties that $\partial\Omega_\mathrm{b} = \Gamma_1 \cup \Gamma_2$, $\Gamma_1 \cap \Gamma_2 = \emptyset$ and $\partial\Omega_\mathrm{t} = \Gamma_{\mathrm{t}_1} \cup \Gamma_{\mathrm{t}_2}$, $\Gamma_{\mathrm{t}_1} \cap \Gamma_{\mathrm{t}_2} = \emptyset$ where Γ_2 and Γ_{t_2} are the regions of boundary in which the bread and the tray are in contact instead Γ_1 and Γ_{t_1} are the regions of boundary in which they are not i.e. the regions where the dough and the tray are in contact with the air contained in the oven respectively. The importance of defining boundaries explicit lies in the fact that we have to manage heat exchanged, e.g. by conduction, between the elements involved during baking or the energy lost by bread for evaporation and boil phenomena. Moreover, we introduce a final time $\tau > 0$ and we set differential equations on the time interval $[0, \tau]$.

This industrial application consists in constructing the digital twin of the baking process to end of monitoring the energy consumption. To do so we have to define a numerical model which describes the physical system governed by differential equations, and to construct two algorithms: one for parameters identification and the other for energy estimation. The subject of the study is a bread dough made of a certain quantity of water, flour and yeasts, contained in a steel baking tray which, in the baking process, is put in a hot oven. In the following the differential equation considered for the numerical model.

- The dough is treated as a hyperelastic material so, under the action of its weight, it has an elastic behavior and it is subject to a deformation then, thanks to the yeasts presence, there will be a positive volume expansion.

  $$E(u) = \int_\Omega \frac{\mu}{2}(\mathrm{tr}\mathbf{C} - 3\sqrt[3]{J_{ref}^2}) + \frac{\lambda}{2}(\log(J) - \log(J_{ref}))^2 dX - \int_\Omega B \cdot u dX \tag{11}$$

  Thus the elasticity equation is $\delta E(u) = 0$.

- During the cooking process the temperature of the dough changes following the heat equation, where the boundary conditions describe the energy interactions between the paste with the baking tray or the oven, and considering the heat lost, by the dough, for the evaporation or the boil phenomena.

$$\begin{aligned}\int_\Omega \rho\ c_p\ J\frac{\partial\theta}{\partial t}\tilde{\theta}dX + \int_\Omega (k\ J\ \mathbf{C}^{-1}\nabla_X\theta)\cdot(\nabla_X\tilde{\theta})dX = \\ = W_{o\longrightarrow d} + W_{bt\longrightarrow d} + W_{loss}\end{aligned} \tag{12}$$

- The temperature of the steel baking tray also evolves when we put into the oven.

$$\begin{aligned}\int_{\Omega_{bt}} \rho_{bt}\ c_{bt}\frac{\partial\theta_{bt}}{\partial t}\tilde{\theta}_{bt}dX + \int_{\Omega_{bt}} (k_{bt}\nabla_X\theta_{bt})\cdot(\nabla_X\tilde{\theta}_{bt})dX = \\ = W_{o\longrightarrow bt} + W_{d\longrightarrow bt}\end{aligned} \tag{13}$$

- Due to the evaporation and the boiling phenomena the water quantity, contained in the paste, decreases and has a diffusive evolution.

$$\int_\Omega J\ \frac{\partial\rho_m}{\partial t}\ \tilde{\rho}_m dX + \int_\Omega (\beta\ J\ \mathbf{C}^{-1}\nabla_X\rho_m)\cdot(\nabla_X\tilde{\rho}_m)dX = M_{loss} \tag{14}$$

- The concentration of yeast, uniformly distributed in the paste, evolves in time, according to temperature reached by the loaf.

$$\begin{cases}\frac{d}{dt}Y = K_y(\theta)Y \quad \Longrightarrow \quad Y(t) = \exp(\ K_y(\theta)\ t)Y_0 \\ Y(0) = Y_0\end{cases} \tag{15}$$

where K_{y} is the growth rate function modeled as a cubic with two zeros in 0°C and in the yeast death temperature θ_{y}

$$K_{\mathrm{y}}(\theta) = y_{cost}\frac{27\ \theta^2(\theta_{\mathrm{y}} - \theta)}{4\ \theta_{\mathrm{y}}^3}.$$

- The metabolism of the yeasts implies a CO_2 production and diffusion in the paste.

$$\int_\Omega J\ \frac{\partial D}{\partial t}\ \tilde{D}dX + \int_\Omega (\beta_{co2}\ J\ \mathbf{C}^{-1}\nabla_X D)\cdot(\nabla_X\tilde{D})dX = \int_\Omega f(\theta)\ Y\ \tilde{D}dX \tag{16}$$

where the CO_2 production rate $f(\theta)$ is modeled as follows

$$f(\theta) = \max\Big\{ - D_{co2}\frac{(\theta-20)(\theta-60)}{400}, 0\Big\}, \tag{17}$$

with D_{co2} the concentration of CO_2 with respect to the concentration of yeasts per second.

- The rate of CO_2 involves a volume expansion because of the proving of the bread under specific conditions on temperature and moisture. We approximate the volume such that it is directly proportional to the temperature with an empirical coefficient of linearity which depends on the the number of moles. From the ideal gasses law we compute the volume of the CO_2 gas bubbles considering the number of moles contained in a gram of CO_2 that is

$$PV = n_{mol} R\theta, \qquad V = \frac{n_{mol} R\theta}{P}. \tag{18}$$

  The elastic stiffness are update according to the CO_2 quantity, so the elasticity equation is rerun to obtain the volume expansion which simulate the leavening.

The partial differential equations that describe the heat exchanges and the evolution of moisture, yeast and carbon dioxide are coupled with the quasi-static evolution of the growing elastic dough.

In order to approach the numerical model we construct the two geometries, related with the two domains of integration (the bread dough and the baking tray) and define the meshes.

Space of functions	**Meshes**	**Functions approximated**
V_h is a Vector Function Space of P1	bread	Placement u.
W_h is a Function Space of P2	bread	Temperature of the dough θ_{b}, density ρ, moisture ρ_m, yeasts rate Y and CO_2 concentration D.
U_h is a Function Space of P1	tray	Temperature of the baking tray θ_{t}.

The spaces of finite elements are defined to the end of approximating functions with polynomials therefore to obtain the spatial discretization of the model instead, for the time discretization, a semi-implicit Euler method is used to deal with the coupling among the different equations.

The numerical model is implemented using the Python library FEniCSx and the results of the simulations are graphically analyzed with the software ParaView.

The results of the simulation provide a starting point for setting up suitable parameter-identification procedures on which we are focusing our attention. In particular the work presented in [2] is preparatory to estimate the CO_2 production during the leavening process.

In our work we presented a general observation on how to replace changes of material properties in limited regions within a domain with fictitious forcing terms in initial- and boundary-value problems associated with wave propagation and diffusion.

Then, by considering a paradigmatic heat conduction problem on a domain with a cavity, we proved that the presence of the void can be replaced by a fictitious heat source with support contained within the cavity.

Introducing the constant mass density $\rho > 0$, specific heat $c > 0$, and thermal conductivity $k > 0$, we can write the heat conduction problem for the temperature field $\theta : \Omega_D \times [0, \tau] \to \mathbb{R}$ as

$$\begin{cases} \rho c \partial_t \theta = k \Delta \theta & \text{on } \Omega_D \times [0, \tau], \\ k \nabla \theta \cdot \boldsymbol{n} = g & \text{on } \partial\Omega_B \times [0, \tau], \\ k \nabla \theta \cdot \boldsymbol{n} = 0 & \text{on } \partial\Omega_C \times [0, \tau], \\ \theta(0, \cdot) = \theta_0(\cdot) & \text{on } \Omega_D, \end{cases} \tag{19}$$

where $\boldsymbol{n}$ is the unit outer normal to $\partial\Omega_D$, g is a prescribed (space- and time-dependent) heat flux on the external boundary, and θ_0 is the initial temperature field. For the sake of simplicity we assume a vanishing heat flux across the internal boundary $\partial\Omega_C$.

Our goal is now to prove that there exists a second differential problem, stated on the filled domain Ω_B, the solution of which is a temperature field $\tilde{\theta} : \Omega_B \times [0, \tau] \to \mathbb{R}$ such that $\tilde{\theta} = \theta$ on $\Omega_D \times [0, \tau]$. This problem takes the form

$$\begin{cases} \rho c \partial_t \tilde{\theta} = k \Delta \tilde{\theta} + f & \text{on } \Omega_B \times [0, \tau], \\ k \nabla \tilde{\theta} \cdot \boldsymbol{n} = g & \text{on } \partial\Omega_B \times [0, \tau], \\ \tilde{\theta}(0, \cdot) = \tilde{\theta}_0(\cdot) & \text{on } \Omega_B, \end{cases} \tag{20}$$

where the initial condition $\tilde{\theta}_0$ coincides with θ_0 on Ω_D and $f : \Omega_B \times [0, \tau] \to \mathbb{R}$ is a fictitious heat source.

Theorem 5 *Given a solution θ of the differential problem (19) with smooth boundary data g and initial condition θ_0, there exists a fictitious source field f such that:*

(i) *the solution $\tilde{\theta}$ of the differential problem (20) coincides with θ on $\Omega_D \times [0, \tau]$;*

(ii) *$f = 0$ in the physical domain $\Omega_D \times [0, \tau]$.*

We illustrated this fact in a situation where the source term can be analytically recovered from the values of the temperature and heat flux at the boundary of the cavity.

Our result provided a strategy to map the nonlinear geometric inverse problem of void identification into a more manageable one, that involves the identification of forcing terms given the knowledge of external boundary data. To set the stage for a systematic study of the inverse problem, we presented algebraic reconstructions, based on a finite-element discretization of the domain, that give an approximation of the fictitious source from different sets of temperature measurements. We showed how the accuracy of the reconstruction is reflected on the void identification.

Simulation of Embedded DT demands new numerical models and algorithms with respect to the virtual prototype because of restricted resources of the embedded systems. The model complexity must be in general substantially smaller than the one of the virtual replica. A surrogate model has the following main characteristics:

- replicates the physical behavior, so mimics the behavior of the simulations as closely as possible,
- is computationally cheaper,
- needs few data as input.

The technical features of such surrogate model allows to run the DT in embedded system. The presence of a surrogate model is motivated by the fact that it has to run online with the real process which is a main target of digital twin for Industry 4.0.

The ongoing development of microprocessors and network technologies allows the connection in real time between real and digital systems, in fact the simultaneity is the key to retrieve or get the information on which you are interested in, through indirect measurements and to use it to perform online tasks on the real system.

Therefore we are working on the building of such surrogate model or reduced order model, based on the numerical one, by using machine learning technique. Especially we focus our attention on the procedure presented in the article [9] by developing Physical Informed Neural Networks (PINNs), which are neural networks that embed in their formulation the physical knowledge coming from the partial differential equations of the model and exploiting the weak formulation of these PDEs.

Conclusions

The challenges in the industrial practice today can be deal with using DTs. With the support of such techniques, the modern industry brings a wide range of online tasks useful to take decisions, to develop architectures, to predict behavior or detect failure. It can be expected that with the rising awareness regarding the economic benefits of adopting DT technology, it will act as the backbone in the Industry 4.0.

There is no turning back from such a trend!

References

[1] E. Chinellato, S. Pierobon, and F. Marcuzzi, *Physics-aware soft sensors for embedded digital twins.* Proceedings of 9th International Congress on Information and Communication Technology (ICICT 2024). Springer LNNS (2024), in press.

[2] G.G. Giusteri, F. Marcuzzi, and L. Rinaldi, *Replacing voids and localized parameter changes with fictitious forcing terms in boundary-value problems.* Results in Applied Mathematics, 20:100402 (2023).

[3] M. Grieves, "Virtually perfect: Driving innovative and lean products through product lifecycle management". Volume 11.

[4] M. Grieves and J. Vickers, *Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems.* Transdisciplinary perspectives on complex systems: New findings and approaches (2017), pp. 85–113.

[5] Y. Jiang, S. Yin, K. Li, H. Luo, and O. Kaynak, *Industrial applications of digital twins.* Philosophical Transactions of the Royal Society A, 379(2207):20200360 (2021).

[6] M.G. Juarez, V.J. Botti, and A.S. Giret, *Digital twins: Review and challenges.* Journal of Computing and Information Science in Engineering, 21(3):030802 (2021).

[7] P.A. Kobryn, E.J. Tuegel, and S.M. Branch, *Condition-based maintenance plus structural integrity (cbm+ si) & the airframe digital twin.* USAF Air Force Research Laboratory, 88ABW-201101428 (2011).

[8] D. Martin, N. Kühl, and G. Satzger, *Virtual sensors.* Business & Information Systems Engineering 63 (2021), 315–323.

[9] D. Patel, D. Ray, M.R. Abdelmalik, T.J. Hughes, and A.A. Oberai, *Variationally mimetic operator networks.* arXiv preprint arXiv:2209.12871 (2022).

[10] A. Quarteroni and S. Quarteroni. "Numerical models for differential problems" (volume 2). Springer, 2009

[11] H. Van der Auweraer and D. Hartmann, *The executable digital twin: merging the digital and the physics worlds.* arXiv preprint arXiv:2210.17402 (2022).

PDE's and Conservation Laws: from the basics to current research

Luca Talamini (*)

Abstract. In the first part of this short note we introduce the reader to the general theory of scalar conservation laws. After reviewing some classical theorems, we delve into some more recent results obtained in [AMT24].

1 Scalar conservation laws

A scalar conservation law in d dimensions is a partial differential equation for a function $y \mapsto u(y) \in \mathbb{R}$, where $y \in \Omega \subset \mathbb{R}^d$, of the form

$$\operatorname{div} g(u) = 0 \qquad \text{in } \mathscr{D}'_y. \tag{1.1}$$

Here $g : \mathbb{R} \to \mathbb{R}^d$ is a given possibly non-linear smooth function

$$g(u) = (g_1(u), \dots, g_d(u)).$$

Time dependent versions of (1.1) can be obtained by factorizing $\mathbb{R}^d = \mathbb{R}_t \times \mathbb{R}^n_x$, $d = n + 1$, and choosing $g_1(u) = u$:

$$\partial_t u(t,x) + \operatorname{div}_x f(u(t,x)) = 0 \qquad \text{in } \mathscr{D}'_{t,x}. \tag{1.2}$$

Here div_x stands for the divergence with respect to the space variables only. In the following we will mainly deal with the evolutionary version (1.2).

The *Cauchy problem* for a scalar conservation law arises when one couples equation (1.2) with an initial condition $u_0(x)$ on the hyperplane $\{t = 0\}$:

$$u(0,x) = u_0(x), \qquad \forall\, x \in \mathbb{R}^n, \qquad\qquad u_0 \in \mathbf{L}^\infty(\mathbb{R}^n). \tag{1.3}$$

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: luca.talamini@phd.unipd.it . Seminar held on 31 January 2024.

Method of characteristic and singularity formation

In presence of a solution $u(t,x)$ of class $\mathscr{C}^1$, defined on some strip $[0,T]\times\mathbb{R}^n$, the method of characteristics, that we now describe, can be applied. In general, this method allows one to reduce a PDE to the resolution of a family of ordinary differential equations. In the particular case of (1.2), a *characteristic* starting at $y_0\in\mathbb{R}^n$ is a function $t\mapsto x(t)$ solving the ordinary differential equation

$$\dot{x}(t) \;=\; f'\big(u(t,x(t))\big), \qquad\qquad x(0)=y_0. \tag{1.4}$$

Consider the map $t\mapsto u(t,x(t))$, which is of class $\mathscr{C}^1$, and differentiate with respect to time to obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}u(t,x(t)) = \partial_t u(t,x(t)) + \nabla_x\, u(t,x(t))\;\cdot\; f'(u(t,x(t)) = \partial_t u(t,x) + \mathrm{div}_x\, f(u(t,x)) = 0$$

This implies that the solution u is constant along characteristics: $u(t,x(t)) = u_0(y_0)$ for every $t\in[0,T]$. In particular, from (1.4), we deduce also that characteristics have constant speed. The same method also gives local existence in a short time interval $[0,\tau]\times\mathbb{R}^n$ starting from an initial datum $u_0\in\mathscr{C}^1$.

Example 1.1 (Formation of singularities) To see that in general a smooth solution does not exist for all times, consider the following basic example. Let $n=1$ and assume that there are $y_1<y_2$ such that $f'(u_0(y_1))>f'(u_0(y_2))$. Then there exists some $T>0$ with the property that, the characteristics $x_1(t), x_2(t)$ starting respectively at y_1, y_2, *meet* at time T: $x_1(T)=x_2(T)$ (see Figure 1). Then, on the one hand, we know that smooth solutions are constant along characteristics. On the other hand, if this was the case, the solution would have two values at the point $(T,x_1(T))=(T,x_2(T))$. This implies that a smooth solution does not exists for times $s>T$.

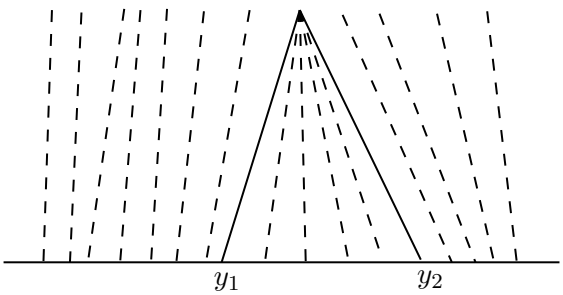


Figure 1: Formation of a singularity for a scalar conservation laws with $n=1$.

The weak formulation (1.2) makes sense also for irregular (not smooth) functions, for which the method of characteristics cannot be applied. However, in general, distributional solutions to the Cauchy problem (1.2), (1.3) are not unique, as we can see in the following example.

Example 1.2 Let $n = 1$ and consider Burgers' equation, namely the conservation law with flux $f = \frac{u^2}{2}$. Let the initial datum be

$$u_0(x) = \begin{cases} 0 & x < 0 \\ 1 & x > 0 \end{cases}$$

One easily checks that

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

is a solution. On the other hand, one can also solve the Cauchy problem using a discontinuous function. For example, using the divergence theorem, it is easy to check that the function

$$\widetilde{u}(t,x) = \begin{cases} 0, & \text{if } x < \frac{1}{2}t, \\ 1, & \text{if } \frac{1}{2}t < x \end{cases}$$

is again a distributional solution.

The following definition is a fundamental step towards a uniqueness result.

Definition 1.3 (Entropy-entropy flux) We say that a pair $(\eta, \mathcal{Q})$ of Lipschitz functions $\eta : \mathbb{R} \to \mathbb{R}$, $\mathcal{Q} : \mathbb{R} \to \mathbb{R}^n$ is an *entropy-entropy flux* pair if

$$\mathcal{Q}'(u) = f'(u) \cdot \eta'(u) \qquad \text{for a.e. } u \in \mathbb{R}.$$

Remark 1.4 Any Lipschitz $\eta : \mathbb{R} \to \mathbb{R}$ can serve as an entropy provided that we define the flux $\mathcal{Q}$ as

$$\mathcal{Q}(u) = \int^u f'(\omega) \cdot \eta'(\omega)\, \mathrm{d}u$$

Now we are ready to give the fundamental definition of *entropy solution* to a conservation law.

Definition 1.5 A function $u \in \mathscr{C}^0\big((0,+\infty), \mathbf{L}^1(\mathbb{R})\big)$ is an *entropy solution* of (1.2) if

$$\partial_t \eta(u) + \operatorname{div}_x \mathcal{Q}(u) \leq 0 \qquad \text{in } \mathscr{D}'_{t,x} \tag{1.5}$$

for all entropy-entropy flux pairs $(\eta, \mathcal{Q})$ such that η is *convex*.

Remark 1.6 If (1.5) holds, then it follows by integrating in x that the function

$$t \mapsto \int_{\mathbb{R}^n} \eta(u(t,x))\, \mathrm{d}x$$

is non-increasing.

Example 1.7 (Shock solution) A *shock* in dimension 1 ($n = 1$) connecting two states $u^- > u^+$ is a function of the form

$$u(t,x) = \begin{cases} u^-, & \text{if } x \le \lambda t, \\ u^+, & \text{if } x > \lambda t, \end{cases} \tag{1.6}$$

Using the divergence theorem, it is an explicit calculation to see that the function in (1.6) is a distributional solution to (1.2) if and only if the *speed* λ satisfies the *Rankine-Hugoniot* relation

$$\lambda = \frac{f(u^+) - f(u^-)}{u^+ - u^-}.$$

Moreover, using again the divergence theorem in equation (1.5) we find that u is also an entropy solution if and only if

$$\mathcal{Q}(u^+) - \mathcal{Q}(u^-) - \lambda\big(\eta(u+) - \eta(u-)\big) \le 0 \qquad \text{for all convex entropies } \eta. \tag{1.7}$$

By using in (1.7) the family of entropies $\eta(\omega) = [\omega - k]^-$ (here $[a]^- = \max\{-a, 0\}$), $k \in (u^+, u^-)$, one obtains that u is an entropy solution if and only if

$$\operatorname{sgn}[u^+ - k]^- \cdot \big(f(u^+) - f(k)\big) - \lambda\big[u^+ - k\big]^- \le 0 \qquad \forall\, k \in (u^+, u^-)$$

that rewrites

$$f(k) - f(u^+) - \lambda(k - u^+) \le 0 \qquad \forall\, k \in (u^+, u^-).$$

The geometrical meaning is that the graph of f is lying below the chord passing through $(u^-, f(u^-)), (u^+, f(u^+))$ (see Figure 4). Symmetric statements hold if $u^- < u^+$. Notice that the function $\widetilde{u}$ in Example 1.2 has a single shock that does *not* satisfy the entropy conditions above.

The goal of the next paragraph will be to sketch the proof of the following fundamental Theorem.

Theorem 1.8 *There exists a unique entropy solution to the Cauchy problem (1.2), (1.3).*

1.1 Existence and uniqueness of entropy solutions

We start this section by motivating the concept of entropy solutions via a classical observation. We consider the *vanishing viscosity approximations*, that are equations of the following form: if ε is positive, one adds a small *viscosity term* in the right hand side of (1.2)

$$\partial_t u^\varepsilon + \operatorname{div}_x f(u^\varepsilon) = \varepsilon\, \Delta u^\varepsilon \tag{1.8}$$

where Δ stands for the Laplacian operator with respect to the space variables $x \in \mathbb{R}^n$. If $\varepsilon \to 0$ then, at least formally, solutions to (1.8) are expected to converge to solutions of

(1.2). The catch is that (1.8) is a parabolic-type equation, and therefore its solutions exist and are smooth at $t > 0$, even if the initial datum u_0 is merely a bounded function. As the following classical Proposition shows, in the limit $\varepsilon \to 0$ entropies allow to detect solutions that come from the vanishing viscosity approximation.

Proposition 1.9 *Assume that $\{u^{\varepsilon_k}\}_{k\in\mathbb{N}}$ is a sequence of uniformly bounded solutions to (1.8) that converges almost everywhere in (t, x) to some function u. Then u is an entropy solution to (1.2).*

Proof. Let $(\eta, \mathcal{Q})$ be any entropy-entropy flux pair and multiply (1.8) by η'. By the chain rule, using the definition of entropy, we obtain

$$\partial_t \eta(u^\varepsilon) + \mathrm{div}_x \mathcal{Q}(u^\varepsilon) = \varepsilon\, \eta'(u^\varepsilon)\Delta u^\varepsilon = \varepsilon\, \Delta\eta(u^\varepsilon) - \varepsilon\eta''(u^\varepsilon)|\nabla u^\varepsilon|^2. \tag{1.9}$$

The first term in the right hand side satisfies, as $\varepsilon \to 0$,

$$\varepsilon\, \Delta\eta(u^\varepsilon) \to 0 \qquad \text{in } \mathscr{D}'_{t,x}.$$

The second term in general does not converge to zero in distributions, but it is negative. Therefore passing to the limit as $\varepsilon \to 0$ we obtain (1.5). □

Notice that Proposition 1.9 also provides a hint on how to prove the existence part of Theorem 1.8 with the following strategy:

(a) Show that the family of solutions to (1.8) $\{u^\varepsilon\}_{\varepsilon>0}$ is bounded and strongly pre-compact in $\mathbf{L}^1$.

(b) Using the compactness given by point (1) extract a subsequence $\{u^{\varepsilon_k}\}_{k\in\mathbb{N}}$ that converges pointwise for a.e. (t, x).

(c) Apply Proposition 1.9 to deduce that the limit is in fact an entropy solution to (1.2).

It turns out that this strategy can actually be implemented and proves the existence part of Theorem 1.8. For more details, see for example [D].

The uniqueness of entropy solutions is a Theorem that dates back to Kružhkov [Kru70].

Theorem 1.10 (Kružhkov) *Let u, v be two entropy solutions with initial data $u_0, v_0 \in \mathbf{L}^1(\mathbb{R}^n) \cap \mathbf{L}^\infty(\mathbb{R}^n)$. Then*

$$\int_{\mathbb{R}^n} |u(t,x) - v(t,x)|\, \mathrm{d}x \leq \int_{\mathbb{R}^n} |u_0(x) - v_0(x)|\, \mathrm{d}x \qquad \forall\, t > 0.$$

Proof. We just sketch the proof. For every k, the function

$$\eta_k(\omega) = |\omega - k| \qquad \omega \in \mathbb{R}$$

is a convex entropy, with entropy flux equal to

$$\mathcal{Q}_k(\omega) \doteq \mathrm{sgn}\,(\omega - k)\big(f(\omega) - f(k)\big).$$

An immediate observation is that if v is a constant solution, the Theorem follows in this case by taking $k = v$ and using Remark 1.6. Now if v is not constant, choosing $k = v(s, y)$ we find

$$\partial_t\, \eta_{v(s,y)}(u) + \partial_x\, \mathcal{Q}_{v(s,y)}(u) \le 0 \qquad \text{in } \mathscr{D}'_{t,x} \text{ for all } (s, y). \tag{1.10}$$

Reversing the roles of u and v, we find in the same way

$$\partial_s\, \eta_{u(t,x)}(v) + \partial_y\, \mathcal{Q}_{u(t,x)}(v) \le 0 \qquad \text{in } \mathscr{D}'_{s,y} \text{ for all } (t, x). \tag{1.11}$$

Then we sum (1.10) and (1.11) in the larger product space $X \doteq \big(\mathbb{R}^+_t \times \mathbb{R}^n_x\big) \times \big(\mathbb{R}^+_s \times \mathbb{R}^n_y\big)$ to obtain

$$\begin{aligned} &(\partial_t + \partial_s)|u(t,x) - v(s,y)| \\ &+ (\partial_x + \partial_y)\Big(\operatorname{sgn}\,(u(t,x) - v(s,y)) \cdot \big(f(u(t,x)) - f(v(s,y))\big)\Big) \le 0 \qquad \text{in } \mathscr{D}'(X). \end{aligned} \tag{1.12}$$

It then follows by "localizing" the distribution (1.12) on the diagonal of X (that is, the set $(t, x) = (s, y)$), that

$$\partial_t|u(t,x) - v(t,x)| + \partial_x\Big(\operatorname{sgn}\,(u(t,x) - v(t,x)) \cdot \big(f(u(t,x)) - f(v(t,x))\big)\Big) \le 0 \qquad \text{in } \mathscr{D}'_{t,x}.$$

and integrating in x this implies that the function $t \mapsto \|u(t,\cdot) - v(t,\cdot)\|_{\mathbf{L}^1}$ is non increasing in time. □

2 Evolution and structure of first order perturbations

Now that we have associated to each initial datum u_0 a unique solution u, that depends Lipschitz continuously on u_0 in $\mathbf{L}^1$ (by Theorem 1.10), a natural question is whether the map $u_0 \mapsto u$ enjoys some differentiability properties, in some weak sense. The aim of this section is to present some results related to this problem, contained in [AMT24]. From now on, we assume that $n = 1$ and that the flux f is a $\mathscr{C}^1$ function $f : \mathbb{R} \to \mathbb{R}$ that it is not affine when restricted to any nontrivial interval.

A general principle in evolutionary PDE's is that perturbations of solutions to a PDE solve a linearized version of the original equation. In our case, for the scalar conservation law (1.2) in space dimension 1, the Cauchy problem reads

$$\begin{aligned} &\partial_t u(t,x) + \partial_x f(u(t,x)) = 0, \qquad t > 0, \quad x \in \mathbb{R} \\ &u(0,x) = u_0(x). \end{aligned} \tag{2.1}$$

It is not restrictive to consider, from now on, solutions that are positive. We let $u \ge 0$ be the entropy solution to the Cauchy problem (2.1) with initial datum $u_0 \ge 0$ and consider $\{u^h\}_{h>0}$, a family of solutions to the same equation with perturbed initial data:

$$u^h(0,\cdot) = u^h_0 = u_0 + h \cdot v^h_0, \qquad \|v^h_0\|_{\mathbf{L}^1} \le 1, \qquad v^h_0 \ge 0, \qquad \lim_{h \to 0^+} v^h_0 = \rho_0. \tag{2.2}$$

The fact that $v_0^h \geq 0$ is just a simplifying assumption, that however can be dropped with no additional difficulties. Even assuming that v_0^h converge to a $\mathbf{L}^1$ function ρ_0, a first basic issue that we face is to identify a suitable functional space X in which we can consider limits of the form

$$\rho(t) \doteq \lim_{h\to 0} (u^h(t) - u(t))\, h^{-1} \in X.$$

Next, we would like to identify a suitable limiting evolution equation for the limiting perturbations $t \mapsto \rho(t)$, in the space X, augmented with the initial datum $\rho(0) = \rho_0$. The main problem is that, even if $\rho_0 \in \mathbf{L}^1$, it might well happen that the perturbation $\rho(t) \notin \mathbf{L}^1$ at later times. To understand why, we start with a basic example (borrowed from [BM95]).

Example 2.1 Consider Burgers equation

$$\partial_t\, u + \partial_x \left(\frac{u^2}{2}\right) = 0$$

and a family of perturbed initial conditions (see Figure 2)

$$u_0^h \doteq (1+h)x \cdot \mathbf{1}_{[0,1]}(x)$$

where $\mathbf{1}_A$ is the characteristic function of a set A, and let u^h be the solution corresponding to the initial datum u_0^h. Using Example 1.7 to calculate the speed of shocks in the solutions u^h, we can calculate explicitly that

$$u^h(t,x) = \frac{(1+h)x}{1+(1+h)t} \cdot \mathbf{1}_{[0,\sqrt{1+h(1+t)}]}(x).$$

Notice that at $t = 0$, the limit is an $\mathbf{L}^1$ function

$$\rho_0 \doteq \lim_{h\to 0} \frac{u_0^h - u_0}{h} = \lim_{h\to 0} v_0^h = x \cdot \chi_{[0,1]}(x) \in \mathbf{L}^1$$

and yet at time t, we have

$$\rho(t) = \lim_{h\to 0} \frac{u^h(t) - u(t)}{h} = \frac{x}{1+t^2} \cdot \mathscr{L}^1 \llcorner (0, \sqrt{1+t}) \;+\; \frac{t}{2(1+t)} \cdot \delta_{\sqrt{1+t}} \qquad \notin \mathbf{L}^1.$$

where δ_a stands for a Dirac mass at the point $a \in \mathbb{R}$. In other words, what at time $t = 0$ was an absolutely continuous perturbation, can concentrate its mass along the shock curve of u.

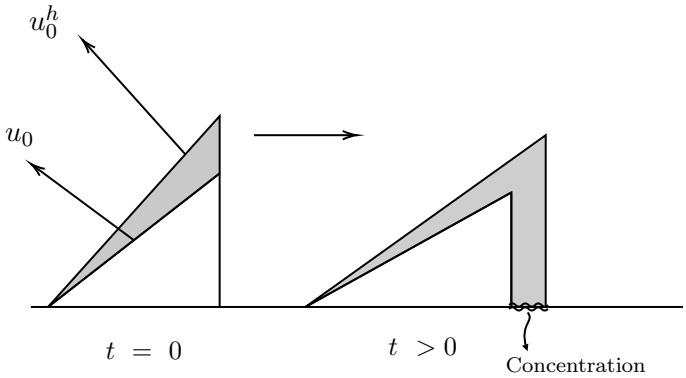


Figure 2: An initial perturbation in $\mathbf{L}^1$ can progressively concentrate on small sets.

Evolution of perturbations

This example suggests that we should take $X = \mathscr{M}^+(\mathbb{R})$, the space of positive finite measures on $\mathbb{R}$. Indeed, we have the following. Since for solutions of (2.1) a comparison principle holds, we have that if u^h is the entropy solution to the problem with initial datum $u_0 + hv_0^h$, then $u^h \geq u$. By the Theorem of Kružhkov 1.10,

$$\limsup_{h\to 0+} h^{-1}\|u^h(t,\cdot) - u(t,\cdot)\|_{\mathbf{L}^1} \leq \|v_0^h\|_{\mathbf{L}^1} \leq 1 \tag{2.3}$$

so that the family $\{h^{-1}(u^h - u)\}_h$ is relatively compact in the space of measures endowed with the weak topology. By Taylor expanding the flux function f around the reference solution $u(t,x)$

$$f(u^h(t,x)) = f(u(t,x)) + f'(u(t,x)) \cdot (u^h(t,x) - u(t,x)) + o(|u^h(t,x) - u(t,x)|)$$

we find that formally, any limiting measure ρ satisfies the *continuity equation*

$$\partial_t \rho + \partial_x(f'(u)\rho) = 0 \qquad \text{in } \mathscr{D}'_{t,x}. \tag{2.4}$$

Defining

$$\Delta(u_0) \doteq \left\{\rho \in \mathscr{M}(\mathbb{R}^+\times\mathbb{R}) \;\middle|\; \rho = \text{w–}\lim_{h\to 0} \frac{u^h - u}{h} \quad u^h(0,\cdot) \text{ as in (2.2), for some } \rho_0 \in \mathscr{M}^+(\mathbb{R})\right\}$$

we state a first Theorem:

Theorem 2.2 *Any $\rho \in \Delta(u_0)$ solves the Cauchy problem for the continuity equation*

$$\begin{cases} \partial_t \rho + \partial_x(\lambda\rho) = 0, & (t,x) \in \mathbb{R}^+ \times \mathbb{R} \\ \rho(0) = \rho_0 \end{cases}$$

where

$$\lambda(t,x) = \begin{cases} f'(u), & \text{if } u \text{ is continuous at } (t,x), \\ \frac{f(u^+) - f(u^-)}{u^+ - u^-}, & \text{if } u \text{ has a jump } u^- = u(t,x-), \quad u^+ = u(t,x+). \end{cases}$$

Notice that λ is just the speed of a characteristic passing at (t,x) if u is continuous, otherwise it is the speed of the shock with states u^-, u^+ (recall Example 1.7). Here λ, thanks to some regularity properties of u (see [BM17]), can be defined in every point (t,x), and therefore the product $\lambda\rho$ is well defined.

Structure of Perturbations

It is well know that entropic solutions to (2.1) satisfy a kinetic equation (see [LPT94])

$$\partial_t \chi + f'(v)\partial_x \chi = \partial_v \mu, \qquad \mu \in \mathscr{M}^+(\mathbb{R}^+ \times \mathbb{R} \times \mathbb{R})$$

where

$$\chi(t,x,v) = \mathbf{1}_{\text{hyp } u}(t,x,v), \qquad (t,x,v) \in \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}.$$

and $\text{hyp}\, u$ denotes the hypograph of the function u. The left hand side is just a linear transport equation, while the right hand side can be thought of as a Lagrange multiplier that acts so as to enforce the constraint that χ should be a characteristic function of an hypograph. By definition of χ, the solution u can be obtained by integrating χ in the "v" variable, that is $u = \int_v \chi$. Define

$$\nu_h \doteq \frac{1}{h}\big(\chi^h - \chi\big), \qquad \chi^h(t,x,v) = \mathbf{1}_{\text{hyp } u^h}(t,x,v). \tag{2.5}$$

Again by the $\mathbf{L}^1$ contractivity given by Theorem 1.10, the sequence ν_h is relatively compact in the space of measures endowed with the weak topology, because $|\nu_h| = h^{-1}\|u^h - u\|_{\mathbf{L}^1}$. Hence we define the set

$$\widetilde{\Delta}(u_0) = \Big\{ \nu \in \mathscr{M}(\mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}) \;\Big|\; \nu = \operatorname*{w\text{–}lim}_{h\to 0} \nu_h \text{ for some sequence } \nu_h \text{ as in (2.5)} \Big\}.$$

Since projections commute with weak-limits, for a subsequence such that both $\rho_h \to \rho$ and $\nu_h \to \nu$ (in such case we call (ρ,ν) a compatible pair), one has $\rho = p_\sharp \nu$, where p is the canonical projection on the space-time variables:

$$\begin{array}{ccc} \nu_h & \xrightarrow{h\to 0} & \nu \\ {\scriptstyle p_\sharp}\downarrow & & \downarrow{\scriptstyle p_\sharp} \\ \rho_h & \xrightarrow{h\to 0} & \rho \end{array}$$

In particular, $p_\sharp \widetilde{\Delta}(u_0) = \Delta(u_0)$ and one can recover ρ from ν but not viceversa: the measure ν yields additional information about perturbations. In the following we will address the natural question of understanding the structure of elements in $\widetilde{\Delta}(u_0)$.

Let (ρ,ν) be a compatible pair and let $(t,x) = y \mapsto a_y \in \mathscr{P}(\mathbb{R}_v)$ be the disintegration of the measure ν w.r.t the projection p (see for example [AFP00]):

$$\nu(y,v) = a_y(v) \otimes \rho(y) \in \mathscr{M}(\mathbb{R}^+ \times \mathbb{R}_x \times \mathbb{R}_v). \tag{2.6}$$

From the point of view of our original problem, $a_{t,x}$ is supposed to detect, with a first order precision, the asymptotic position of the graph of $u^h(t)$ with respect to the graph of $u(t)$, near points x in which u has a shock u^-, u^+ (see Figure 3).

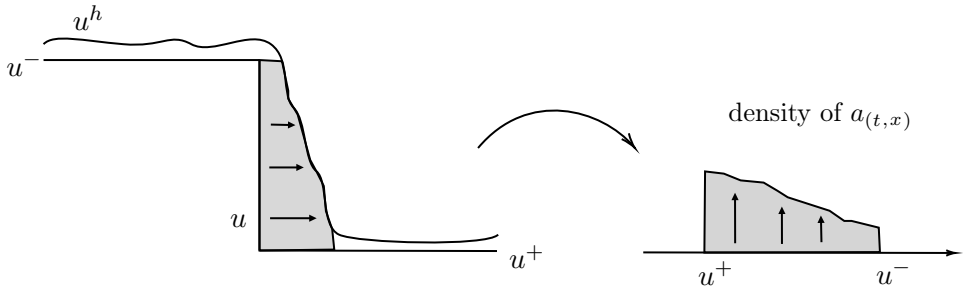


Figure 3: Geometrical meaning of the probability measure $a_{(t,x)}$.

The following Definition will be fundamental to understand the structure of the measures $a_y(v)$ (and hence of ν).

Definition 2.3 Let $u^- > u^+$ be connected by an entropy admissible shock with speed λ. We let $\mathscr{I}_1(u^-, u^+)$ be the set of maximal connected intervals of the set

$$J \doteq \Big\{ v \in (u^+, u^-) \;\Big|\; f(v) - f(u^+) - \lambda(v - u^+) < 0 \Big\}. \tag{2.7}$$

and we let $\mathscr{I}_2$ be the union of $\{u^-\}$, $\{u^+\}$ if $f'(u^-) = \lambda$ or $f'(u^+) = \lambda$, respectively. We set

$$\mathscr{I}(u^-, u^+) \doteq \mathscr{I}_1(u^-, u^+) \cup \mathscr{I}_2(u^-, u^+).$$

Finally, we let

$$\mathscr{K}(u^-, u^+) = (u^+, u^-) \setminus \mathscr{J}$$

If $u^- < u^+$, $\mathscr{I}(u^-, u^+)$ and $\mathscr{K}(u^-, u^+)$ are defined symmetrically.

Remark 2.4 The connected components $\mathscr{J}_1(u^-, u^+)$ represent the minimal entropy admissible shocks into which one can split the shock u^-, u^+.

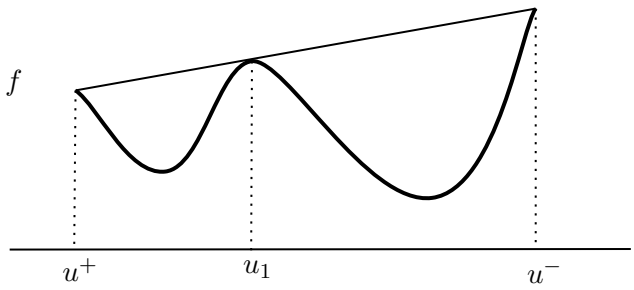


Figure 4: Example of the set $\mathscr{J}(u^-, u^+)$ for an admissible shock u^-, u^+. In this case $\mathscr{J}(u^-, u^+) = \{(u^+, u_1), (u_1, u^-)\}$.

Now we can state our second main result. With $I(u^-, u^+)$ we denote the open interval with endpoints u^- and u^+. We let $u^\pm(y)$ be the left and right limits of u at the point $y = (t, x)$: $u^\pm(t, x) = u(t, x\pm)$.

Theorem 2.5 *Any compatible pair (ρ, ν) disintegrates*

$$\nu = a_y(v) \otimes \rho(y) \in \mathscr{M}(\mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}), \qquad a_y \in \mathscr{P}\Big(\mathrm{clos}\big(I(u^-(y), u^+(y))\big)\Big)$$

where, for ρ-almost every y

(1) $(a_y) \llcorner I(u^-(y), u^+(y)) = \boldsymbol{g}_y \llcorner \mathscr{L}^1$, *where* $\boldsymbol{g}_y \in \mathbf{L}^1\big(I(u^-(y), u^+(y))\big)$ *is nonincreasing.*

(2) $D\boldsymbol{g}_y \in \mathscr{M}^-\big(I(u^-(y), u^+(y))\big)$ *is concentrated on the set* $\mathscr{K}(u^-(y), u^+(y))$.

Corollary 2.6 (Convex case) *Let f be strictly convex or concave. Any compatible pair (ρ, ν) disintegrates*

$$\nu = a_y(v) \otimes \rho(y) \in \mathscr{M}(\mathbb{R}^+ \times \mathbb{R} \times \mathbb{R})$$

where

$$a_y = \frac{1}{|u^- - u^+|}\, \mathscr{L}^1 \llcorner I(u^+(y), u^-(y)) \qquad \text{for } \rho\text{-almost every } y.$$

In particular, for every $\rho \in \Delta(u_0)$ there exists a unique lift $\nu \in \widetilde{\Delta}(u_0)$.

For cubic cases shocks cannot be split, but contact discontinuities (i.e. shocks in which $\lambda = f'(u^-)$ or $\lambda = f'(u^+)$) are present. Therefore specializing Theorem 2.5 in this case we obtain

Corollary 2.7 (Cubic case) *Let $f(u) = u^3$. Any compatible pair (ρ, ν) disintegrates*

$$\nu = a_y(v) \otimes \rho(y) \in \mathscr{M}(\mathbb{R}^+ \times \mathbb{R} \times \mathbb{R})$$

where for some measurable functions $y \mapsto \rho_1(y) \in \mathbb{R}$, $y \mapsto \rho_\pm(y) \in \mathbb{R}$, we have

$$a_y = \varrho_1(y) \cdot \mathscr{L}^1 \llcorner I(u^+(y), u^-(y)) + \varrho_+(y) \cdot \delta_{\{u^+\}} + \varrho_-(y) \cdot \delta_{\{u^-\}} \qquad \text{for } \rho\text{-almost every } y$$

and

$$f'(u^\pm(t,x)) \neq \lambda(t,x) \quad \Longrightarrow \quad \varrho_\pm(t,x) = 0 \qquad \text{for } \rho\text{-almost every } (t,x).$$

References

[AFP00] L. Ambrosio, N. Fusco, and D. Pallara, "Functions of Bounded Variation and Free Discontinuity Problems". Oxford Science Publications. Clarendon Press, 2000.

[AMT24] F. Ancona, E. Marconi and L. Talamini, "A differential structure for scalar conservation laws". In preparation.

[BM17] S. Bianchini and E. Marconi, *On the structure of L^∞ entropy solutions to scalar conservation laws in one-space dimension-entropy solutions to scalar conservation laws in one-space dimension.* Archive for Rational Mechanics and Analysis (Jun 2017).

[BM95] A. Bressan and A. Marson, *A variational calculus for discontinuous solutions of systems of conservation laws.* Communications in Partial Differential Equations 20/9-10 (1995), 1491–1552.

[D] C. Dafermos. "Hyperbolic Conservation Laws in Continuum Physics", Fourth edition.Springer-Verlag, Berlin, 2016

[LPT94] P.-L. Lions, B. Perthame, and E. Tadmor, *A kinetic formulation of multidimensional scalar conservation laws and related equations.* J. Amer. Math. Soc. 7/1 (1994), 169–191.

[Kru70] S. N. Kružkov, *First order quasilinear equations with several independent variables.* Mat. Sb. (N.S.) 81/123 (1970), 228–255.

A differential game model for sponsored content

Chiara Brambilla (*)

Abstract. Consider a communication platform that publishes two distinct types of advertising: traditional and native. Native advertising is a marketing technique that emulates the typical content of the platform where it is displayed. The strong similarity between native advertising and platform content enhances its effectiveness, as consumers may not identify the sponsored message. Nevertheless, when consumers become aware of the commercial nature of native advertising, they may feel misled and may loose trust in the platform. In our framework, a firm invests in both traditional and native advertising on a media channel known for its high-quality content. So, the media outlet has to consider the trade-off between generating revenue through ad placements and the impact of native advertising on its credibility. The problem is defined as a linear state differential game over an infinite-horizon. We search for an open-loop Stackelberg equilibrium and investigate the long-term sustainability of native advertising for the media.

1 Introduction

Native advertising is a form of communication that mimics the standard content of the platform where it is displayed [13]. This kind of advertising is common in digital media and includes ads that looks like search engine results, news, videos, [13] etc. The format of native advertising makes more challenging for consumers to distinguish the sponsored content [12]. As a result, native advertising is very effective [11]; however, discovering content as sponsored may cause negative reactions from consumers. If consumers feel misled, they may lose trust in the media's credibility [1, 2, 3]. The media outlet must search for a balance between the revenue generated by publishing native advertising and the subsequent credibility loss. Our study seeks to formalise this trade-off for a media platform using an approach derived from dynamic advertising models theory [7] as a part of a differential game [6,8]. Given the asymmetrical roles of the media and the firm, we formulate a differential game played á la Stackelberg, with the media as the Leader. We aim to identify an open-loop equilibrium in an infinite-horizon game.

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: brambill@math.unipd.it. Seminar held on 8 November 2023.

We consider a two-players differential game. A player is the firm that invests in advertising to maintain its brand value through a marketing campaign on the media outlet. Two types of advertising are available, the traditional and the native. If $G(t)$ denotes the brand value of the firm at time t, by Nerlove–Arrow advertising model, the evolution of this variable can be defined as

$$\dot{G}(t) = -\delta G(t) + \gamma_a a(t) + \gamma_n n(t)\,. \tag{1}$$

where $\delta > 0$ is the natural decay parameter, $a(t)$ denotes the standard advertising flow, and $n(t)$ denotes the native advertising flow. Moreover, γ_a and γ_n are positive efficiencies. There exists a unique positive solution to (1) for any $a(\cdot)$, $n(\cdot) \in L^1([0,+\infty);[0,+\infty))$ and initial condition $G(0) = G_0 > 0$.

As in [5], we assume for the firm a linear in goodwill and quadratic in costs profit. We consider the following discount profit with discount factor $\rho > 0$:

$$J_F = \int_0^{+\infty} \left(\pi G(t) - \frac{\kappa_a a^2(t)}{2} - \frac{\kappa_n n^2(t)}{2}\right) e^{-\rho t}\, dt \tag{2}$$

where π, κ_a, $\kappa_n > 0$.

As in [3], Even though native advertising positively influences the brand value of the firm, platform's consumers may perceive it as deceptive. Hence, native advertising may cause a loss of trust in the media's credibility [3].

Currently, no dynamic models address the credibility issue within an optimal control framework. We propose to formalise the evolution of credibility similarly to goodwill's one. Hence, the credibility of a media outlet diminishes over time if its content is not regularly updated [9] (at a rate $\varepsilon > 0$). In our model, the media credibility motion equation is given by

$$\dot{C}(t) = -\varepsilon C(t) + w(t) - \alpha n(t)\,. \tag{3}$$

The above equation describes the high-quality content investment to maintain the credibility by the flow $w(\cdot) \in L^1([0,+\infty);[0,+\infty))$, and the negative effect caused by native advertising by the parameter $\alpha > 0$ Marketing research findings ([1, 2, 12]) support this assumption. In our model, standard advertising does not affect media credibility being this content easily recognisable by consumers.

The media profit is defined similarly to the firm's one, with an additional revenue given by the publication of advertising content:

$$J_M = \int_0^{+\infty} \left(\eta C(t) - \frac{\sigma w^2(t)}{2} + \frac{\kappa_a a^2(t)}{2} + \frac{\kappa_n n^2(t)}{2}\right) e^{-\rho t}\, dt \tag{4}$$

where all parameters are positive. By (3), the impact of native advertising may result in a credibility equal to zero. As a result, native advertising damage also the media payoff (4). Hence, to ensure a positive profit, we assume the media to set a constraint by limiting native advertising on its platform. Let $N(\cdot) \in L^1([0,+\infty);[0,+\infty))$ be a control for the media outlet such that

$$n(t) \in [0, N(t)] \qquad \forall t \in [0,+\infty)\,. \tag{5}$$

2 Open-loop Stackelberg equilibrium

Searching for an open-loop Stackelberg equilibrium [6, Ch. 5, p. 113], we consider the necessary and sufficient conditions [10, Ch. 3, p. 234, Th. 12, Th. 13].

The research of an open-loop Stackelberg equilibrium is based on the following steps. First, the Leader states his strategy $(\overline{w}(\cdot), \overline{N}(\cdot))$. Then, the Follower solves his optimal control problem

$$
\begin{aligned}
\max_{(a,n)} \quad & J_F = \int_0^{+\infty} \left(\pi G(t) - \frac{\kappa_a a^2(t)}{2} - \frac{\kappa_n n^2(t)}{2} \right) e^{-\rho t}\, dt \\
& \dot{G}(t) = -\delta G(t) + \gamma_a a(t) + \gamma_n n(t), \qquad G(0) = G_0 \\
& \dot{C}(t) = -\varepsilon C(t) + \overline{w}(t) - \alpha n(t), \qquad C(0) = C_0 \\
& a(t) \geq 0, \qquad n(t) \in [0, \overline{N}(t)]
\end{aligned} \tag{6}
$$

by taking into account the Leader's strategies. Finally, the Leader solve its optimal control problem given the Follower's best response functions.

To solve (6), we apply the Pontryagin Maximum Principle to the current Hamiltonian function of the Follower

$$
\begin{aligned}
\mathrm{H}_F^c(G, C, a, n, \overline{w}, \overline{N}, \lambda_1, \lambda_2) = \pi G - \frac{\kappa_a}{2} a^2 - \frac{\kappa_n}{2} n^2 + \\
+ \lambda_1(-\delta G + \gamma_a a + \gamma_n n) + \lambda_2(-\varepsilon C + \overline{w} - \alpha n)\,,
\end{aligned}
$$

with co-state functions $\lambda_1(\cdot)$, $\lambda_2(\cdot)$.

Proposition 1 *Given the Leader's controls $(\overline{w}(\cdot), \overline{N}(\cdot))$, the Follower's best response functions are*

$$
a^*(t) \equiv \frac{\pi \gamma_a}{\kappa_a(\rho + \delta)} \quad \textit{and} \quad n^*(t) = \min\left\{ \frac{\pi \gamma_n}{\kappa_n(\rho + \delta)}, \overline{N}(t) \right\} \tag{7}
$$

for all $t \in [0, +\infty)$.

We can now study the Leader's problem

$$
\begin{aligned}
\max_{w,N} \quad & J_M = \int_0^{+\infty} \left(\eta C(t) - \frac{\sigma w^2(t)}{2} + \frac{\kappa_a (a^*(t))^2}{2} + \frac{\kappa_n (n^*(t))^2}{2} \right) e^{-\rho t}\, dt \\
& \dot{G}(t) = -\delta G(t) + \gamma_a a^*(t) + \gamma_n n^*(t), \qquad G(0) = G_0 \\
& \dot{C}(t) = -\varepsilon C(t) + w(t) - \alpha n^*(t), \qquad C(0) = C_0 \\
& w(t) \geq 0, \qquad N(t) \geq 0\,.
\end{aligned} \tag{8}
$$

where $(a^*(\cdot), n^*(\cdot))$ are as in (7). To solve (8), we reason as before. Consider the the current Hamiltonian of the Leader

$$
\begin{aligned}
\mathrm{H}_L^c(G, C, w, N, \mu_1, \mu_2) = \eta C - \frac{\sigma}{2} w^2 + \frac{\kappa_a}{2} (a^*)^2 + \frac{\kappa_n}{2} (n^*(N))^2 + \\
+ \mu_1(-\delta G + \gamma_a a^* + \gamma_n n^*(N)) + \mu_2(-\varepsilon C + w - \alpha n^*(N))\,,
\end{aligned}
$$

where $\mu_1(\cdot)$ and $\mu_2(\cdot)$ are the co-state functions. We say that native advertising is **admissible** for the media when the upper bound imposed by the media is strictly positive, that is, $N(t) > 0$ for all $t \in [0, +\infty)$.

Proposition 2 *Native advertising is admissible for the media outlet if and only if*

$$\alpha < \bar{\alpha} = \frac{\gamma_n \pi(\rho + \varepsilon)}{2\eta(\rho + \delta)}\,. \tag{9}$$

The Leader's optimal control functions are

$$w^*(t) \equiv \frac{\eta}{\sigma(\rho + \varepsilon)} \quad \textit{and} \quad N^*(t) \in \begin{cases} \left[\frac{\pi\gamma_n}{\kappa_n(\rho+\delta)}, +\infty\right) & \alpha < \bar{\alpha} \\ \{0\} \cup \left[\frac{\pi\gamma_n}{\kappa_n(\rho+\delta)}, +\infty\right) & \alpha = \bar{\alpha} \\ \{0\} & \alpha > \bar{\alpha} \end{cases} \tag{10}$$

for all $t \in [0, +\infty)$.

If the damage parameter α is low, then the media accepts native advertising to increase its profit, while if α is too high, the media maintain its credibility by choosing $N^*(\cdot) = 0$. We recall that η is the the marginal revenue with respect to credibility, and σ is the marginal cost of the investment in credibility. By (9), we note that $\bar{\alpha}$ decreases in η. On the other hand, the investment in credibility $w^*(\cdot)$ increases in η and decreases in σ.

At this point, we can compute explicitly the Follower's optimal controls, and obtain the OLSE.

Proposition 3 *The open-loop Stackelberg equilibrium of our problem is $((w^*, N^*), (a^*, n^*))$, where*

$$w^*(t) \equiv \frac{\eta}{\sigma(\rho + \varepsilon)}, \qquad N^*(t) \in \begin{cases} \left[\frac{\pi\gamma_n}{\kappa_n(\rho+\delta)}, +\infty\right) & \alpha < \bar{\alpha} \\ \{0\} \cup \left[\frac{\pi\gamma_n}{\kappa_n(\rho+\delta)}, +\infty\right) & \alpha = \bar{\alpha} \\ \{0\} & \alpha > \bar{\alpha} \end{cases}$$

$$a^*(t) \equiv \frac{\pi\gamma_a}{\kappa_a(\rho + \delta)}, \qquad n^*(t) = \begin{cases} \frac{\pi\gamma_n}{\kappa_n(\rho+\delta)} & \alpha < \bar{\alpha} \\ \min\left\{\frac{\pi\gamma_n}{\kappa_n(\rho+\delta)}, N^*(t)\right\} & \alpha = \bar{\alpha} \\ 0 & \alpha > \bar{\alpha} \end{cases} \tag{11}$$

for all $t \in [0, +\infty)$, with $\bar{\alpha} = \dfrac{\gamma_n \pi(\rho + \varepsilon)}{2\eta(\rho + \delta)}$.

The above open-loop Stackelberg equilibrium is time consistent. This because of the noncontrollability of the optimal co-state functions of Follower [6, Def. 5.1].

Proposition 4 *Assume $\alpha < \bar{\alpha}$. The media outlet's credibility remains positive if and only if*

$$\alpha < \frac{\eta\kappa_n(\rho + \delta)}{\gamma_n \pi\sigma(\rho + \varepsilon)} =: \alpha_S\,. \tag{12}$$

Since κ_n and σ appear in the threshold α_S, but not in the admissibility threshold $\bar{\alpha}$, we cannot compare the two thresholds. Even though, when admissible, native advertising is always profitable to the media, the credibility at the steady-state C^*_{SS} may become negative.

3 Conclusions

We consider the problem from a media outlet's prospective. The main goal was to analyse the trade-off between gaining profit by publishing native advertising and the consequent credibility loss. We address the problem as a hierarchical differential game where the media acts as the Leader. We compute a time-consistent open-loop Stackelberg equilibrium, and obtain the conditions for the admissibility of native advertising. This admissibility depends on the damage parameter to credibility: if the negative effect is low, it is optima for the media outlet to accept native advertising, while if the damage is high, the media safeguards its credibility and avoid native advertising. Furthermore, we discover that being admissible is not a sufficient condition on native advertising for guaranteeing also positive credibility in the long term.

References

[1] Amazeen, M.A. & Muddiman, A.R., *Saving Media or Trading on Trust?*. Digital Journalism 6 (2018), 176–195. https://doi.org/10.1080/21670811.2017.1293488.

[2] Amazeen, M.A. & Wojdynski, B.W., *The effects of disclosure format on native advertising recognition and audience perceptions of legacy and online news publishers.* Journalism 21 (2020), 1965–1984. https://doi.org/10.1177/1464884918754829.

[3] Bachmann, P., Hunziker, S., & Rüedy, T., *Selling their souls to the advertisers? How native advertising degrades the quality of prestige media outlets.* Journal of Media Business Studies 16 (2019), 95–109. https://doi.org/10.1080/16522354.2019.1596723.

[4] Brambilla C., Buratto A. & Grosset L., *A differential game model for sponsored content.*. Journal of the Operational Research Society (2024).

[5] Buratto, A. & Zaccour, G., *Coordination of Advertising Strategies in a Fashion Licensing Contract.* Journal of Optimization Theory and Applications 142 (2009), 31–53.

[6] Dockner, E.J., Jörgensen, S., Van Long, N. & Sorger, G., "Differential games in economics and management science". Cambridge University Press, 2000. https://doi.org/10.1017/CBO9780511805127.

[7] Huang, J., Leng, M., & Liang, L., *Recent developments in dynamic advertising research.* European Journal of Operational Research 220 (2012), 591–609. https://doi.org/10.1016/j.ejor.2012.02.031.

[8] Jörgensen, S. & Zaccour, G., "Differential Games in Marketing". International Series in Quantitative Marketing. Springer, 2004. https://doi.org/10.1007/978-1-4419-8929-1.

[9] Saleh H.F., *Developing New Media Credibility Scale: A Multidimensional Perspective*. International Journal of Humanities and Social Sciences 10 (2016), 1351–1364. World Academy of Science, Engineering and Technology.

[10] Seierstad, A. & Sydsæter, K., *Optimal Control Theory with Economic Applications*. In: Advanced Textbooks in Economics 24 (1987). North-Holland.

[11] Tutaj, K. & van Reijmersdal, E.A., *Effects of online advertising format and persuasion knowledge on audience reactions*. Journal of Marketing Communications 18 (2012), 5–18. https://doi.org/10.1080/13527266.2011.620765.

[12] Wojdynski, B.W., *The Deceptiveness of Sponsored News Articles: How Readers Recognize and Perceive Native Advertising*. American Behavioral Scientist 60 (2016), 1475–1491.

[13] Wojdynski, B.W., *Advertorials and Native Advertising*. In Vos, T.P., Hanusch, F., Dimitrakopoulou, D., Geertsema-Sligh, M., & Sehl, A. (Eds.), *The International Encyclopedia of Journalism Studies* (2019), pp. 1–6. Wiley-Blackwell. https://doi.org/10.1002/9781118841570.iejs0062.

Classical Modular Forms and the k-square problem

Marco Baracchini (*)

1 Introduction

The idea of this seminar is to speak about modular forms. We would like to convince the audience that modular forms could be useful in order to study some arithmetic problems, give the rigourous definition of classical modular forms of a certain weight and level and give an idea of possible ways to study modular forms.

In the first Section we will speak about two arithmetic problems showing that they could be related to studying some modular forms.

In the second Section we will give the definition of what a modular form of a certain weight and level is.

In the last Section we briefly give an idea of a geometric interpretation of modular forms and how one can study modular forms using $p-$adic analysis instead of the complex one.

2 Motivations

To have an intuition you can think of modular form as a formal power series

$$\sum_{n\in\mathbb{N}} a_n q^n \in \mathbb{C}[[q]]$$

with complex coefficients that satisfy some invariance properties (that we will see in another section). The general idea is to prove that some numbers a_n that are related to an arithmetic problem can be seen as coefficients of a modular form; then one can study the rigidity of the modular form in order to deduce some properties of the numbers a_n and of the related arithmetic problem.

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: marco.baracchini@math.unipd.it . Seminar held on 28 February 2024.

2.1 k-squares problem

The first arithmetic problem that I would like to introduce is the k-square problem.

Fix $k \in \mathbb{N}$, one can ask if a natural number $n \in \mathbb{N}$ could be written as sum of k squares, i.e. if there are integer numbers $s_1, \dots, s_k \in \mathbb{Z}$ such that

$$n = s_1^2 + \cdots + s_k^2.$$

For example, if one fixes $k = 2$, this problem has a positive answer for the number $5 = 2^2 + 1^2$ but a negative answer for 6. More generally one can inquire about the number of ways to write a natural number $n \in \mathbb{N}$ as sum of k squares, i.e. compute

$$S_k(n) := \#\{(s_1, \dots, s_k) \in \mathbb{Z}^k \mid \sum_{i=1}^{k} s_i^2 = n\}.$$

It is possible to define the following power series using the numbers above:

$$f_k := \sum_{n \in \mathbb{N}} S_k(n) q^n \in \mathbb{Z}[[q]] \subset \mathbb{C}[[q]].$$

Via an easy computation one can show that $f_k = \theta^k$ where

$$\theta := \sum_{s \in \mathbb{Z}} q^{s^2}$$

is the Jacobi-Theta function. It turns out that θ^{2k} is a modular form of weight $k \in \mathbb{N}$ and level $\Gamma_0(4)$. Then one can approach the $k-$square problem using modular forms: one can write $\theta^4 = f_4$ and $\theta^8 = f_8$ as products of modular forms whose coefficients are known. This gives explicit formulas for the coefficients of f_4 and f_8:

Theorem 2.1 $S_4(n) = 8 \sum_{1 \leqslant d \leqslant n, 4 \nmid d \mid n} d, \quad S_8(n) = 16 \sum_{1 \leqslant d \leqslant n, d \mid n} d^3.$

Corollary 2.2 *For each $n \in \mathbb{N}$ there are four integers $s_1, s_2, s_3, s_4 \in \mathbb{Z}$ s.t.*

$$n = s_1^2 + s_2^2 + s_3^2 + s_4^2.$$

Proof. For the proof of the Theorem see for example the Section §3.1 of [Zag08] or the more introductory Master Thesis [Var]. The Corollary is an easy consequence: via the Theorem for each $n \in \mathbb{Z}_{>0}$ the number $S_4(n)$ is higher than 7, in particular it is non zero. □

One can also prove with modular form techniques a more general Theorem.

Theorem 2.3 *Let F be a totally real number field. Fix a choice of $F_{>0} \subset F$. If there is a $k \in \mathbb{N}$ s.t. for each $x \in \mathcal{O}_F \cap F_{>0}$ there are $s_1, \cdots, s_k \in \mathcal{O}_F$ with*

$$x = s_1^2 + \cdots + s_k^2,$$

then

$$F = \mathbb{Q} \quad or \quad F = \mathbb{Q}(\sqrt{5}).$$

Proof. Look at the Theorem 3 in [JW07] proved in the article [Maa41]. □

In order to understand this Theorem one has to define what a totally real number field is and the ring of integers of a number field.

A number field is a finite extension of the rational numbers $\mathbb{Q}$.

A totally real number field F is a number field such that for each morphism of fields $\sigma : F \longrightarrow \mathbb{C}$ one has that the image of the morphism is contained in the real numbers, i.e. $\sigma(F) \subset \mathbb{R}$.

Each number field F is an algebraic extension of $\mathbb{Q}$, hence for each number field there exists an embedding $\sigma : F \to \mathbb{C}$. For a totally real field one can fix an embedding like that and one knows that this embedding splits as

$$F \hookrightarrow \mathbb{R} \hookrightarrow \mathbb{C},$$

hence one can define $F_{>0} := F \cap \sigma^{-1}(\mathbb{R}_{>0})$. The choice of $F_{>0}$ is not unique as it depends on the choice of σ.

For example $F = \mathbb{Q}(\sqrt{2}) \cong \mathbb{Q}[x]/(x^2 - 2)$ is a totally real number field and the choice of $F_{>0}$ corresponds to the (not unique) choice of $\sqrt{2} \in F$. Two examples of non totally real fields are $\mathbb{Q}(i) \cong \mathbb{Q}[x]/(x^2 + 1)$ and $\mathbb{Q}(\sqrt[3]{2}) \cong \mathbb{Q}[x]/(x^3 - 2)$.

If F is a number field, the ring of integers $\mathcal{O}_F$ is the integral closure of $\mathbb{Z}$ in F, i.e.

$$\mathcal{O}_F := \{x \in F \mid \exists p = \sum_{n=0}^{N} a_n X^n \in \mathbb{Z}[X] \text{ s.t. } a_N = 1,\ p(x) = 0\}.$$

For example if $F = \mathbb{Q}(i) \cong \mathbb{Q}[x]/(x^2 + 1)$, then $\mathcal{O}_F = \mathbb{Z}[i] = \mathbb{Z}[x]/(x^2 + 1)$.

2.2 Studying rational points of elliptic curves

For more details about the topic treated in this section one can look at the book of Silverman [Sil09] and the book by Diamond and Shurman [DS05].

The arithmetic problem treated in this section is to find rational solutions of elliptic curves defined over the rational numbers. An elliptic curve is the set of points that solve a smooth cubic equation with rational coefficients. For example points $(x, y) \in \mathbb{Q}^2$ such that

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

Looking for rational solutions of the above equation is the same as looking for projective integral solutions $[x : y : z] \in \mathbb{P}^3(\mathbb{Q}) \setminus \{[0 : 1 : 0]\}$

$$zy^2 + yz^2 = x^3 - zx^2.$$

The advantage of the projective plane is that

$$\mathbb{P}^3(\mathbb{Q}) = \mathbb{P}^3(\mathbb{Z}),$$

so one can look for integer solutions (that means $x, y, z \in \mathbb{Z}$).

Once one has a problem over the integer numbers one can approximate this problem reducing modulo a prime number p: one can reduce the equation modulo p and can study the solutions in $\mathbb{Z}/p\mathbb{Z}$ varying the prime p. It is known that the number of solutions of a cubic equation modulo p is around p, more precisely

$$\mid p - \#\{(\bar{x}, \bar{y}) \in (\mathbb{Z}/p\mathbb{Z})^2 \mid \bar{y}^2 + \bar{y} = \bar{x}^3 - \bar{x}^2\} \mid \leqslant 2\sqrt{p}.$$

The knowledge of the number of solutions modulo p for each p is equivalent to the knowledge of

$$b_p := p - \#\{(\bar{x}, \bar{y}) \in (\mathbb{Z}/p\mathbb{Z})^2 \mid \bar{y}^2 + \bar{y} = \bar{x}^3 - \bar{x}^2\},$$

and it is known that two elliptic curves over $\mathbb{Q}$ with the same numbers b_p have some common properties (they are isogenous, i.e. one is a finite quotient of the other). Then knowing the numbers b_p one knows the isogeny class of an elliptic curve.

The difficult point in this procedure is that the numbers b_p are infinitely many, then it is not possible to compute them all without knowing some extra information.

Theorem 2.4 *There is a modular form $f = \sum_{n\in\mathbb{N}} a_n q^n \in \mathbb{C}[[q]]$ of weight 2 and level $\Gamma_0(N)$ for an integer $N \in \mathbb{Z}$ such that for all primes p of good reduction (all the primes that do not divide the discriminant of the cubic equation)*

$$a_p = b_p, \quad a_0 = 0.$$

Thanks to this famous Theorem we can compute a_p for a finite number of primes p, compute N (that is called the conductor of the elliptic curve), compute the space of modular forms of weight 2 and level $\Gamma_0(N)$; there are some databases on the net, as in `https://www.lmfdb.org/ModularForm/`. Then you can look for modular forms with a_p as $p-$coefficients. If you compute enough a_p you will be able to identify the modular form, hence you get all the a_p of your elliptic curve (not only the finitely many that you computed).

In our case the modular form attached to the equation $y^2 + y = x^3 - x^2$ is

$$f = q \prod_{n\in\mathbb{N}_{>0}} (1 - q^{11n})^2 (1 - q^n)^2$$

of weight 2 and level $\Gamma_0(11)$. Then you can expand this product in order to get a_p for $p \neq 11$, since 11 is the only prime of bad reduction of the elliptic curve of our example.

3 Definition of modular forms

In this section we want to define what a modular form of level $\mathrm{SL}_2(\mathbb{Z})$ and weight $k \in \mathbb{Z}$ is. In order to do this we need some tools. For more details look at the Diamond-Shurman [DS05]. Moreover for the level $\mathrm{SL}_2(\mathbb{Z})$ case there is a nice introduction in Chapter VII of [Ser93], written by J.P. Serre.

3.1 Algebraic tools

Let us consider

$$\mathrm{SL}_2(\mathbb{Z}) := \{\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mid a, b, c, d \in \mathbb{Z} \ det(\gamma) = ad - bc = 1\}$$

and the upper half complex plane

$$\mathcal{H} := \{\tau \in \mathbb{C} \mid \mathrm{Im}(\tau) > 0\}.$$

We can define an action of $\mathrm{SL}_2(\mathbb{Z})$ on $\mathcal{H}$: for each $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ and $\tau \in \mathcal{H}$

$$\gamma \cdot \tau := \frac{a\tau + b}{c\tau + d}.$$

One can verify that $\gamma_1 \cdot \tau \in \mathcal{H}$ and $\gamma_1 \cdot (\gamma_2 \cdot \tau) = (\gamma_1\gamma_2) \cdot \tau$ for each $\tau \in \mathcal{H}$, $\gamma_1, \gamma_2 \in \mathrm{SL}_2(\mathbb{Z})$.

3.2 Analytic tools

Let $f : \mathcal{H} \to \mathbb{C}$ be an holomorphic function.

Lemma 3.1 *If $f(\tau) = f(\tau + 1)$, then f has a Fourier expansion that converges absolutely at f, i.e.*

$$f(\tau) = \sum_{n \in \mathbb{Z}} a_n e^{2\pi i n \tau},$$

where $a_n \in \mathbb{C}$.

Proof. Fix $q := e^{2\pi i \tau} \in B_{\mathbb{C}}(0, 1) \setminus \{0\}$. We can define a logarithmic (biholomorphic) function

$$\log : B_{\mathbb{C}}(0, 1) \setminus \mathbb{R}_{\leqslant 0} \longrightarrow \{\tau \in \mathbb{C} \mid 1/2 < \mathrm{Re}(\tau) \leqslant 1/2\}$$

such that $\log(e^{2\pi i \tau}) = \tau$ for each $\tau \in \{z \in \mathbb{C} \mid 1/2 < \mathrm{Re}(\tau) \leqslant 1/2\}$.

Let $g(q) := f \circ \log(q)$, then g is a holomorphic function and $f(\tau) = g(e^{2\pi i \tau})$. Since $f(\tau) = f(\tau+1)$ one can show that g could be (uniquely) analytically continued in $B_{\mathbb{C}}(0, 1) \setminus \{0\}$ and one get that $f(\tau) = g(e^{2\pi i \tau})$ for each $\tau \in \mathcal{H}$.

Now if one looks at the Laurent expansion of g at 0 one gets that

$$g(q) = \sum_{n \in \mathbb{Z}} a_n q^n,$$

with $a_n \in \mathbb{C}$, then

$$f(\tau) = g(e^{2\pi i \tau}) = \sum_{n \in \mathbb{Z}} a_n e^{2\pi i n \tau}.$$

□

Definition A holomorphic function $f : \mathcal{H} \to \mathbb{C}$ s.t. $f(\tau) = f(\tau + 1)$ is **holomorphic at** $i\infty$ iff the coefficients of its Fourier expansion $a_n = 0$ for all negative $n \in \mathbb{Z}_{<0}$.

Observe that via the change of coordinate $q(\tau) = e^{2\pi i\tau}$ we get

$$0 = \lim_{\mathbb{R}\ni\alpha\to\infty} e^{2\pi(i\alpha)} = \lim_{\mathbb{R}\ni\alpha\to\infty} q(\tau = i\alpha)$$

hence when the coordinate τ approaches $i\infty$ the coordinate q tends to 0. This could explain why we say that f is holomorphic at $i\infty$ if g is holomorphic at 0.

3.3 Definition of level $\mathrm{SL}_2(\mathbb{Z})$ and weight k modular forms

Now we have all the tools in order to define modular forms.

Definition A **modular form** of weight $k \in \mathbb{Z}$ and level $\mathrm{SL}_2(\mathbb{Z})$ is a holomorphic function $f : \mathcal{H} \to \mathbb{C}$ s.t.

- $f(\gamma \cdot \tau) = (c\tau + d)^k f(\tau)$ for each $\gamma \in \mathrm{SL}_2(\mathbb{Z})$;
- f is holomorphic at $i\infty$.

Observe that since $f(\tau) = 1^k f(\tau) = f(\left(\begin{smallmatrix}1 & 1\\ 0 & 1\end{smallmatrix}\right)\cdot\tau = f(\tau + 1)$ it makes sense to ask that f is holomorphic at $i\infty$.

The series $f = \sum_{n\in\mathbb{N}} a_n e^{2\pi i n\tau} = \sum_{n\in\mathbb{N}} a_n q^n \in \mathbb{C}[[q]]$ is called the $q-$expansion of f.

3.4 Congruence subgroups $\Gamma \leqslant \mathrm{SL}_2(\mathbb{Z})$

The level of a modular form is usually denoted with a Γ and it is a congruence subgroup of $\mathrm{SL}_2(\mathbb{Z})$. Let us define what a congruence subgroup is and investigate some basic properties. Let $N \in \mathbb{N}_{>1}$, we denote

$$\Gamma(N) := \{\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) \mid (a-1) \equiv b \equiv c \equiv (d-1) \equiv 0 \bmod N\} \leqslant \mathrm{SL}_2(\mathbb{Z})$$

the principal congruence subgroup of level N. It is formed by all the matrices that are congruent to the identity matrix modulo N.

Definition A subgroup $\Gamma \leqslant \mathrm{SL}_2(\mathbb{Z})$ is a **congruence subgroup** (of level N) if there exists an $N \in \mathbb{N}_{>0}$ such that $\Gamma(N) \leqslant \Gamma$.

For example $\mathrm{SL}_2(\mathbb{Z}) = \Gamma(1)$ is the principal congruence subgroup of level 1. Any congruence subgroup Γ has the following properties (for more details look at [DS05]):

- Γ has finite index in $\mathrm{SL}_2(\mathbb{Z})$;
- the quotient $\Gamma\backslash\mathcal{H}$ is a (not compact) Riemann variety;
- one can add a finite number of extra points (called cusps) to the quotient $\Gamma\backslash\mathcal{H}$ such that it becomes a compact Riemann variety;

- the set of the cusps is $\mathrm{SL}_2(\mathbb{Z})/\Gamma$;
- for any $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ the group $\gamma^{-1}\Gamma\gamma \leqslant \mathrm{SL}_2(\mathbb{Z})$ is a congruence subgroup.

In the case of $\Gamma = \mathrm{SL}_2(\mathbb{Z})$ there is only one cusp: the point at $i\infty$. This cusp corresponds to the unique element of $\mathrm{SL}_2(\mathbb{Z})/\mathrm{SL}_2(\mathbb{Z}) = \{[I_2]\}$. Observe that $\gamma_N := \left(\begin{smallmatrix} 1 & N \\ 0 & 1 \end{smallmatrix}\right) \in \Gamma(N)$, hence for any congruence subgroup Γ there exists an $N \in \mathbb{N}$ such that $\gamma_N \in \Gamma$. Observe that the action of γ_N on $\mathcal{H}$ is given by

$$\gamma \cdot \tau = \tau + N.$$

3.5 Definition of level Γ and weight k modular forms

Fix a congruence subgroup Γ, let $f : \mathcal{H} \to \mathbb{C}$ be a holomorphic function such that

$$f(\gamma \cdot \tau) = (c\tau + d)^k f(\tau), \quad \text{for each } \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma.$$

Since there is $\gamma_N \in \Gamma$ we know that

$$f(N(\tau+1)) = f(\gamma_N \cdot (N\tau)) = (0\tau + 1)^k \cdot f(N\tau) = f(N\tau).$$

Via the analytic tool we get that

$$f(\tau) = g(e^{\frac{2\pi i \tau}{N}}) = \sum_{n \in \mathbb{Z}} a_n q^{\frac{n}{N}}.$$

If $a_n = 0$ for all $n < 0$ we say that f is holomorphic at $i\infty$. One would like to check that f is holomorphic at all the cusps, not only at $i\infty$. For a general cusp $[\mathsf{c}] \in \mathrm{SL}_2(\mathbb{Z})/\Gamma$, where $\mathsf{c} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$ we define

$$f_{[\mathsf{c}]_k}(\tau) := (C\tau + D)^{-k} f(\mathsf{c} \cdot \tau).$$

Via the $\Gamma-$invariance of f and with an explicit computation one can check that for any $\gamma \in \mathsf{c}^{-1}\Gamma\mathsf{c}$,

$$f_{[\mathsf{c}]_k}(\gamma \cdot \tau) = (c\tau + d)^k f_{[\mathsf{c}]_k}(\tau).$$

Since $\mathsf{c}^{-1}\Gamma\mathsf{c}$ is a congruence subgroup there exists an $N_{\mathsf{c}} \in \mathbb{N}$ such that $\gamma_{N_{\mathsf{c}}} \in \Gamma(N_{\mathsf{c}})$. Then $f_{[\mathsf{c}]_k}(\tau + N_{\mathsf{c}}) = f_{[\mathsf{c}]_k}(\tau)$ and via the analytic tool

$$f_{[\mathsf{c}]_k}(\tau) = \sum_{n \in \mathbb{Z}} a_{\mathsf{c},n} e^{\frac{2\pi i n \tau}{N\gamma_1}}.$$

Then f is holomorphic at the cusp $[c]$ iff $f_{[\mathsf{c}]_k}$ is holomorphic at $i\infty$, i.e. if

$$a_{\mathsf{c},n} = 0 \quad \text{for all } n \in \mathbb{Z}_{<0}.$$

Definition A **modular form** of level Γ and weight $k \in \mathbb{N}_{>0}$ is a holomorphic function $f : \mathcal{H} \to \mathbb{C}$ s.t.

- $f(\gamma \cdot \tau) = (c\tau + d)^k f(\tau)$ for each $\gamma \in \Gamma$;
- f is holomorphic at all the cusps of $\Gamma\backslash\mathcal{H}$.

We denote the $\mathbb{C}$-vector space of modular forms of weight k and level Γ as $M_k(\Gamma)$

For completion we now write the definition of the congruence subgroup $\Gamma_0(N)$ that we cited in the motivation sections. Let $N \in \mathbb{N}$,

$$\Gamma_0(N) := \{\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) \mid c \equiv 0 \bmod N\} \leqslant \mathrm{SL}_2(\mathbb{Z}). \tag{1}$$

It is the group of matrices in $\mathrm{SL}_2(\mathbb{Z})$ that reduce to a matrix of the form $\begin{pmatrix} * & * \\ 0 & * \end{pmatrix}$ modulo N. Observe that it is a congruence subgroup since $\Gamma(N) \leqslant \Gamma_0(N)$.

4 p-adic analysis and modular forms

If someone would like to study more about the modern techniques I suggest to read the introduction written by R. Brasca in the article [Bra12]. In this introduction Brasca makes a summary of the various definitions of modular forms. In this last Section we would like to motivate the study of p-adic modular forms: I would like to explain why $p-$adic analysis could help in studying modular forms over $\mathbb{C}$, even if these modular forms belong to the complex analysis.

The starting point is that one can define modular forms over any ring, not just over the complex numbers. A possible way to understand this fact is following the article [Kat73] by Nicholas Katz. He defines modular forms in a geometric way. His idea uses the fact that one can parametrize all the elliptic curves with a fixed differential via the Rieman surface $\mathcal{H}$. More precisely, for each $\tau \in \mathcal{H}$ one can build the torus

$$E_\tau := \frac{\mathbb{C}}{\mathbb{Z} \oplus \mathbb{Z}\tau}.$$

It turns out that E_τ corresponds to an elliptic curve over $\mathbb{C}$ and we can fix $\mathrm{d}z$ as the canonical differential. Katz uses this fact in order to see a modular form as a rule F that associates to an elliptic curve E defined over $\mathbb{C}$ with a fixed differential ω, a complex number $F(E, \omega)$. One has to ask for some nice properties of this rule in order to recover an equivalent definition of a modular form. Denote with $\mathcal{M}_\Gamma$ the modular curve that parametrizes the elliptic curves over $\mathbb{C}$ with some structure (that depends on the level Γ). Katz showed that there is a sheaf $\underline{\omega}$ on $\mathcal{M}_\Gamma$ such that the modular forms of weight k and level Γ are exactly the global section of $\underline{\omega}^{\otimes k}$, i.e.

$$M_k(\Gamma) = \underline{\omega}^{\otimes k}(\mathcal{M}_\Gamma).$$

Thanks to this geometric interpretation, one can define modular forms over any ring where the modular curve is defined. Also in this geometric setting one gets a notion of $q-$expansion that coincides with the “classical" one.

For example one can define modular forms over the p-adic numbers $\mathbb{Q}_p$, use $p-$adic analysis (and not only complex/real-analysis) and deduce some properties of the "classical" modular forms (i.e. modular forms over $\mathbb{C}$). We denote the $\mathbb{Q}_p-$vector space of modular forms of weight $k \in \mathbb{Z}$ and level Γ as $M_k(\Gamma, \mathbb{Q}_p)$

Using $p-$adic analysis in the study of modular forms gives a big advantage: the weight space $\mathbb{Z}$ is no more discrete! Indeed $\mathbb{Z}$ inside $\mathbb{R}$ is discrete, but $\mathbb{Z}$ inside $\mathbb{Q}_p$ is no more discrete. Hence one can work in the (infinite dimensional) $\mathbb{Q}_p-$vector space

$$\bigcup_{k\in\mathbb{Z}} M_k(\Gamma, \mathbb{Q}_p) \subset \mathbb{Q}_p[[q^{\frac{1}{N}}]].$$

This vector space is completed with respect to the $\infty-$norm

$$||\sum_{n\in\mathbb{N}} a_n q^{\frac{n}{N}}||_\infty := \max_{n\in\mathbb{N}} |a_n|$$

defined on

$$\mathbb{Q}_p < q^{\frac{1}{N}} >:= \{\sum_{n\in\mathbb{N}} a_n q^{\frac{n}{N}} \in \mathbb{Q}_p[[q^{\frac{1}{N}}]] \ \mid \ |a_n| \to 0\}$$

In this p-adic setting one can hope to find a family of modular forms

$$f_k := \sum_{n\in\mathbb{N}} a_{n,k} q^{\frac{n}{N}} \in \mathbb{Q}[[q]]$$

where f_k varies continuously on $k \in \mathbb{Z}$, i.e.

$$||f_{k_1} - f_{k_2}||_\infty \overset{(k_1-k_2)\to 0}{\longrightarrow} 0.$$

Since $\mathbb{Z}$ is not discrete with respect to the p-adic norm, this condition gives some rigidity to a family of modular forms and for example one can hope to prove some statements on classical modular forms (for example modular forms with rational coefficients) that fit in a p-adic family.

References

[Bra12] Riccardo Brasca, *p-adic modular forms of non-integral weight over Shimura curves*. Compositio Mathematica 149/1 (November 2012), 32–62.

[DS05] Fred Diamond and Jerry Shurman, "A first course in modular forms". Volume 228 of Graduate Texts in Mathematics. Springer, 2005.

[JW07] Chun-Gang Ji and Da-Sheng Wei, *Sums of integral squares in cyclotomic fields*. Comptes Rendus Mathematique 344/7 (2007), 413–416.

[Kat73] Nicholas M. Katz, *p-adic properties of modular schemes and modular forms.* In Modular Functions of One Variable III, pp. 69–190, Springer, Berlin (1973).

[Maa41] Hans Maass, *Uber die Darstellung total positiver Zahlen des Körpers* $\mathbb{R}(\sqrt{5})$ *als Summe von drei Quadraten.* Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg 1 (1941), 185–191.

[Ser93] Jean-Pierre Serre, "A course in arithmetic". Springer, 1993.

[Sil09] Joseph H. Silverman, "The arithmetic of elliptic curves". Volume 106 of Graduate Texts in Mathematics. Springer, 2nd edition, 2009.

[Var] Ila Varma, "Sums of Squares, Modular Forms, and Hecke Characters". Master Thesis.

[Zag08] Don Zagier. "Elliptic Modular Forms and Their Applications" (pages 1–103).Springer Berlin Heidelberg, Berlin, Heidelberg, 2008

p-adic numbers and characteristic p

Pietro Vanni (*)

Abstract. For each prime p, p-adic numbers form an extension of the rational numbers that, being topologically complete, allows one to use analytic methods in arithmetic. In this talk I will introduce p-adic numbers outlining their basic properties and the role they play in number theory. Then I will give an idea on how one can employ p-adic numbers to study algebraic varieties (i.e. systems of polynomial equations) in characteristic p.

1 Notation

In this section we fix some notation:

- we denote by $\mathbb{N}$ the set of natural numbers.
- $\mathbb{Z}$ is the ring of integers.
- p will be a fixed prime number.
- $\mathfrak{P}$ will be the set of prime numbers.
- If $n \in \mathbb{Z}$, we denote by $\nu_p(n)$ the p-adic valuation of n, i.e. the biggest exponent $m \in \mathbb{N}$ such that p^m divides n.
- $\mathbb{F}_p$ will denote the finite field with p elements.
- $\mathbb{Q}$ will be the field of rational numbers.
- $\mathbb{R}$ will be the field of real numbers.
- $\mathbb{C}$ will be the field of complex numbers.
- K will be used to denote a general field.
- $K[x]$ will denote the ring of polynomial functions on K, with variable x.
- $K[\![x]\!]$ will be the ring of formal power series with coefficients in K and variable x.

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: vanni@math.unipd.it . Seminar held on 13 March 2024.

- $K(\!(x)\!)$ will denote the ring of formal Laurent series with coefficients in K, and with variable x.
- If $d \in \mathbb{N}$, $\mathbf{A}_K^d$ will be the affine space of dimension d over K.

2 Introduction to p-adic numbers

In this section we define p-adic numbers and we look at some or their properties.

Let $|\cdot|$ denote the real absolute value on $\mathbb{Q}$. It is well known that if we complete $\mathbb{Q}$ for the metric topology induced by $|\cdot|$ we obtain $\mathbb{R}$.
However in $\mathbb{Q}$ we can have different kinds of absolute values.

Definition 2.1 An *absolute value* on $\mathbb{Q}$ is a function $\mathfrak{v}(\cdot) : \mathbb{Q} \to \mathbb{R}$ such that:

(a) $\mathfrak{v}(r) \geq 0$ for each $r \in \mathbb{Q}$.

(b) $\mathfrak{v}(r) = 0$ if and only $r = 0$.

(c) $\mathfrak{v}(r, r') \leq \mathfrak{v}(r) + \mathfrak{v}(r')$ for each r and r' in $\mathbb{Q}$.

(d) $\mathfrak{v}(rr') = \mathfrak{v}(r)\mathfrak{v}(r')$ for each r and r' in $\mathbb{Q}$.

Note that every absolute value on $\mathbb{Q}$ induces a distance d on $\mathbb{Q}$ in the following way:

$$d(r, r') = \mathfrak{v}(r - r'). \tag{1}$$

We know give a more "exotic" example.

Example 2.2 Let $r = \frac{a}{b} \in \mathbb{Q}$, with a and b in $\mathbb{Z}$. The *p-adic absolute value* of r is defined as

$$|r|_p = p^{\nu_p(b) - \nu_p(a)}.$$

The function $|\cdot|_p : \mathbb{Q} \to \mathbb{R}$ is easily seen to be an absolute value on $\mathbb{Q}$. Note that a rational number is as small for this absolute value as much as it is divisible by p. For example the sequence $\{|p^n|\}_{n \in \mathbb{N}}$ converges to 0. This absolute value induces a metric topology on $\mathbb{Q}$ that we will call the *p-adic topology*.

Indeed it is easy to see that the p-adic absolute value satisfies the follwing strengthening of inequality 3:

$$\left|r + r'\right|_p \leq \max(|r|_p, \left|r'\right|_p). \tag{2}$$

An absolute value that satisfies such property is called *non archimedean*. This is because inequality (2) implies that $|n|_p \leq 1$ for each $n \in \mathbb{Z}$, and thus the archimedean property of the real absolute value does not hold for $|\cdot|_p$.
One can wonder if there exist other absolute values on $\mathbb{Q}$, but this is not the case by the following theorem.

Theorem 2.3 (Ostrowski) *The nontrivial absolute values of* $\mathbb{Q}$ *are, up to equivalence,* $\{|\cdot|\} \cup \{|\cdot|_q\}_{q\in\mathfrak{P}}$.

Here the trivial absolute value is the value sending 0 to 0 and every other rational number to 1. Two absolute values are called equivalent if they induces equivalent topologies on $\mathbb{Q}$. See [Kob84, Theorem I.1] for a proof of the theorem above.
We can now define the field of p-adic numbers, in analogy with the definition of $\mathbb{R}$ as the completion of $\mathbb{Q}$ for the real absolute value.

Definition 2.4 The field of *p-adic numbers* is defined as the completion of $\mathbb{Q}$ for the p-adic absolute value (see [Lan02], Proposition XII.2.1 for the definition and construction of completions of fields). We will denote it by $\mathbb{Q}_p$.

The p-adic valuation extends to $\mathbb{Q}_p$ by continuity. Let's give an example of a p-adic number.

Example 2.5 Is it easy to see that the sum

$$\sum_{n\in\mathbb{N}} p^n$$

converges in $\mathbb{Q}_p$. So it defines a p-adic number.

Indeed it can be show that every p-adic number admits an unique expansion of the form

$$\sum_{-\infty<<m}^{+\infty} a_m p^m,$$

where $a_m \in \{0, 1, \dots, p-1\}$ (see [Gou20, Corollary 4.3.4]). Such expansions are called *p-adic expansions*. This characterization of p-adic numbers seems to hint that $\mathbb{Q}_p$ "looks like" the field $\mathbb{F}_p(\!(p)\!)$ (as a set). Indeed this intuition can be made rigorous by the theory of Witt vectors (see [Ser79, section II.6]). In fact $\mathbb{Q}_p = W(\mathbb{F}_p)[\frac{1}{p}]$.
The theory of p-adic expansions suggests an obvious notion of p-adic integers.

Definition 2.6 The ring of *p-adic integers* is the subring of $\mathbb{Q}_p$ constituted by p-adic expansions with no negative powers of p. We will denote this ring by $\mathbb{Z}_p$.

The p-adic number in Example 2.5 is a p-adic integer. Note that p-adic integers can be equivalently characterized as elements in $\mathbb{Q}_p$ with valuation less or equal than 1. Moreover we have $\mathbb{Q}_p = \mathbb{Z}_p[\frac{1}{p}]$.
Again one can say that p-adic integers "look like" $\mathbb{F}_p[\![p]\!]$. We will try to motivate this analogy more in the next section.

Remark 2.7 Note that the quotient ring $\mathbb{Z}_p/p\mathbb{Z}_p$ is identified with $\mathbb{F}_p$.

3 Geometric interpretation of $\mathbb{Z}_p$

In this section we give a geometric interpretation of $\mathbb{Z}_p$, in relation with the algebro-geometric viewpoint on the ring $\mathbb{Z}$.

To give a gemetric meaning to $\mathbb{Z}_p$, we must first realize $\mathbb{Z}$ as a geometric space. This realization is better explained using an analogy with the classical description of the space $\mathbf{A}^1_{\mathbb{C}}$ in algebraic geometry, that we recall below.
Consider the couple $(\mathbf{A}^1_{\mathbb{C}}, \mathbb{C}[x])$, i.e. the affine line endowed with the complex polynomial functions over it. The maximal ideals of $\mathbb{C}[x]$ are the principal ideals generated by the monials $(x - z)$, where $z \in \mathbb{C}$. This follows from the Fundamental Theorem of Algebra together with the fact that $\mathbb{C}[x]$ is a principal ideal domain. Thus it is clear the maximal ideals of $\mathbb{C}[x]$ correspond to the points of $\mathbf{A}^1_{\mathbb{C}}$. So we can say that $\mathbf{A}^1_{\mathbb{C}}$ is the space of maximal ideals of $\mathbb{C}[x]$. One can also give an algebraic meaning to the operation of evaluating a polynomial function $f(\cdot) \in \mathbb{C}[x]$ in a point $z \in \mathbb{C}$. This operation correspond to take the image of f trough the projection

$$\mathbb{C}[x] \to \mathbb{C}[x]/(x - z)\mathbb{C}[x] = \mathbb{C},$$

as it is easily seen taking the Taylor expansion of $f(\cdot)$ in z.
We want to define a geometric space $\mathrm{MaxSpec}(\mathbb{Z})$ associated to the ring $\mathbb{Z}$ along this lines. To do so we define $\mathrm{MaxSpec}(\mathbb{Z})$ to be the set of maximal ideals of $\mathbb{Z}$, that correspond to the set of prime numbers $\mathfrak{P}$. The ring of "functions" on this space is $\mathbb{Z}$. To evaluate a function $m \in \mathbb{Z}$ in the point $p \in \mathfrak{P}$ we take the image of m along the projection

$$\mathbb{Z} \to \mathbb{Z}/p\mathbb{Z} = \mathbb{F}_p.$$

Note that in this case the "functions" on $\mathrm{MaxSpec}(\mathbb{Z})$ take values in different fields for different points.
As already mentioned, one can expand functions on $\mathbf{A}^1_{\mathbb{C}}$ on at the point $z \in \mathbb{C}$ using the Taylor expansion.
In the same way one can expand a function on $\mathrm{MaxSpec}(\mathbb{Z})$ at the point p taking its p-adic expansion.
If we formally complete such expansions (i.e. we let these expansions go to infinity) we obtain the formal power series $\mathbb{C}[\![x]\!]$ in the case of $\mathbf{A}^1_{\mathbb{C}}$ and the ring $\mathbb{Z}_p$ in the case of $\mathrm{MaxSpec}(\mathbb{Z})$. In a way that can be precised by formal algebraic geometry (see [Har77, section II.9]), $\mathbb{C}[\![x - z]\!]$ can be regarded as the ring of functions defined in an infinitesimal neighborhood of the point z. Similarly $\mathbb{Z}_p$ should be regarded as the ring of functions on $\mathrm{MaxSpec}(\mathbb{Z})$ that are defined in an infinitesimal neighborhood of p. So $\mathbb{Z}_p$ represents a sort of "infinitesimal thickening" of $\mathbb{F}_p$, because $\mathbb{F}_p$ constitute the functions defined on the point p (the constant functions).
We report these analogies in the following table.

Space	$\mathbf{A}^1_{\mathbb{C}}$	$\mathrm{MaxSpec}(\mathbb{Z})$
Functions	$\mathbb{C}[x]$	$\mathbb{Z}$
Point	$(x - z)$	p
Evaluation	$\mathrm{mod}\ (x - z)$	$\mathrm{mod}\ p$
Values	$\mathbb{C}$	$\mathbb{F}_p$
Infinitesimal functions	$C[\![x - a]\!]$	$\mathbb{Z}_p$

4 Cohomology of varieties in characteristic p

In this section we see how p-adic numbers can be used to define cohomology theories for affine algebraic varieties over $\mathbb{F}_p$.
This section is a bit more technical so we assume that the reader is familiar with basic algebraic geometry (the content of [Har77] for example), and with rigid geometry (like the one explained in [Ber90]). Anyway we will try to keep the exposition as much down to earth as possible.

Definition 4.1 An *affine algebraic variety* X over K is the zero locus of a polynomial system in K^d, for $d \in \mathbb{N}$.

Namely, if $m \in \mathbb{N}$ and if $f_1, \dots, f_m \in K[x_1, \dots, x_d]$, we can define an algebraic variety X as the zero set of the system

$$\begin{cases} f_1(x_1, \dots, x_d) = 0 \\ \qquad\qquad\qquad \vdots \\ f_m(x_1, \dots, x_d) = 0. \end{cases}$$

The system defining the variety corresponds to the ringed space

$$\mathrm{Spec}(K[x_1, \dots, x_d]/(f_1, \dots, f_m)), \tag{3}$$

The space (3) is constructed in a similar way as the space $\mathrm{MaxSpec}(\mathbb{Z})$ of section 3, as the space of prime ideals of $K[x_1, \dots, x_d]/(f_1, \dots, f_m)$, with ring of functions given by $K[x_1, \dots, x_d]/(f_1, \dots, f_m)$. We will identify X with such object.

Example 4.2 The affine variety $\mathrm{Spec}(\mathbb{F}_p[x]/(x))$ is the point corresponding to the origin in $\mathbf{A}^1_{\mathbb{F}_p}$.

By a cohomology theory on X we mean a sequence of algebraic invariants $\{\mathrm{H}^i(X)\}_{i\in\mathbb{N}}$, usually vector spaces, associated to X.
The construction of a cohomology theory cannot be accomplished via the usual methods of algebraic topology, because
$\mathrm{Spec}(K[x_1, \dots, x_d]/(f_1, \dots, f_m))$ is totally disconnected. What one can do is defining de Rham cohomology for X.

4.1 Algebraic de Rham cohomology

One can define a de Rham complex $\Omega^\bullet_{X/K}$ as in [Sta24, Tag 07HX]. This is done as in the differential geometry setting, starting from the module of differential 1-forms $\Omega^1_{X/K}$. Namely $\Omega^1_{X/K}$ is the set whose elements are formal expressions of the type

$$\sum_{i=1\dots,d} g_i \mathrm{d}x_i,$$

where g_i is a function on X, for $i = 1, \dots, d$.

Example 4.3 The module of differential forms of degree one over $\mathbf{A}^1_{\mathbb{F}_p}$ is $\mathbb{F}_p[x]\mathrm{d}x$, i.e. the free module of rank 1 over $\mathbb{F}_p[x]$ generated by the differential $\mathrm{d}x$. The de Rham complex for $\mathbf{A}^1_{\mathbb{F}_p}$ is then

$$0 \to \mathbb{F}_p[x] \xrightarrow{\partial} \mathbb{F}_p[x]dx \to 0,$$

where ∂ is the differential defined by $\partial(f) = \frac{\mathrm{d}}{\mathrm{d}x} f \mathrm{d}x$, for $f \in \mathbb{F}_p[x]$.

Definition 4.4 Let X be an affine algebraic variety over K. The *de Rham cohomology* of X, denoted by $\{\mathrm{H}^i_{\mathrm{dR}}(X)\}_{i\in\mathbb{N}}$, is defined to be the cohomology of the de Rham complex of X. Explicitly

$$\mathrm{H}^i_{\mathrm{dR}}(X) = \mathrm{H}^i(\Omega^\bullet_{X/\mathbb{F}_p}).$$

But this is not a satisfactory theory in characteristic p. In fact if X is a variety over $\mathbb{F}_p$, the cohomology groups $\mathrm{H_{dR}}^i(X)$ are all torsion, being vector spaces over $\mathbb{F}_p$. This implies in particular that de Rham cohomology for varieties over $\mathbb{F}_p$ is not a Weil cohomology theory (see [Sta24, Tag 0FGS], for the definition of a Weil cohomology theory). This basically means that de Rham cohomology is not suitable for doing arithmetic computations in characteristic p.
Moreover de Rham cohomology behaves in an unpleasant way in characteristic p, as the following example illustrates.

Example 4.5 One has that $\mathrm{H}^1_{\mathrm{dR}}(\mathbf{A}^1_{\mathbb{F}_p}) \neq 0$, because one cannot integrate differential forms like $x^{p-1}\mathrm{d}x$ (since $p = 0$). But one would expect the first cohomology group of an affine line to be 0 as in topology.

To solve this problem one could then try to "lift" varieties over $\mathbb{F}_p$ to the infinitesimal thickening of $\mathbb{F}_p$, i.e. $\mathbb{Z}_p$, because $\mathbb{Z}_p$ lives in characteristic 0.

4.2 Rigid convergent cohomology

Let $X = \mathrm{Spec}(\mathbb{F}_p[x_1, \dots, x_d]/(f_1, \dots, f_m))$ be an affine algebraic variety over $\mathbb{F}_p$. To lift this variety to characteristic 0 one can choose polynomials $h_1, \dots, h_s$ in $\mathbb{Z}_p[x_1, \dots, x_d]$ such that $h_i = f_i \pmod p$ for $i = 1, \dots, d$. Then define the variety $\mathcal{X} = \mathrm{Spec}(\mathbb{Q}_p[x_1, \dots, x_d]/(h_1, \dots, h_m))$. Note that the points of $\mathcal{X}$ in ${\mathbb{Z}_p}^d$ are sended to the points of X in ${\mathbb{F}_p}^d$ along the projection

$$ {\mathbb{Z}_p}^d \to {\mathbb{F}_p}^d. \tag{4}$$

Thus $\mathcal{X}$ is a kind of infinitesimal thickening of X in characteristic 0. So it can be seen as an infinitesimal "tube" around X that lives in characteristic 0. In particular it has the same "shape" as X, and so we can expect it to have the same geometric properties of X, in some sense.

Example 4.6 For the variety $X = \mathrm{Spec}(\mathbb{F}_p[x]/(x))$, we can choose the lift $x \in \mathbb{Z}_p[x]$ of $x \in \mathbb{F}_p$, and so $\mathcal{X}$ would be the origin in $\mathbf{A}^1_{\mathbb{Q}_p}$.

So one could try to define cohomology groups $\{\mathrm{H}^i_{\mathrm{lift}}(X)\}_{i\in\mathbb{N}}$ on X as

$$\mathrm{H}^i_{\mathrm{lift}}(X) = \mathrm{H}^i_{\mathrm{dR}}(\mathcal{X}), \tag{5}$$

for $i \in \mathbb{N}$.
Unfortunately this construction does not work, because the cohomology groups $\{\mathrm{H}^i_{\mathrm{lift}}(X)\}_{i\in\mathbb{N}}$ depend on the choice of the h_is in general (but not if $\mathcal{X}$ is a smooth proper -so not affine- scheme over $\mathbb{Z}_p$, see [BO83, Corollary 2.5].
This problem is solved enlarging this infinitesimal tube, using rigid geometry. Let $X \hookrightarrow P$ be an embedding of X into an affine smooth formal scheme over $\mathbb{Z}_p$, and let $t_1, \dots, t_m$ be respective lifts of $f_1, \dots, f_m$ in P. Then we define the *tube* of X in P as the set

$$\{p \in P_{\mathbb{Q}_p} : |t_i(x)| < 1,\ i = 1, \dots, d\}, \tag{6}$$

where $P_{\mathbb{Q}_p}$ is the generic fiber of P. It is denoted by $]X[_P$. This is an actual tube as one can see in the following example.

Example 4.7 The tube of $\mathrm{Spec}(\mathbb{F}_p[x]/(x))$ in $\mathbf{A}^1_{\mathbb{Q}_p}$ is the open unit disk centered in the origin: $\mathbf{B}^1(0, 1^-)$.

The tube for a more general X is formed attaching to X open unit disks centered in the points of X.

Using this rigid analytic tube one can define the convergent rigid cohomology on X.

Definition 4.8 The *rigid convergent cohomology* groups of X are defined as

$$\mathrm{H}^i_{\mathrm{rig}}(X) = \mathrm{H}^i_{\mathrm{dR}}(]X[_P),$$

for $i = 1, \dots, n$.

It is indeed true that this definition does not depend on the embedding $X \hookrightarrow P$ (this basically follows from [LS07, Corollary 2.3.16]).

Example 4.9 The cohomology group $\mathrm{H}^1_{\mathrm{rig}}(\mathrm{Spec}(\mathbb{F}_p[x]/(x)))$ is equal to 0.
In fact the functions on $\mathbf{B}^1(0, 1^-)$ are the convergent series

$$\mathbb{Q}_p\{\{x\}\} = \left\{ \sum_{i\in\mathbb{N}} a_i x^i : lim_{i\to\infty} |a_i|_p \varepsilon^i = 0, \text{ for some } \varepsilon \in \mathbb{R} \right\}.$$

So the de Rham complex of $\mathbf{B}^1(0, 1^-)$ is

$$0 \to \mathbb{Q}_p\{\{x\}\} \xrightarrow{\partial} \mathbb{Q}_p\{\{x\}\}dx \to 0,$$

where ∂ is the differential defined by $\partial(f) = \frac{\mathrm{d}}{\mathrm{d}x} f \mathrm{d}x$, for $f \in \mathbb{Q}_p\{\{x\}\}$. But the convergent series can be integrated, and so $\mathrm{H}^1_{\mathrm{dR}}(\mathbf{B}^1(0, 1^-)) = 0$.

In forthcoming work in collaboration with Federico Bambozzi and Bruno Chiarellotto we will define a derived version of convergent rigid cohomology, for which the functions on the disks forming the tubes will be the tempered functions. These are defined for $\mathbb{Q}_p$ as

$$\mathbb{Q}_p[\![x]\!]_{\text{temp}} = \left\{ \sum_{i\in\mathbb{N}} a_i x^i : lim_{i\to\infty} |a_i|_p (i+1)^{-n} = 0 \text{ for some } n \in \mathbb{N} \right\}.$$

These kind of functions do not correspond to an open set of $\mathbf{A}^1_{\mathbb{Q}_p}$ as a rigid analytic space. But they are related to an open disk of the *derived analytic space* associated to $\mathbf{A}^1_{\mathbb{Q}_p}$ in the sense of [BBB16, Subsection 2.4].

References

[BBB16] Federico Bambozzi and Oren Ben-Bassat, *Dagger geometry as Banach algebraic geometry.* J. Number Theory 162 (2016), 391–462.

[Ber90] Vladimir G. Berkovich, "Spectral theory and analytic geometry over non-Archimedean fields". volume 33 of Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 1990.

[BO83] P. Berthelot and A. Ogus, *F-isocrystals and de Rham cohomology. I.* Invent. Math. 72/2 (1983), 159–199.

[Gou20] Fernando Q. Gouvêa, "p-adic numbers". Universitext. Springer, Cham, third edition, 2020. An introduction.

[Har77] Robin Hartshorne, "Algebraic geometry". Volume No. 52 of Graduate Texts in Mathematics. Springer-Verlag, New York-Heidelberg, 1977.

[Kob84] Neal Koblitz, "p-adic numbers, p-adic analysis, and zeta-functions". Volume 58 of Graduate Texts in Mathematics. Springer-Verlag, New York, second edition, 1984.

[Lan02] Serge Lang, "Algebra". Volume 211 of Graduate Texts in Mathematics. Springer-Verlag, New York, third edition, 2002.

[LS07] Bernard Le Stum, "Rigid cohomology". Volume 172 of Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, 2007.

[Ser79] Jean-Pierre Serre, "Local fields". Volume 67 of Graduate Texts in Mathematics. Springer-Verlag, New York-Berlin, 1979. Translated from the French by Marvin Jay Greenberg.

[Sta24] The Stacks project authors, *The stacks project.* `https://stacks.math.columbia.edu` (2024).

A sphere rolling on a plane: a journey into nonholonomic mechanics

Mariana Costa Villegas [(*)]

1 Introduction to mechanical systems

Lagrangian mechanics describes motion in a mechanical system by means of the configuration space. A Lagrangian mechanical system is given by a manifold (the configuration space) and a function on its tangent bundle (the Lagrangian).

Definition 1.1 The *configuration space* of a mechanical system is an n-dimensional smooth manifold Q whose coordinates $q^1, ..., q^n$ specify the configuration of the system.

The tangent bundle TQ is the space of positions and velocities and is $2n$-dimensional. Some examples of configuration manifolds for classical mechanical systems are:

(a) pendulum: $Q = S^1$

(b) double pendulum: $Q = S^1 \times S^1$

(c) rigid body: $Q = \mathrm{SO}(3)$

The Lagrangian $L : TQ \to \mathbb{R}$ is a function on the tangent bundle of the configuration manifold given by the kinetic minus the potential energy of the system. The motion of the system is described by the Euler-Lagrange equations

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}^i} - \frac{\partial L}{\partial q^i} = 0, \qquad i = 1, ..., n.$$

The Euler-Lagrange equations are equivalent to a variational principle on a space of smooth paths called Hamilton's principle. This principle states that a path is a solution of the Euler- Lagrange equations if and only if it is a stationary point of an action functional. This fact has many important consequences in the study of Lagrangian and Hamiltonian systems.

[(*)]Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: `mariana.costavillegas@math.unipd.it` . Seminar held on 10 April 2024.

1.1 Constraints

Holonomic constraints. Holonomic systems are mechanical systems that are subject to constraints which limit their possible configurations. Such constraints, when given as constraints on the velocity, may be integrated and expressed as constrains on the configuration variables. An example of an holonomic system is the pendulum, constrained by the length of the string.

Nonholonomic constraints. Nonholonomic systems are systems with nonintegrable constraints on their velocities. The constraints are given as constraints on the velocities and cannot be expressed as constraints on the configuration variables. Essentially, nonholonomic constraints restrict types of motion but not position.

An example of a nonholonomic system is the Chaplygin sleigh which consists of a rigid body (a sleigh) in the plane supported by three points, two of which slide freely without friction while the third is a knife edge. The knife does not allow motion perpendicular to its edge but can slide in the direction parallel to it.

Since they restrict the possible velocities of the system, nonholonomic constraints are described by nonintegrable distributions on the configuration space Q. We recall the definition of integrable distributions below.

Definition 1.2 A distribution $\mathcal{D}$ on Q is an assignment $q \mapsto \mathcal{D}_Q$ where $\mathcal{D}_q$ is a subspace of T_qQ such that around every point $q_0 \in Q$ there is a neighbourhood U of q_0 such that for all $q \in U$ $\mathcal{D}_q = \text{span}\{X_1(q), ...X_k(q)\}$ for smooth vector fields $X_1, ...X_k$.

Definition 1.3 A distribution is integrable if there exists an integral manifold passing through each $q \in Q$.

We remark that if the distribution is integrable, then the constraints are holonomic.

2 Linear nonholonomic systems

Nonholonomic systems are a generalization of classical Lagrangian and Hamiltonian systems in which one allows nonholonomic constraints. A nonholonomic system with linear constraints is determined by

- A configuration space Q: an n-dimensional manifold
- A Lagrangian: a function on the tangent space of the configuration manifold $L : TQ \to \mathbb{R}$ given by $L = T - U$ kinetic minus potential energy
- A constraint distribution $\mathcal{D} \subset TQ$: a non integrable regular linear distribution of rank $r < n$ defined by

$$\sum_{k=1}^{n} \beta_k^a(q)\dot{q}^k = 0, \qquad a = 1, ..., r$$

Lagrange D'Alembert principle. The equations of motion are given by

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) - \frac{\partial L}{\partial q^i} = R_i, \qquad i = 1, ...n,$$

where R_i are the components of the reaction force. In order to obtain an expression for R_i, we need to use the Lagrange D'Alembert principle which assumes that the reaction force R annihilates any possible displacement of the system. Namely, if $\dot{q}$ satisfies the constraints, then $R_i \dot{q}^i = 0$.

Energy. If we define $E(q, \dot{q}) = \dot{q}^i \frac{\partial L}{\partial \dot{q}^i} - L$, then if the constraints are satisfied, using $R_i \dot{q}^i = 0$, it can be shown that E is a constant of motion. So in nonholonomic systems with linear constraints the energy is preserved.

Remarks. In nonholonomic systems there is no variational principle and there is not a Hamiltonian formulation. This gives rise to important differences between Hamiltonian and nonholonomic mechanics. Whereas in Hamiltonian mechanics, Noether's theorem gives a way to link symmetries and first integrals, the mechanisms that relate symmetries with conserved quantities in nonholonomic systems are not completely understood. For instance, momentum is not always preserved for systems with symmetries in the nonholonomic setting. Furthermore, unlike Hamiltonian systems, nonholonomic systems need not preserve volume in the phase space. The Chaplygin sleigh is a classical example of a nonholonomic system which exhibits dissipation.

2.1 Example: A sphere rolling without slipping on a plane

As a classical example of a nonholonomic system, let us consider the motion of a sphere of radius r and mass m rolling without slipping on a horizontal plane.

The system. To study the problem, we fix a spatial frame $\Sigma_s = \{O; \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ such that the horizontal plane Π contains the origin O and is spanned by the vectors $\mathbf{e}_1, \mathbf{e}_2$. We also fix a body frame $\Sigma_b = \{C; \mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3\}$ whose origin is the center of mass C of the sphere and such that the vectors $\mathbf{E}_i$ are aligned with its principal axes of inertia.

We define the following points and vectors:

- C' is the geometric center of the sphere
- C is the center of mass
- O is the origin of the space frame
- P is the contact point
- $\boldsymbol{u} \in \mathbb{R}^3$ are the coordinates of the vector $\overrightarrow{OC}$, connecting the origin of the spatial frame and the center of mass, with respect to the spatial frame Σ_s
- $\boldsymbol{x} \in \mathbb{R}^3$ are the coordinates of the vector $\overrightarrow{OP}$, connecting the origin of the spatial frame and the contact point, with respect to the spatial frame Σ_s

- $\boldsymbol{\rho} \in \mathbb{R}^3$ are the coordinates of the vector $\overrightarrow{CP}$, connecting the center of mass and the contact point, with respect to the body frame Σ_b
- $\boldsymbol{\gamma} = B^{-1}\mathbf{e}_3 \in \mathbb{R}^3$ are the coordinates of the unitary vector normal to the plane at the contact point P, with respect to the body frame Σ_b
- $\omega \in \mathbb{R}^3$ are the coordinates of the angular velocity vector with respect to the spatial frame Σ_s
- $\boldsymbol{\Omega} \in \mathbb{R}^3$ are the coordinates of the angular velocity vector with respect to the body frame Σ_b

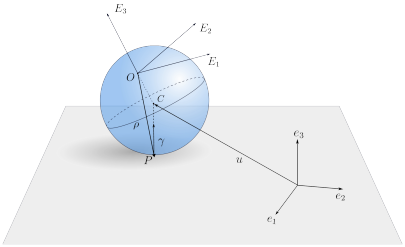


Figure 1

The configuration manifold of the system is $Q = \mathrm{SO}(3) \times \mathbb{R}^2$ and a configuration is determined by the pair $(\boldsymbol{x}, B) \in Q$ where since $\boldsymbol{x} \in \Pi$, we may consider $\boldsymbol{x} = (x_1, x_2, 0) \in \mathbb{R}^2 \times \{0\}$, and the attitude matrix $B \in \mathrm{SO}(3)$ determines the orientation of the body (i.e. it is the change of basis matrix between the bases $\{\boldsymbol{e}_i\}$ and $\{\boldsymbol{E}_i\}$ of $\mathbb{R}^3$). We recall (see [19]) that the space and body coordinate representations of the angular velocity are defined by the left and right trivializations:

$$B^{-1}\dot{B} = \hat{\boldsymbol{\Omega}}, \qquad \dot{B}B^{-1} = \hat{\boldsymbol{\omega}},$$

where, for $\boldsymbol{a} \in \mathbb{R}^3$, the notation $\hat{\boldsymbol{a}}$ stands for the unique 3×3 skew-symmetric real matrix such that $\hat{\boldsymbol{a}}\boldsymbol{b} = \boldsymbol{a} \times \boldsymbol{b}$ for all $\boldsymbol{b} \in \mathbb{R}^3$, where $\times$ is the cross product in $\mathbb{R}^3$. It is well-known that the mapping $\hat{}\colon (\mathbb{R}^3, \times) \to \mathfrak{so}(3)$ is a Lie algebra isomorphism. From the definitions of the vectors $\boldsymbol{x}, \boldsymbol{u}$ and $\boldsymbol{\rho}$, we have that

$$\boldsymbol{x} = \boldsymbol{u} - r\mathbf{e}_3.$$

The tangent bundle of the configuration manifold has dimension 10 and is described by $TQ = \mathrm{SO}(3) \times \mathbb{R}^2 \times \mathbb{R}^3 \times \mathbb{R}^2$, where we identify $T\mathrm{SO}(3) \equiv \mathbb{R}^3$ by considering the left trivialization of the Lie algebra of $\mathrm{SO}(3)$. A point on TQ is given by $(B, \boldsymbol{x}, \boldsymbol{\Omega}, \dot{\boldsymbol{x}})$.

The constraint of rolling without slipping is obtained by making the velocity at the contact point equal to zero and is described by

$$\dot{\boldsymbol{u}} = B(\boldsymbol{\rho} \times \boldsymbol{\Omega}).$$

It can be seen that this constraint defines two independent nonholonomic constraints. Therefore, the phase space has dimension 8.

Reduction. The system has an SE(2)-symmetry that corresponds to translations and rotations of the plane Π. It can be checked that the Lagrangian L and the distribution $\mathcal{D}$ of the system are invariant under the lift of the action of SE(2) on Q so the equations of motion can be reduced by this symmetry. The reduced phase space has dimension 5.

Equations of motion. Using the Lagrange D'Alembert principle, we can find that the reduced equations of motion are

$$\begin{aligned}\dot{\boldsymbol{M}} &= \boldsymbol{M} \times \boldsymbol{\Omega} + m\dot{\rho} \times (\boldsymbol{\Omega} \times \boldsymbol{\rho}) + mg\boldsymbol{\rho} \times \boldsymbol{\gamma} \\ \dot{\boldsymbol{\gamma}} &= \boldsymbol{\Omega} \times \boldsymbol{\gamma},\end{aligned}$$

where $\boldsymbol{M} = \mathbb{I}\boldsymbol{\Omega} + m\boldsymbol{\rho} \times (\boldsymbol{\Omega} \times \boldsymbol{\rho})$, $\mathbb{I} = \operatorname{diag}(I_1, I_2, I_3)$ is the inertia tensor and $\|\boldsymbol{\gamma}\|^2 = 1$.

Energy. The system has the energy first integral

$$E = \frac{1}{2}\langle \boldsymbol{M}, \boldsymbol{\Omega} \rangle - mg\langle \boldsymbol{\rho}, \boldsymbol{\gamma} \rangle.$$

Integrable cases. To prove integrability of the system, according to the Euler-Jacobi theorem, we would need two additional first integrals and a smooth invariant measure. In the general case, without additional symmetries, there is no invariant measure and no additional first integrals. The system is generally chaotic, however, there are some known integrable cases. The case where the sphere is axially symmetric, known as Routh's sphere, was studied by Routh in 1884 and is known to be integrable [20]. The case of the dynamically balanced sphere, where the center of mass corresponds to the geometric center of the sphere, is known as the Chaplygin sphere and was proven to be integrable by Chaplygin in 1903 [9].

A particular case of those two integrable cases, is the homogeneous sphere, a dynamically balanced sphere with equal moments of inertia, $I_1 = I_2 = I_3$. This system is integrable and it can be shown that the homogeneous sphere moves in a straight line on the plane.

Variations. We may consider variations of this problem. For example, we may consider a sphere rolling on a plane which is rotating or a sphere with a rotating shell rolling on a fixed plane. This type of systems are nonholonomic, but the constraints are not linear in the velocities but affine. There are some differences between linear and affine nonholonomic systems which we will discuss in the following section.

3 Affine nonholonomic systems

A nonholonomic system with affine constraints is determined by a configuration space, a Lagrangian and a constraint distribution. In this case, the distribution $\mathcal{A} \subset TQ$ is a non integrable regular affine distribution defined by

$$\mathcal{A} = \mathcal{D} + Z, \qquad \mathcal{M}_q = \mathcal{D}_q + Z(q),$$

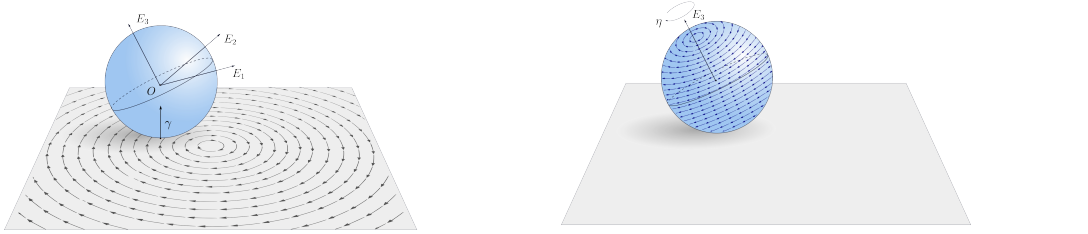
where $\mathcal{D}$ is the linear distribution and $Z \in \mathfrak{X}(Q)$ is a vector field on the configuration manifold.

Moving energy. Nonholonomic systems with affine constrains do not in general preserve the energy. However, as noticed in [13], if the affine terms correspond to the infinitesimal generator of a continuous symmetry of the Lagrangian, then a modification of the energy, which we term *moving energy* in accordance with [13, 11], arises as a first integral.

Remarks. More in general, the understanding of nonholonomic systems whose constraints are affine in the velocities, is much less developed than that of the linear nonholonomic constraints. There is still a lot to be understood on the existence of momentum type integrals and existence of invariant measure. Furthermore, there are less examples to illustrate and guide these investigations.

3.1 Examples

Sphere rolling on a rotating plane. We may consider the system of a homogeneous sphere rolling on a steadily rotating horizontal plane. This system has been largely considered in the literature, see [13] and the references therein. The nonholonomic constraint is determined by setting the velocity at the contact point to be the velocity of the plane at that point. This defines a nonholonomic affine constraint. The system may be reduced by the SO(3) symmetry of the sphere, to a system with a 5-dimensional phase space. It can be seen that the system is integrable and in fact, that the solutions of the reduced system are periodic. The reduced system possesses four independent first integrals, one of which is the moving energy. It can be shown that the trajectory of the sphere on the plane describes a circle.



(a) Sphere rolling on a rotating plane (b) Sphere with a rotating shell rolling on a fixed plane

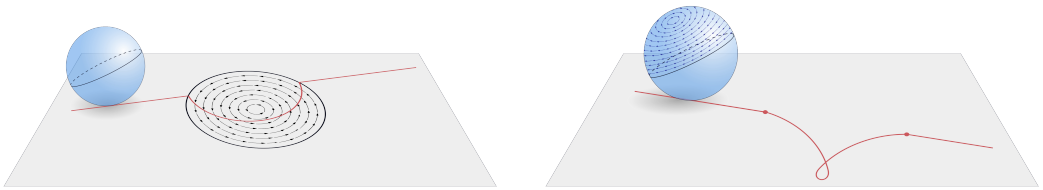
Figure 2

The sphere with a rotating shell rolling on a fixed plane. We consider an analogous problem. We consider a homogeneous sphere rolling on a fixed horizontal plane and we assume that the sphere has a thin massless shell which rotates steadily as shown in Figure 2b. The system may be reduced by the SE(2)-symmetry of the plane to a 5-dimensional phase space and it can be seen that the reduced system has three first integrals, one of which is the moving energy, and a smooth invariant measure. It is therefore also integrable.

3.2 Anais billiard phenomenon

We consider the following problem described in [18] and shown in Figure 3a. Throw an homogeneous ball rolling without slipping on a horizontal plane which has a rotating circular disc and suppose that the ball goes in and out of the rotating disc. As we have seen in Section 2.1, while the sphere is rolling on the fixed part of the plane, its trajectory will follow a straight line. Meanwhile, as we saw in Section 3.1, when the ball enters the rotating disc on the plane, it will follow a circular trajectory. When the ball goes out of the disc it will go back to a linear trajectory and it has been shown that this trajectory will be the exact prolongation of the initial one.

We consider an analogous problem shown in Figure 3b. Suppose that a sphere has a thin shell of negligible mass, and assume that half of the shell is fixed while the other half rotates steadily with respect to a reference frame attached to the center of the sphere. We have shown that an analogous phenomenon occurs: the sphere is rolling in a rectilinear trajectory until the rotating part of the shell comes in contact with the plane, then the trajectory changes but when the contact point comes out of the rotating part, it goes back to the exact initial rectilinear trajectory.



(a) (b)

Figure 3: Graphic representation of the Anais billiard phenomenon and its generalization.

4 Generalization

We now consider a convex rigid body with smooth surface $\mathcal{S}$ on the infinite horizontal plane Π subject to the following constraint. We assume that there are two given vector fields, $V \in \mathfrak{X}(\Pi)$ and $W \in \mathfrak{X}(\mathcal{S})$, which determine the velocity of the contact point. This constraint is a generalization of the nonholonomic constraint of rolling without slipping. The system is illustrated in Figure 4.

If both vector fields V and W vanish, we recover the problem of a convex body rolling without slipping on the plane. In particular, if the convex body is a sphere we recover the problem of a sphere rolling without slipping on a plane described in Section 2.1. Furthermore, with specific choices of V and W we may recover the examples described in Section 3.1.

Our motivation to consider the problem in its full generality (i.e. for arbitrary convex body and arbitrary vector fields V and W) is to illustrate dynamic phenomena that could guide the development of the theory for existence of invariant measures, existence

of first integrals, integrability and chaotic behaviour of mechanical systems with affine nonholonomic constraints.

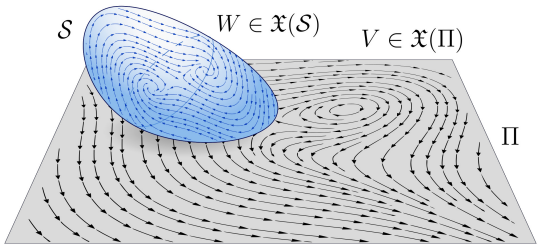


Figure 4: Graphic representation of the vector fields V on the plane Π and W on the surface $\mathcal{S}$ of the convex body. The nonholonomic constraint specifies that the velocity of the contact point equals the sum of both vector fields at that point.

The system. To study the system we consider the same definitions as in Section 2.1 and include the following observation that allows us to consider a more general shape of the body. In this case, the vector $\boldsymbol{u}$ and $\boldsymbol{x}$ are related by the equation

$$\boldsymbol{x} = \boldsymbol{u} + B\boldsymbol{\rho}. \tag{1}$$

Following the approach of [6], since the surface $\mathcal{S}$ of the body is smooth and convex, it guarantees that the Gauss map $\boldsymbol{n}_b$ is a diffeomorphism. We may use the Gauss map $\boldsymbol{n}_b : \mathcal{S} \to S^2 \subset \mathbb{R}^3$ of the surface of the body to obtain a functional relation between $\boldsymbol{\rho}$ and $\boldsymbol{\gamma}$:

$$\boldsymbol{n}_b(\boldsymbol{\rho}) = -\boldsymbol{\gamma}, \qquad \boldsymbol{\rho} = \boldsymbol{n}_b^{-1}(-\boldsymbol{\gamma}). \tag{2}$$

We think of the vector field $V \in \mathfrak{X}(\Pi)$ as the restriction to $\Pi \subset \mathbb{R}^3$ of a vector field on $\mathbb{R}^3$ which is tangent to Π. For this reason, for each $\boldsymbol{x} \in \Pi$, we will write $\boldsymbol{V}_s(\boldsymbol{x}) = (V_1(\boldsymbol{x}), V_2(\boldsymbol{x}), 0) \in \mathbb{R}^3$, as the coordinate expression of the vector field V with respect to the spatial frame Σ_s. Similarly, we think of the vector field $W \in \mathfrak{X}(\mathcal{S})$ as the restriction to $\mathcal{S} \subset \mathbb{R}^3$ of a vector field on $\mathbb{R}^3$ which is tangent to $\mathcal{S}$. For this reason, denoting by $\boldsymbol{X}$ the coordinates of vectors with respect to the body frame Σ_b, for each $\boldsymbol{X} \in \mathcal{S} \subset \mathbb{R}^3$, we will write $\boldsymbol{W}_b(\boldsymbol{X}) = (W_1(\boldsymbol{X}), W_2(\boldsymbol{X}), W_3(\boldsymbol{X})) \in \mathbb{R}^3$, where the tangency condition $\langle \boldsymbol{W}_b(\boldsymbol{X}), \boldsymbol{n}_b(\boldsymbol{X}) \rangle = 0$ holds for all $\boldsymbol{X} \in \mathcal{S} \subset \mathbb{R}^3$.

We emphasize that the coordinate expressions for the vector fields V and W are given in distinct reference frames. V is naturally written the space frame Σ_s whereas W is naturally written in the body frame Σ_b.

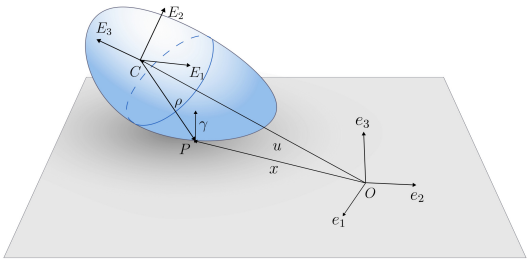


Figure 5: Graphic representation of the vectors $\boldsymbol{\rho}, \boldsymbol{\gamma} \in \mathbb{R}^3$ (which are written with respect to the body frame $\Sigma_b = \{C; \boldsymbol{E}_1, \boldsymbol{E}_2, \boldsymbol{E}_3\}$) and $\boldsymbol{x}, \boldsymbol{u} \in \mathbb{R}^3$ (which are written with respect to the spatial frame $\Sigma_s = \{O; \boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3\}$).

Constraints. The velocity of the material point in contact with the plane, written in the space frame Σ_s, is given by $\dot{\boldsymbol{u}} + B(\boldsymbol{\Omega} \times \boldsymbol{\rho})$. Therefore, the nonholonomic constraint is:

$$\dot{\boldsymbol{u}} = B(\boldsymbol{\rho} \times \boldsymbol{\Omega}) + \boldsymbol{V}_s(\boldsymbol{x}) + B\boldsymbol{W}_b(\boldsymbol{\rho}), \tag{3}$$

where $\boldsymbol{x}$ is expressed in terms of $\boldsymbol{u}$, B and $\boldsymbol{\rho}$ by (1). It can be seen that (3) defines two independent nonholonomic constraints. Therefore, the affine distribution $\mathcal{A}$ has dimension 8. We express $\mathcal{A} = \mathcal{D} + Z$ where $\mathcal{D} \subset TQ$ is the linear distribution and $Z \in \mathfrak{X}(Q)$ is a vector field. These can be taken as

$$\mathcal{D} = \{(\boldsymbol{u}, \dot{\boldsymbol{u}}, B, \boldsymbol{\Omega}) \in \mathbb{R}^3 \times \mathbb{R}^3 \times \mathrm{SO}(3) \times \mathbb{R}^3 \,:\, \dot{\boldsymbol{u}} = B(\boldsymbol{\rho} \times \boldsymbol{\Omega}) \text{ and } u_3 = -\langle \boldsymbol{\rho}, \boldsymbol{\gamma} \rangle\}, \tag{4a}$$

$$Z(\boldsymbol{u}, B) = (\boldsymbol{V}_s(\boldsymbol{x}) + B\boldsymbol{W}_b(\boldsymbol{\rho}), \boldsymbol{0}). \tag{4b}$$

Equations of motion. Using the Lagrange D'Alembert principle we get the equations of motion as the following.

Proposition 4.1 *The equations of motion of the problem are*

$$\dot{\boldsymbol{M}} = \boldsymbol{M} \times \boldsymbol{\Omega} + m\dot{\boldsymbol{\rho}} \times (\boldsymbol{\Omega} \times \boldsymbol{\rho}) + mg\boldsymbol{\rho} \times \boldsymbol{\gamma} + m(B^{-1}\boldsymbol{V}_s(\boldsymbol{x}) + \boldsymbol{W}_b(\boldsymbol{\rho})) \times (\dot{\boldsymbol{\rho}} + \boldsymbol{\Omega} \times \boldsymbol{\rho}), \tag{5a}$$

$$\dot{B} = B\hat{\boldsymbol{\Omega}}, \tag{5b}$$

$$\dot{\boldsymbol{u}} = B(\boldsymbol{\rho} \times \boldsymbol{\Omega}) + \boldsymbol{V}_s(\boldsymbol{x}) + B\boldsymbol{W}_b(\boldsymbol{\rho}), \tag{5c}$$

where $\boldsymbol{M} = \mathbb{I}\boldsymbol{\Omega} + m\boldsymbol{\rho} \times (\boldsymbol{\Omega} \times \boldsymbol{\rho} - B^{-1}\boldsymbol{V}_s(\boldsymbol{x}) - \boldsymbol{W}_b(\boldsymbol{\rho}))$ *and* $\boldsymbol{x} = \boldsymbol{u} + B\boldsymbol{\rho}$.

5 A dynamically balanced sphere

We consider the special case in which the surface of the convex body is spherical, with radius $r > 0$, and the center of mass coincides with the geometric center. If both V and W

vanish, we recover the classical Chaplygin ball problem [9] which is known to be integrable. Other cases considered previously for non-vanishing V, W are found in [2, 17, 3]. Here we consider the general case.

The relation (2) between $\boldsymbol{\rho}$ and $\boldsymbol{\gamma}$ is

$$\boldsymbol{\rho} = -r\boldsymbol{\gamma}, \tag{6}$$

and (1) becomes

$$\boldsymbol{x} = \boldsymbol{u} - r\boldsymbol{e}_3. \tag{7}$$

In view of (6), we have $\boldsymbol{\gamma} \times \boldsymbol{\rho} = 0$ and $\dot{\boldsymbol{\rho}} = \boldsymbol{\rho} \times \boldsymbol{\Omega}$, so equation (5a) simplifies to

$$\dot{\boldsymbol{M}} = \boldsymbol{M} \times \boldsymbol{\Omega} \tag{8}$$

where in this case $\boldsymbol{M} = \mathbb{I}\boldsymbol{\Omega} + mr^2\boldsymbol{\gamma} \times (\boldsymbol{\Omega} \times \boldsymbol{\gamma}) + mr\boldsymbol{\gamma} \times (B^{-1}\boldsymbol{V}_s(\boldsymbol{x}) + \boldsymbol{W}_b(\boldsymbol{\rho}))$. This simplification implies that the vector $\boldsymbol{M}$, as seen in the spatial frame Σ_s is constant. As a consequence, we have.

Proposition 5.1 *For any $V \in \mathfrak{X}(\Pi)$ and $W \in \mathfrak{X}(\mathcal{S})$, the system has first integrals*

$$\langle \boldsymbol{M}, \boldsymbol{\alpha} \rangle, \qquad \langle \boldsymbol{M}, \boldsymbol{\beta} \rangle \qquad \textit{and} \qquad \langle \boldsymbol{M}, \boldsymbol{\gamma} \rangle,$$

where the Poisson vectors $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ are the rows of the attitude matrix $B \in \mathrm{SO}(3)$.

5.1 The case $V = 0$

When $V = 0$ the system has an SE(2)-symmetry and we can consider the reduced system. The reduced equations of motion are

$$\dot{\boldsymbol{M}} = \boldsymbol{M} \times \boldsymbol{\Omega}, \qquad \dot{\boldsymbol{\gamma}} = \boldsymbol{\gamma} \times \boldsymbol{\Omega}, \tag{9}$$

with $\boldsymbol{M} = \mathbb{I}\boldsymbol{\Omega} + mr^2\boldsymbol{\gamma} \times (\boldsymbol{\Omega} \times \boldsymbol{\gamma}) + mr\boldsymbol{\gamma} \times \boldsymbol{W}_b(\boldsymbol{\rho})$. As a consequence of Proposition 5.1, the reduced system (9) has first integrals

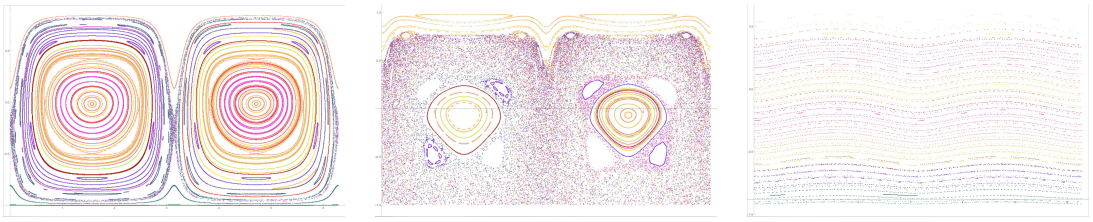
$$\|\boldsymbol{M}\|^2, \qquad \langle \boldsymbol{M}, \boldsymbol{\gamma} \rangle \qquad \text{and} \qquad \|\boldsymbol{\gamma}\|^2 = 1. \tag{10}$$

These first integrals are insufficient to conclude integrability of (9) using the Jacobi last multiplier theorem [1], which would require existence of an additional independent first integral and a smooth invariant measure.

5.1.1 Example: Chaplygin sphere with a rotating shell

Below we treat the simplest non-zero choice of $W \in \mathfrak{X}(\mathcal{S})$, corresponding to the uniform rotation of a light shell around a principal axis of inertia, which we assume to be the third one. The corresponding form of $\boldsymbol{W}_b$ is given by $\boldsymbol{W}_b(\boldsymbol{\rho}) = -r\sigma\boldsymbol{\gamma} \times \boldsymbol{E}_3$.

Poincaré map If $\boldsymbol{M}$ and $\boldsymbol{\gamma}$ are not parallel, the first integrals (10) are independent and their level sets are 3-dimensional submanifolds of the phase space $\mathbb{R}^3 \times S^2$. The dynamics can be numerically investigated using a 2-dimensional Poincaré map. Below we present some numerical experiments assuming $\langle \boldsymbol{M}, \boldsymbol{\gamma} \rangle = 0$ which lead us to conjecture that the dynamics is chaotic.



(a) (b) (c)

Figure 6: Poincaré map for the dynamically balanced sphere with a rotating shell at $\langle M, \gamma \rangle = 0$ with $I_1 = 0.5$, $I_2 = 2.5$, $I_3 = 3$, $m = 1$, $r = 5$, $\sigma = 10$.

Limit cases of the dynamics The numerical experiments in Figure 6 suggest that the dynamics is approximately integrable when the non-dimensional parameter ε is taken sufficiently large or small. Actually, we may prove that when $\varepsilon \to \infty$ the system is the classical Chaplygin sphere rolling on a plane which is known to be integrable [9] and when $\varepsilon \to 0$ the system has an additional first integral and a smooth invariant measure, it is therefore integrable in virtue of Jacobi's last multiplier theorem [1].

5.2 The case $W = 0$

Under the assumption that the vector field $\boldsymbol{V}_s$ is divergence free, the system possesses an invariant measure.

Proposition 5.2 *Suppose* $\operatorname{div}_{\mathbb{R}^2} \boldsymbol{V}_s = 0$. *Then*

$$\frac{1}{\sqrt{1 - mr^2 \langle \boldsymbol{\gamma}, A\boldsymbol{\gamma} \rangle}} \, d\boldsymbol{M} \, d\boldsymbol{u} \, d\boldsymbol{\alpha} \, d\boldsymbol{\beta} \, d\boldsymbol{\gamma}$$

is an invariant measure.

Assuming distinct moments of inertia, I_j, and non-zero $\boldsymbol{V}_s$, we do not expect additional first integrals and we expect the dynamics to be chaotic. We mention that [3] and [17] describe examples of this system for particular vector fields $V \in \mathfrak{X}(\Pi)$. In [3] the authors considered a rotating plane, while [17] considers a horizontally vibrating plane. In both cases, based on numerical explorations, those references concluded that the dynamics is chaotic.

6 A homogeneous sphere

We now assume that our convex body is a homogeneous sphere which puts us in the framework of Section 5 with the additional hypothesis of equal moments of inertia $I := I_1 = I_2 = I_3$. The equations of motion (5) may be rewritten as

$$\begin{aligned} &\dot{\boldsymbol{M}} = \boldsymbol{M} \times \boldsymbol{\Omega}, \qquad \dot{\boldsymbol{\alpha}} = \boldsymbol{\alpha} \times \boldsymbol{\Omega}, \qquad \dot{\boldsymbol{\beta}} = \boldsymbol{\beta} \times \boldsymbol{\Omega}, \qquad \dot{\boldsymbol{\gamma}} = \boldsymbol{\gamma} \times \boldsymbol{\Omega}, \\ &\dot{\boldsymbol{u}} = -rB(\boldsymbol{\gamma} \times \boldsymbol{\Omega}) + \boldsymbol{V}_s(u) + B\boldsymbol{W}_b(\boldsymbol{\gamma}), \end{aligned} \tag{11}$$

where the Poisson vectors $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ are the rows of the attitude matrix $B \in \mathrm{SO}(3)$ and we have used equations (6) and (7) to write $\boldsymbol{V}_s$ and $\boldsymbol{W}_b$ as functions of $\boldsymbol{u}$ and $\boldsymbol{\gamma}$.

The following proposition gives sufficient conditions for $\boldsymbol{V}_s$ and $\boldsymbol{W}_b$ to guarantee the existence of an invariant measure whose form coincides with the one of the linear system.

Proposition 6.1 *Suppose that* $\mathrm{div}_{\mathbb{R}^2}\boldsymbol{V}_s(\boldsymbol{x})$ *and* $\mathrm{div}_{S^2}\boldsymbol{W}(\boldsymbol{\gamma})$ *identically vanish, then the system (11) possesses the invariant measure* $d\boldsymbol{M}\, d\boldsymbol{u}\, d\boldsymbol{\alpha}\, d\boldsymbol{\beta}\, d\boldsymbol{\gamma}$.

References

[1] Arnol'd, V.I., Kozlov, V.V., Neishtadt, A., "Mathematical Aspects of Classical and Celestial Mechanics". Dynamical Systems, III. Encyclopaedia Math. Sci., vol. 3, Berlin: Springer, 1993.

[2] Bizyaev, I.A., Borisov, A.V., Mamaev, I.S., *Different models of rolling for a robot ball on a plane as a generalization of the Chaplygin ball problem.* Regul. Chaotic Dyn. 24 , 560–582 (2019).

[3] Bizyaev, I.A., Borisov, A.V., Mamaev, I.S., *Dynamics of the Chaplygin ball on a rotating plane.*. Russ. J. Math. Phys. 25, 423–433 (2018).

[4] Bloch, Anthony M., Krishnaprasad, P. S., Marsden, Jerrold E., Murray, Richard M., *Nonholonomic mechanical systems with symmetry.* Arch. Rational Mech. Anal. 136, 21–99 (1996).

[5] Bloch, A.M., "Nonholonomic Mechanics and Controls". Interdisciplinary Applied Mathematics, vol. 24, Systems and Control, New-York: Springer Verlag, 2003.

[6] Borisov, A.V., Mamaev, I.S., *The rolling motion of a rigid body on a plane and a sphere. Hierarchy of dynamics.* Regul. Chaotic Dyn. 7, 177–200 (2002).

[7] Borisov, A.V., Mamaev, I.S., Kilin, A.A., *Rolling of a ball on a surface. New integrals and hierarchy of dynamics.* Regul. Chaotic Dyn. 7, 201–219 (2002).

[8] Borisov, A.V., Mamaev, I.S. and Bizyaev, I.A., *The hierarchy of dynamics of a rigid body rolling without slipping and spinning on a plane and a sphere.* Regul. Chaotic Dyn. 18, 277–328 (2013).

[9] Chaplygin, S.A., *On a motion of a heavy body of revolution on a horizontal plane.* Reg. Chaotic Dyn. 7, 131–148 (2002) [original paper in Mathematical Collection of the Moscow Mathematical Society 24, 139–168 (1903)].

[10] Dalla Via, M., Fassò, F., Sansonetto, N., *On the dynamics of a heavy symmetric ball that rolls without sliding on a uniformly rotating surface of revolution.* J. Nonlinear Sci. 32, 84 (2022).

[11] Fassò, F., García-Naranjo, L.C., Sansonetto, N., *Moving energies as first integrals of nonholonomic systems with affine constraints.* Nonlinearity 31, 755–782 (2018).

[12] Fassò, F., Giacobbe, A., Sansonetto, N., *Gauge conservation laws and the momentum equation in nonholonomic mechanics.* Rep. Math. Phys. 62, 345–367 (2008).

[13] Fassò, F., Sansonetto, N., *Conservation of "moving" energy in nonholonomic systems with affine constraints and integrability of spheres on rotating surfaces.* J. Nonlinear Sci. 26, 519–544 (2016).

[14] Fassò, F., Sansonetto, N., *An elemental overview of the nonholonomic Noether theorem.* Int. J. Geom. Methods Mod. Phys. 6, 1343–1355 (2009).

[15] Fass'o, F., Sansonetto, N., *Conservation of energy and momenta in nonholonomic systems with affine constraints.* Regul. Chaotic Dyn. 20, 449–462 (2015).

[16] García-Naranjo L.C., Montaldi J., *Gauge Momenta as Casimir functions of nonholonomic systems.* Arch. Rat. Mech. Anal. 228, 563–602 (2018).

[17] Kilin, A.A., Pivovarova, E.N., *A particular Integrable case in the nonautonomous problem of a Chaplygin sphere rolling on a vibrating plane.* GiornaleAnnoPagine.

[18] Levy-Leblond, J.M., *The ANAIS billiard table.* Eur. J. Phys. 7, 252 (1986).

[19] Marsden J.E., Ratiu T.S., "Introduction to Mechanics and Symmetry". Texts in Applied Mathematics vol. 17, New York: Springer, (1999).

[20] Routh, E.D., "Dynamics of a System of Rigid Bodies". Dover Publications, Inc., New York 1960 (7th ed., revised and enlarged).

[21] Tokieda, T., "Roll Models". The American Mathematical Monthly 120/3, 265–282 (2013).

Collective periodic behaviors in large-volume interacting particle systems

Elisa Marini [(*)]

Abstract. In these notes, we will give an overview of collective periodic behaviors in large systems of interacting components. Loosely speaking, such phenomena consist in nearly-periodic oscillations which characterize the long-time dynamics of some macroscopic quantity of the system, and which cannot be ascribed to any external periodic forcing applied to the system, nor to any oscillatory behavior of its components, but rather arise from the interaction among these latter. Although they are ubiquitous in real-world systems (they are observed for instance in neural networks, predator-prey dynamics, epidemiology), such behaviors are still poorly understood from a theoretical standpoint. We will present a toy model of interacting diffusions displaying collective oscillations. This will serve as an example of the mechanisms which may originate collective periodic behaviors and to give an idea of the mathematics involved in the rigorous study of such phenomena.

1 Introduction

There exist a number of real-world systems - natural or artificial - comprised of a large number of microscopic units (particles or individuals) which are able to self-organise and give rise to coherent collective dynamics which are only perceived on a macroscopic scale, that is, when one does not observe the single unit but the entire system. Local interactions among the microscopic units of the system might originate collective dynamics which are qualitatively different from those of the individual constituents of the system, and therefore are not predictable by studying these latter individually.

A very common and particularly important example of such dynamics are collective periodic behaviors. These occur when some macroscopic quantity (observable) of the system shows almost-periodic oscillations which are persistent in time, despite the fact that its individual components neither have any tendency to behave periodically on their own, nor are subject to any external periodic forcing.

Similar behaviors are observed in a variety of real-world systems, for instance, in biology, ecology, socioeconomics, neuroscience (see Figure 1, which provides an example of

[(*)]Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: elisa.marini@math.unipd.it . Seminar held on 24 April 2024.

collective periodicity observed in this last context).

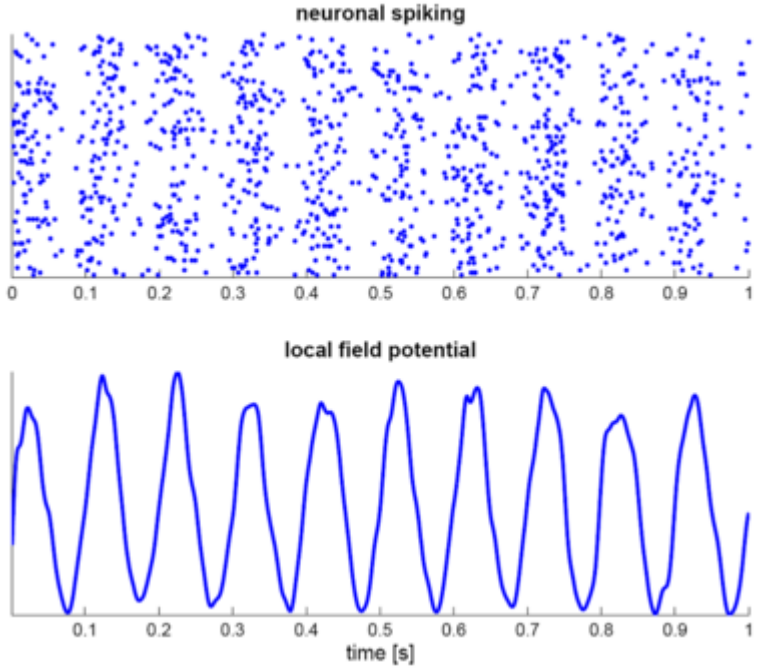


Figure 1: Rest-state neural activity: when an individual is not engaged in focused mental work, neural activity exhibits periodic oscillations. Roughly speaking, the state of each neuron can be described by a quantity called membrane potential and neurons interact by exchanging signals. Each time the membrane potential of a neuron grows beyond a certain threshold, that neuron emits a signal towards other neurons connected to it - we say that it fires, or emits a spike. This affects the states of the receiving neurons and, after a spike, the membrane potential of the neuron which has fired drops. Picture above: time on the x-axis and (labeled) neurons on the y-axis. A blue dot is present at (t, i) whenever neuron i fires at time t. Picture below: average activity of the system of neurons against time. Despite the absence of external stimuli and the fact the individual neurons do not have periodic activity, neurons tend to synchronize.

However, their origin is still poorly understood from a theoretical standpoint.

From a mathematical perspective, all such systems are naturally modeled by considering large families consisting of many interacting components (particles, individuals, ...). One reasonably starts by prescribing each component's dynamics: roughly speaking, the state x_i of the i-th component of the system will be described by a dynamical system, possibly with some random term, of the kind

$$\dot{x}_i(t) = \dot{x}_i(t) = f_i(x_i(t)) + \text{noise}_i \qquad\qquad i = 1, \dots, N$$

where here and in the sequel N will denote the number of components of the system (the system size), f_i stands for the deterministic vector field of particle i, and "noise$_i$" denotes some - for the moment generic - randomness in the evolution of particle i. So, we obtain a system of SDEs for each particle. Then one can couple individual dynamics by adding

interaction terms:

$$\dot{x}_i(t) = f_i(x_i(t)) + \text{noise}_i + \text{ interaction}_i(x_1(t), \dots, x_N(t)).$$

This is a sketch of what an interacting particle system is. We remark that x_i can take both discrete or continuous values (e.g. $\mathbb{R}^d$) and we will consider here a continuous-time dynamics, but in general it is possible to study also discrete-time particle systems.

Within the framework of interacting particle systems, we will focus in these notes on the following question: *how* can a family of interacting units generate a rhythm which is inscribed nowhere in the single unit and without the aid of external periodic forcing?

In particular, we will give a concrete example of an interacting particle system displaying macroscopic periodic behaviors due to the joint action of noise and specific interaction structure. But first, we will go through an overview on possible mechanisms which might be responsible of these behaviors.

2 Overview of the literature

Various stylized models have been proposed to unveil the key mechanisms behind the emergence of collective self-sustained rhythms. However, in most of them rigorous results are hard to obtain, as the study ends up in looking for stable attractors of nonlinear infinite-dimensional dynamical systems [8, 10].

Analytically tractable models can be obtained by considering mean-field interactions. Within this context, the existence of periodic collective behaviors has been proved for some classes of mean-field systems derived as perturbation of classical reversible models from statistical physics by adding *dissipation* in the interaction term [3, 5, 6]. Within a continuous-time Markovian dynamics, dissipation dampens the strength of the interaction among particles during the time when no transition occurs and breaks the time-reversibility of the system, which is incompatible with limit cycles. The simplest spin system within this class is the dissipative mean-field Ising model proposed in [5]. Coupled diffusions with dissipation have been considered in [3]. A contact process with dissipation has been investigated in [6]. Beyond the mean-field setting, in [2] dissipation has been added to a one-dimensional Ising model with nearest-neighbours interactions and emergence of rhythmic behaviors was proved in that case as well.

Besides dissipation, and back to the mean-field framework, *delay* in the interactions may also produce rhythmic behaviors, as highlighted in [7] for interacting Hawkes processes and in [4] for spin models.

Finally, an interesting family of models, which has naturally emerged in applications, is a multi-species extension of the Curie-Weiss model. It has been proved that having two groups of spins with possibly different sizes and different inter- and intra-population interactions suffices for the emergence of macroscopic oscillations [4]. In this regard, it is interesting to examine how the *interplay between interaction and noise* can lead to the appearance of persistent oscillatory behaviors. Indeed, on the one hand it has been pointed out ([4, 7, 14]) that a *specific network structure* may favor the emergence of collective rhythms. On the other hand, many works ([12, 13, 10]) have shown how *noise* can lead to the emergence of periodic laws in systems whose deterministic counterparts do not display

any periodic behavior. Generally speaking, noise added to a deterministic dynamical system may destabilize fixed points of the deterministic dynamics and create of favour limit cycles.

The joint action of noise and the topology of the interaction network in generating collective periodic behaviors is the object of the present notes. We will examine it in more detail by considering the toy model studied in [11].

3 A frustrated network of interacting diffusions

In this section, we combine a specific topology of the interaction network with noise and we present a toy model of frustratedly interacting diffusions that shows *noise-induced periodicity*: in the infinite volume limit, in a certain range of interaction strengths, although the system has no periodic behavior in the zero-noise limit, a moderate amount of noise may generate an attractive periodic law. We stress right away that the peculiar feature of the model under consideration is that the structure of the interaction network depends on the noise in that it is the noise that switches on the interaction terms, thus leading to periodic dynamics.

This model provides a concrete and intuitive example of an interacting particle system featuring emergent collective periodic behavior.

3.1 Description of the interacting particle system

Let us consider a system of N particles moving on $\mathbb{R}$. We divide the N particles into two disjoint communities I_1 and I_2 of sizes N_1, N_2 respectively, and we denote by $(x_i^{(N)}(t))_{i=1}^{N_1}$ the positions of the particles of the first population at time t and by $(y_j^{(N)}(t))_{j=1}^{N_2}$ the positions of the particles of the second population at time t, so that

$$\mathbf{z}^{(N)}(t) = \Big(\overbrace{x_1^{(N)}(t), x_2^{(N)}(t), \dots, x_{N_1}^{(N)}(t)}^{\text{Community } I_1}, \overbrace{y_1^{(N)}(t), y_2^{(N)}(t), \dots, y_{N_2}^{(N)}(t)}^{\text{Community } I_2} \Big)$$

represents the state of the whole system at time t.

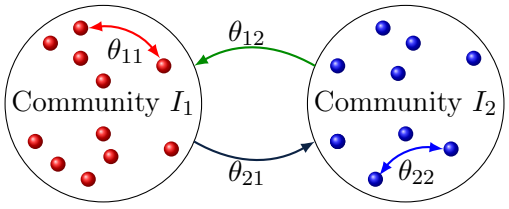


Figure 2: Sketch of the interaction network. We divide particles into two populations, I_1 and I_2. θ_{11} (resp. θ_{22}) is the parameter tuning the strength of the interaction between any pair of particles in population I_1 (resp. I_2). That is, these parameters tune the strength of intra-population interactions. θ_{12} (resp. θ_{21}) tunes the strength of the influence of any particle of the second (resp. first) population over any particle of the first (resp. second) one. We say that these parameters tune inter-population interactions. All interactions parameters are constant in time.

The interaction network is sketched in Figure 2, and in particular we will consider here cooperative intra-population interactions (i.e. θ_{11}, $\theta_{22} > 0$) and frustrated inter-population interactions (i.e. $\theta_{12}\theta_{21} < 0$).

We also define two macroscopic quantities which are the average positions of the two populations at time t and which we will show to exhibit a periodic behavior, in the thermodynamic limit:

$$m_1^{(N)}(t) := \frac{1}{N_1}\sum_{j=1}^{N_1} x_j^{(N)}(t) \qquad m_2^{(N)}(t) := \frac{1}{N_2}\sum_{j=1}^{N_2} y_j^{(N)}(t)$$

The dynamics of the particles are described by the following SDEs:

$$\begin{aligned} dx_j^{(N)} &= \left(-\left(x_j^{(N)}\right)^3 + x_j^{(N)}\right)dt - \alpha\theta_{11}\left(x_j^{(N)} - m_1^{(N)}\right)dt \\ &\quad - (1-\alpha)\,\theta_{12}\left(x_j^{(N)} - m_2^{(N)}\right)dt + \sigma dw_j, \qquad \text{for } j = 1,\dots,N_1 \\ dy_j^{(N)} &= \left(-\left(y_j^{(N)}\right)^3 + y_j^{(N)}\right)dt - (1-\alpha)\,\theta_{22}\left(y_j^{(N)} - m_2^{(N)}\right)dt \\ &\quad - \alpha\theta_{21}\left(y_j^{(N)} - m_1^{(N)}\right)dt + \sigma dw_{N_1+j}, \qquad \text{for } j = 1,\dots,N_2 \end{aligned} \tag{1}$$

where $\alpha := N_1/N$ is the fraction of particles in the first population, $(w_i)_{i=1}^N$ are N independent standard Brownian motions and we assume without loss of generality that $\theta_{12} < 0$, while $\theta_{21} > 0$: particles in the first population want to stay close to the average position of the particles in the second population, and particles in the second population want to do the contrary. Moreover, $\sigma \geq 0$ tunes the amount of noise in the system. This is a concrete example of an interacting particle system.

Remark 3.1 Existence and uniqueness of a strong solution to (1) can be established via the Khasminskii criterion ([9]), by taking the norm-like function

$$V(\mathbf{z}^{(N)}) = \frac{1}{N_1}\sum_{i=1}^{N_1}\left[\frac{(x_i^{(N)})^4}{4} + \frac{(x_i^{(N)})^2}{2}\right] + \frac{1}{N_2}\sum_{i=1}^{N_2}\left[\frac{y_i^{(N)}}{4} + \frac{y_i^{(N)}}{2}\right].$$

To have an insight of what we might expect from the dynamics in (1), let us consider two representative particles, one for each population, whose positions $x_1^{(N)}$ and $y_1^{(N)}$ obey the first and second equations in (1) respectively. In the first place, they will be both subject to a deterministic vector field coming from a double well potential. In the absence of any other term, particles would be attracted towards one of the equilibria of this deterministic vector field ($(0,0)$, $(1,1)$, $(-1,-1)$). Second, each of the two particles under consideration will tend to conform to the average position of the respective population, $m_i^{(N)}$, $i = 1,2$ (since θ_{11}, $\theta_{22} > 0$, cooperative intra-population interactions). Then, $x_1^{(N)}$ will be attracted towards the average position of the second population - as the coefficient $-(1-\alpha)\theta_{12} < 0$ -, whereas $y_1^{(N)}$ will tend to steer away from the average position of the first population

- as $-\alpha\theta_{21} > 0$. Last, the diffusive effect of the Brownian noise, which tends to spread particles in independent random directions, is tuned by the diffusion coefficient σ.

Now, imagine to initialize all the particles at the same position, namely, $x_i^{(N)}(0) = y_j^{(N)}(0) = m_1^{(N)}(0) = m_2^{(N)}(0) = z_0$ for all $i = 1, \dots, N_1$, $j = 1, \dots, N_2$, and imagine that $\sigma = 0$. Then all the interaction terms in (1) will remain equal to zero and particles will move together following the deterministic vector field and ending up in one of its equilibria. On the contrary, if $\sigma > 0$, even if initialized all at one of the stable fixed points of the cubic vector field, particles will start to diffuse away from the equilibrium in different, independent directions. In turn, the interaction terms will become different from zero, and, as the inter-population interaction coefficients have opposite signs, the rest states of the two populations will become incompatible, so they will form a frustrated pair of systems.

Overall, intuitively, it is the interplay between the frustrated interaction network and noise which might generate collective periodic behaviors in this model. However, the role of the noise seems to be crucial in this sense, since it is the noise that "switches on" the interaction terms, generating *dynamical* frustration.

This feature is a hallmark of the so-called phenomenon of noise-induced periodicity, and numerical simulations corroborate and complete this picture, as shown in Subsection 3.2.

3.2 Numerical evidence in favor of the emergence of collective rhythms

Numerical simulations of the dynamics in (1) give evidence of the noise-induced periodicity phenomenon: in appropriate parameter regimes, where the noise intensity σ is tuned at an intermediate value, the average positions of the two populations display an oscillatory behavior. An example of this is reported in Figure 3.

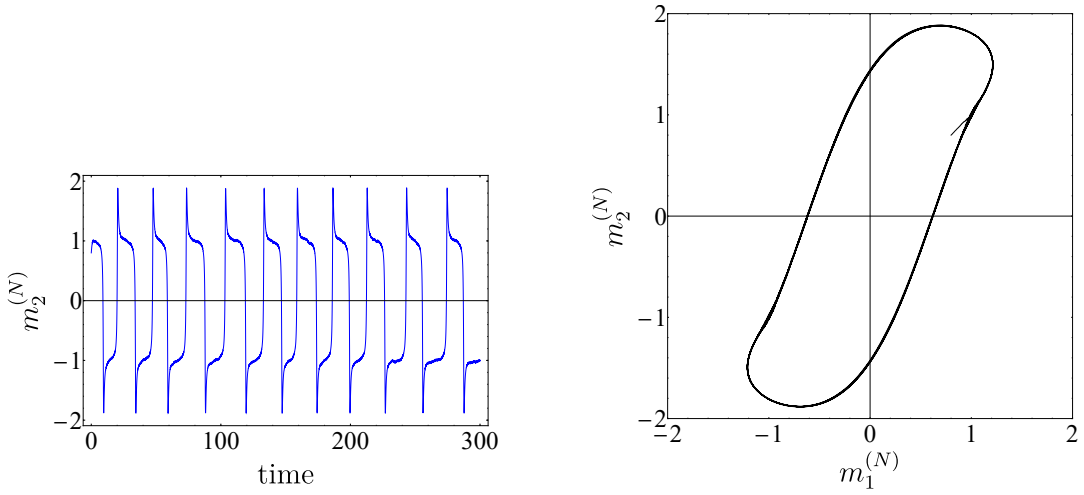


Figure 3: Numerical simulations of the particle system (1). Left: sample trajectory of $m_2^{(N)}$ against time. Right: $(m_1^{(N)}(t), m_2^{(N)}(t))$ in the phase space of the empirical averages. Parameters: 10^6 iterations, time-step $dt = 0.005$, $N = 1000$, $\alpha = 0.5$, $\theta_{11} = \theta_{22} = 8$, $\theta_{12} = 4$, $\theta_{21} = -8$, $\sigma = 0.1$.

On the contrary, when $\sigma = 0$, no oscillatory behaviors are observed, and the system is attracted to a fixed point, confirming the heuristics given above (see Figure 4).

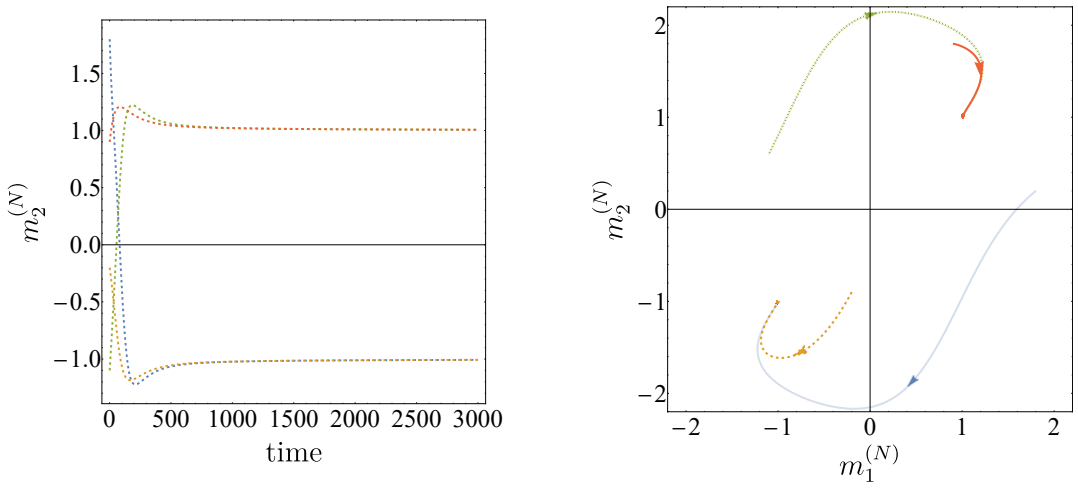


Figure 4: Numerical simulations of (1) with $\sigma = 0$. Same parameters as in Figure 3. Different instances of initial conditions.

Last, letting $\sigma \gg 1$ completely alters the dynamics, which essentially becomes a Brownian motion.

3.3 The thermodynamic limit: propagation of chaos

The picture given by numerical simulations can be made rigorous if we study the limit of (1) as the system size N tends to infinity (thermodynamic limit).

Consider the following system:

$$
\begin{aligned}
dx &= \left[-x^3 + x - \alpha\theta_{11}\left(x - \mathbb{E}[x]\right) - (1-\alpha)\theta_{12}\left(x - \mathbb{E}[y]\right)\right] dt + \sigma dw_1 \\
dy &= \left[-y^3 + y - \alpha\theta_{21}\left(y - \mathbb{E}[x]\right) - (1-\alpha)\theta_{22}\left(y - \mathbb{E}[y]\right)\right] dt + \sigma dw_2
\end{aligned} \tag{2}
$$

where $\mathbb{E}$ stands for the expectation with respect to the probability measure $\mathcal{L}(x(t), y(t))$ and $(w_1(t))_{t\geq 0}$ and $(w_2(t))_{t\geq 0}$ are two independent standard Brownian motions.

Remark 3.2 The existence and pathwise uniqueness of a strong solution to (2) can be proved essentially via standard Picard iteration and contraction arguments (Gronwall's lemma) respectively. See [11].

We can prove via coupling arguments (see [11]) the following theorem.

Theorem 3.1 (Propagation of Chaos)
Fix $T > 0$. Let $(x_1^{(N)}(t), \dots, x_{N_1}^{(N)}(t), y_1^{(N)}(t), \dots, y_{N_2}^{(N)}(t))_{t\in[0,T]}$ be the solution to (1) with an initial condition satisfying the following requirements:

- *the collection* $(x_1^{(N)}(0), \dots, x_{N_1}^{(N)}(0), y_1^{(N)}(0), \dots, y_{N_2}^{(N)}(0))$ *is a family of independent random variables.*
- *the random variables* $(x_1^{(N)}(0), \dots, x_{N_1}^{(N)}(0))$ *(resp.* $(y_1^{(N)}(0), \dots, y_{N_2}^{(N)}(0))$*) are identically distributed with law* λ_x *(resp.* λ_y*). We assume that* λ_x *and* λ_y *have finite second moment.*
- *the random variables* $x_j^{(N)}(0)$ *and* $y_k^{(N)}(0)$ *are independent of the Brownian motions* $((w_i(t))_{t\in[0,T]})_{i=1}^N$ *for all* $j = 1, \dots, N_1$ *and* $k = 1, \dots, N_2$.

Moreover, let $(x_1(t), \dots, x_{N_1}(t), y_1(t), \dots, y_{N_2}(t))_{t\in[0,T]}$ *be the process whose entries are independent and such that* $((x_j(t))_{t\in[0,T]})_{j=1}^{N_1}$ *(resp.* $((y_k(t))_{t\in[0,T]})_{k=1}^{N_2}$*) are copies of the solution to the first (resp. second) equation in (2), with the same initial conditions and the same Brownian motions used to define system (1). Here, "the same" means componentwise equality.*
Define the index sets $\mathcal{I} = \{i_1, \dots, i_{k_1}\} \subseteq \{1, \dots, N_1\}$*, with* $|\mathcal{I}| = k_1$*, and* $\mathcal{J} = \{j_1, \dots, j_{k_2}\} \subseteq \{1, \dots, N_2\}$*, with* $|\mathcal{J}| = k_2$*. Then, we have*

$$\lim_{N\to+\infty} \mathbb{E}\left[\sup_{t\in[0,T]} \left|\mathbf{z}_{k_1,k_2}^{(N)}(t) - \mathbf{z}_{k_1,k_2}(t)\right|\right] = 0, \tag{3}$$

with $|\mathbf{z}|$ *the* ℓ^1*-norm of a vector* $\mathbf{z}$*,* $\mathbf{z}_{k_1,k_2}^{(N)}(t) = (x_{i_1}^{(N)}(t), \dots, x_{i_{k_1}}^{(N)}(t), y_{j_1}^{(N)}(t), \dots, y_{j_{k_2}}^{(N)}(t))$ *and* $\mathbf{z}_{k_1,k_2}(t) = (x_1(t), \dots, x_{k_1}(t), y_1(t), \dots, y_{k_2}(t))$.

Remark 3.3 Theorem 3.1 claims that, for all $T > 0$ and $t \in [0, T]$, any random vector of the form $(x_{i_1}^{(N)}(t), \dots, x_{i_{k_1}}^{(N)}(t), y_{j_1}^{(N)}, \dots, y_{j_{k_2}}^{(N)}(t))$ converges in distribution, as $N \to \infty$, to a vector $(x_1(t), \dots, x_{k_1}(t), y_1(t), \dots, y_{k_2}(t))$, whose entries are independent random variables such that $(x_i)_{i=1}^{k_1}$ are copies of the solution to the first equation in (2) and $(y_j)_{j=1}^{k_2}$ are copies of the solution to the second equation in (2). This is usually referred to as the phenomenon of the Propagation of Chaos: starting from i.i.d. initial conditions, as $N \to \infty$, independence propagates in time, in the sense that the evolution of each particle remains independent of the evolution of any finite subset of the others, despite interactions. This is coherent with the fact that individual units interact only through the empirical means of the two populations, over which the influence of a finite number of particles becomes negligible when taking the infinite volume limit (mean-field interaction).

Furthermore, Theorem 3.1 yields in the limit two representative equations, one for each population, such that the trajectory of any particle inside population I_1 (resp. I_2) is well-approximated, in the limit as $N \to +\infty$, by the trajectory of a single, representative particle obeying the first (resp. second) equation in (2).

System (2) is a system of two McKean-Vlasov SDEs, hence, it can have solutions with periodic law ([12, 13]). It is however very hard to gain insight into its long-time behavior or to find periodic solutions as the problem is infinite dimensional, due to the presence of nonlinearity and noise.

Despite this, as a first step, we can study system (2) in the absence of noise and show that, when $\sigma = 0$, no limit cycle is present, hence the solution to system (2) has no periodic behaviors ([11]).

Furthermore, we can have insight into the behavior of system (2) for $\sigma > 0$ by means of a suitable approximation, as explained in the next subsection.

3.4 Gaussian approximation

It can be easily seen, by applying Itô's formula, that the p-th moments of $x(t)$ and $y(t)$ solutions to (2) - which we will denote by $m_p^x(t) \coloneqq \mathbb{E}[x^p(t)]$ and $m_p^y(t) \coloneqq \mathbb{E}[y^p(t)]$ -, satisfy the following equations:

$$
\begin{aligned}
\frac{dm_p^x}{dt} &= -pm_{p+2}^x + pm_p^x - \alpha\theta_{11}p\left(m_p^x - m_1^x\, m_{p-1}^x\right) \\
&\quad - (1-\alpha)\,\theta_{12}p\left(m_p^x - m_1^y\, m_{p-1}^x\right) + \tfrac{\sigma^2}{2}p(p-1)m_{p-2}^x \\
\frac{dm_p^y}{dt} &= -pm_{p+2}^y + pm_p^y - \alpha\theta_{21}p\left(m_p^y - m_1^x\, m_{p-1}^y\right) \\
&\quad - (1-\alpha)\theta_{22}p\left(m_p^y - m_1^y\, m_{p-1}^y\right) + \tfrac{\sigma^2}{2}p(p-1)m_{p-2}^y.
\end{aligned} \tag{4}
$$

Then, we have the following result, proved in [11].

Theorem 3.2 (Gaussian approximation) *Fix $T > 0$. Let $((x(t), y(t)), 0 \le t \le T)$ solve Equation (2) with deterministic initial conditions $x(0) = x_0$ and $y(0) = y_0$. There exists a Gaussian Markov process $((\tilde{x}(t), \tilde{y}(t)), 0 \le t \le T)$ with $\tilde{x}(0) = x_0$ and $\tilde{y}(0) = y_0$ satisfying the properties:*

1. *The first two moments of $\tilde{x}(t)$ and $\tilde{y}(t)$ satisfy the respective equations in (4) for $p = 1, 2$.*
2. *For all $T > 0$, there exists a constant $C_T > 0$ such that, for every $\sigma > 0$, it holds*

$$
\mathbb{E}\left[\sup_{t\in[0,T]}\{|x(t) - \tilde{x}(t)| + |y(t) - \tilde{y}(t)|\}\right] \le C_T\sigma^2.
$$

This means that the processes $(\tilde{x}(t), 0 \le t \le T)$ and $(\tilde{y}(t), 0 \le t \le T)$ are simultaneously σ-closed to the solutions of (2).

Theorem 3.2 claims that, in the presence of an appropriate amount of noise, the solution $(x(t), y(t))_{t\ge0}$ to system (2) can be approximated - over any finite time interval $[0, T]$ - by a pair of independent Gaussian processes $(\tilde{x}(t), \tilde{y}(t))_{t\in[0,T]}$ whose means and variances obey the same equations satisfied by the means and variances of x and y ([11]). It also provides the explicit (deterministic) equations for the mean and variance of those processes.

In this way, we can reduce the study of the infinite-dimensional system (2) to the study of a four-dimensional system of ODEs - the equations $\mathbb{E}[\tilde{x}(t)]$, $\mathbb{E}[\tilde{y}(t)]$, $Var[\tilde{x}(t)]$, $Var[\tilde{y}(t)]$. Such dynamical system undergoes a Hopf bifurcation as σ crosses a critical

value (i.e., moving σ across some critical threshold, a stable fixed point loses stability and a stable periodic orbit arises in its place). Hence, as a consequence, in a certain range of the noise intensity, system (4) has a limit cycle as a long-time attractor, implying that the laws of the Gaussian processes are periodic. This confirms the presence of a periodic law for system (2) and yields a good qualitative description of the emergence of the self-sustained oscillations observed for system (1).

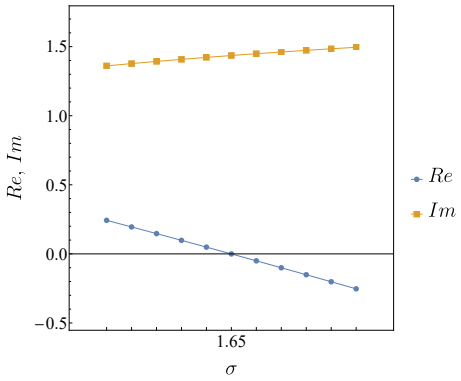


Figure 5: A Hopf bifurcation for a dynamical system at an equilibrium point can be detected by checking whether a pair of complex eigenvalues of the linearized system around the equilibrium crosses the imaginary axis as some parameter of the system changes. Here we plot the real and imaginary parts of one eigenvalue of the linearized system for the means and variances of the Gaussian processes $\tilde{x}$ and $\tilde{y}$. See [11] for further details.

4 Conclusion

We have seen that collective periodic behaviors are ubiquitous in real-world systems. Nonetheless, the identification of minimal hypotheses needed to observe them is still an object of research.

Through the analysis of a toy model of interacting diffusions, we have deepen the study of how the interplay between a simple interaction network and noise might generate sustained rhythms. In particular, we have seen that the role of the noise is crucial, in this model, to generate interesting behaviors: despite the frustrated interaction network, no collective behavior is observed in the zero-noise limit system, whereas this latter can have a periodic law when the noise intensity crosses a certain threshold. This phenomenon goes under the name of noise-induced periodicity.

References

[1] M. Aleandri, I.G. Minelli, *Opinion dynamics with Lotka-Volterra type interactions.* Electron. J. Probab. 2019.

[2] R. Cerf, P. Dai Pra, M. Formentin, D. Tovazzi, *Rhythmic behavior of an Ising Model with dissipation at low temperature.* ALEA. 2021.

[3] F. Collet, P. Dai Pra, M. Formentin, *Collective periodicity in mean-field models of cooperative behavior.* NoDEA. 2015.

[4] F. Collet, M. Formentin, D. Tovazzi, *Rhythmic behavior in a two-population mean-field Ising model.* Phys. Rev. E. 2016.

[5] P. Dai Pra, M. Fischer, D. Regoli, *A Curie-Weiss model with dissipation.* J. Stat. Phys. 2013.

[6] P. Dai Pra, E. Marini, *Noise-induced oscillations for the mean-field dissipative contact process.* arXiv:2403.03783.

[7] S. Ditlevsen, E. Löcherbach, *Multi-class oscillating systems of interacting neurons.* Stoch. Proc. Appl. 2015.

[8] G. Giacomin, C. Poquet, *Noise, interaction, nonlinear dynamics and the origin of rhythmic behaviors.* Braz. J. Prob. Stat. 2015.

[9] R. Khasminskii, "Stochastic Stability of Differential Equations". Springer, Berlin (2012).

[10] B. Lindner, J. Garcia-Ojalvo, A. Neiman, L. Schimansky-Geier, *Effects of noise in excitable systems.* Phys. Rep. 2004.

[11] E. Marini, L. Andreis, F. Collet, M. Formentin, *Noise-induced periodicity in a frustrated network of interacting diffusions.* NODEA. 2023.

[12] M. Scheutzow, *Noise can create periodic behavior and stabilize nonlinear diffusions.* Stochastic Process. Appl. 1985.

[13] M. Scheutzow, *Some examples of nonlinear diffusion processes having a time-periodic law.* Ann. Probab. 1985.

[14] J.D. Touboul, *The hipster effect: When anti-conformists all look the same.* Discrete and Continuous Dynamical Systems-B. 2019.

Differential games and large population limits beyond the classic Mean-Field setting

Davide Francesco Redaelli (*)

Abstract. Differential game theory is a branch of mathematics that touches many fields such as control and game theories, probability, stochastic and partial differential equations. An interesting aspect of it is studying strategies of the players that are optimal in that they produce a situation of equilibrium, for example in the famous sense due to Nash, and also seeing what happens when the number of players grows and possibly becomes infinite. These pages try to give a brief introduction to the theory aimed at a wide audience of mathematicians possibly unaware of the subject, with the final purpose of presenting the main topics which my doctoral research focused on.

1 What's a differential game?

Consider a *controlled* equation of the form

$$\begin{cases} \dfrac{\mathrm{d}X_t}{\mathrm{d}t} = b(t, X_t, \alpha_t), \quad t \in [0,T] \\ X_0 = x_0, \end{cases}$$

where $X\colon [0,T] \to \mathbb{R}$ is the *state* variable, $b\colon [0,T] \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is the *drift* (given) and $\alpha\colon [0,T] \to \mathbb{R}$ is the *control*. The idea behind this terminology is that it is possible to *choose* a control in order to affect the trajectory of X, but why should one want to? Because the controlled equation is coupled with a *cost* of the form

$$J(\alpha) \stackrel{\text{def}}{=} \int_0^T L(t, X_t, \alpha_t)\,\mathrm{d}t,$$

for some function $L\colon [0,T] \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ (referred to as the *running cost*), and one wishes to minimise such a cost, the way to do that being through the choice of α. So we can think that the state variable can represent any quantity that affects the cost and imagine

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: redaelli@math.unipd.it . Seminar held on 9 May 2024.

that there's a *player* that can choose its[(1)] control function in order to reduce its cost by modifying its state.

Suppose now that we are given N controlled equations

$$\begin{cases} \dfrac{\mathrm{d}X_t^i}{\mathrm{d}t} = b^i(t, X_t^1, \dots, X_t^N, \alpha_t^i), \quad t \in [0,T] \\ X_0^i = x_0^i, \end{cases} \qquad i = 1, \dots, N,$$

and corresponding N costs

$$J^i(\alpha) \stackrel{\text{def}}{=} \int_0^T L^i(t, X_t^1, \dots, X_t^N, \alpha_t^i)\,\mathrm{d}t, \qquad i = 1, \dots, N,$$

for some $b^i, L^i \colon [0,T] \times \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}$. This is what we call an N-player *differential game*. Each index i corresponds to a different player, who can choose its control α^i in order to minimise its cost J^i. This is done *simultaneously* by all players, so the "game"-component consists in the players competing with each other in order to minimise their costs, which depend on all states and thus all controls. The "differential"-component consists in the underlying differential equations governing the states of the players: the game is not *one-shot* but develops on the whole time interval $[0,T]$ in such a way that each player has to choose its strategy *at any time*, possibly *adapting* it to the behaviour of the other players.

1.1 Nash equilibria

At this point it should be fairly clear that when studying differential games the focus is placed on the *strategies* of the players, namely *how they choose their controls*. In particular, we are usually interested in strategies that are *optimal*, in a suitable sense which was first formalised by J. F. Nash, Jr within the theory of noncooperative games. We say that a set of strategies $(\alpha^1, \dots, \alpha^N)$ is a *Nash equilibrium* for the game if, for all $i \in \{1, \dots, N\}$ and for any strategy β adopted by the i-th player,

$$J^i(\alpha^{-i}, \alpha^i) \leq J^i(\alpha^{-i}, \beta),$$

where α^{-i} is the vector α deprived of the i-th coordinate. This means that we are in a situation of Nash equilibrium if *no player can have an advantage* (that is, lessen its cost) *by unilaterally changing its strategy*.

This notion of equilibrium strongly depends upon the pieces of information available to the players and how they can use them; that is, the above definition of a Nash equilibrium can only make sense once we have properly defined the nature of the *frozen* strategies α^{-i}, so it is necessary to specify how the players *update* their controls when one of them changes its strategy. We shall consider two main different models: the *open-loop* model and the (Markovian) *closed-loop* model.

[(1)] As stated in [3, Remark 2.1], the biological genders of the players 'have no bearing on what we are interested in, and keeping track of grammatical genders can only be a hindrance and a distraction. [...] As a result, we shall treat the players as *genderless*.'

1.1.1 Open-loop Nash equilibria

If the players cannot update their strategies by adapting their controls to the states of the other players, we talk about open-loop strategies. This means that the *admissible* controls are of the form

$$\alpha_t^i = \phi^i(t, x_0),$$

for some function $\phi^i \colon [0,T] \times \mathbb{R}^N \to \mathbb{R}$: that is, they depend *only* on the instant t in time and on the *initial position* x_0 of the players. In this case, when the i-th player changes its strategy from α^i to β (that is, from ϕ^i to some other $\tilde{\phi}^i$) while the other players keep their strategies α^{-i}, the state of the system will be affected (note that b^j depends on X^i even if $j \neq i$), yet the controls of the other players won't, as their strategies don't see the state X_t^i.

1.1.2 Closed-loop Nash equilibria

If the players can update their strategies by adapting their controls to the states of the other players, we talk about closed-loop strategies. This means that the admissible controls are of the form

$$\alpha_t^i = \phi^i(t, x_0, X_t^1, \dots, X_t^N)$$

for some $\phi^i \colon [0,T] \times \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$. In this case, when the i-th player changes its strategy by using a different ϕ^i, the controls of the other players will be affected, even if they don't change their strategies! This happens because, for the j-th player ($j \neq i$), *keeping a strategy* means *fixing* ϕ^j, and not the resulting control α^j, which will in fact change because of the changes in the value of the state, since ϕ^j is a function of X_t^i.

It is important to emphasise that in open-loop models, when a player makes a decision, it may not be able to take into account the plays of its opponents since its decisions can only be functions of the initial state. In closed-loop models, instead, past plays, as they impact on the values of the state, become part of the common knowledge of the players, who can then react to them. This should give an insight on why open-loop equilibria are less realistic but also more mathematically tractable than closed-loop equilibria; indeed, in the former case, players need not consider how their opponents would react to deviations from the equilibrium. On the other hand, one should expect that when the impact of the players on their opponents' costs is small, open-loop and closed-loop equilibria should be almost the same.

2 How's a differential game related to partial differential equations?

2.1 The 1-player game

Let's go back to the starting controlled system, which is basically a 1-player game. Given a set $\mathcal{A}$ of admissible controls (for example, open-loop controls or closed-loop controls, with values in some compact set $A \subset \mathbb{R}$) we define the *value function* u of the player as follows:

for any $t \in [0,T]$ and $x \in \mathbb{R}$,

$$u(t,x) \stackrel{\text{def}}{=} \inf_{\alpha \in \mathcal{A}} \int_t^T L(s, X_s, \alpha_s)\, \mathrm{d}s,$$

where X solves

$$\begin{cases} \dfrac{\mathrm{d}X_s}{\mathrm{d}s} = b(s, X_s, \alpha_s), \quad s \in [t,T] \\ X_t = x. \end{cases} \tag{2.1}$$

The value function is the main tool that indicates how the player should choose its control in order to play optimally. The starting point to show that is the so-called *dynamic programming principle*: it states that

$$u(t_1, X_{t_1}) = \inf_{\alpha \in \mathcal{A}} \left(\int_{t_1}^{t_2} L(s, X_s, \alpha_s)\, \mathrm{d}s + u(t_2, X_{t_2}) \right) \quad \forall\, t_1 < t_2.$$

Its interpretation is that, in order to play optimally at time t_1, the player does not need to predict the whole future strategy if it knows what would be the minimum cost at some future time t_2: in this case it is enough to focus on the optimisation between t_1 and t_2.

Clearly this is interesting as t_1 and t_2 can be taken arbitrarily close to one another, so that in the limit as $t_2 - t_1 \to 0^+$ the fundamental consequence is that the player only needs to know the current state (that is, the state at time $t = t_1$) and play accordingly. Indeed, if we take $t_1 = t$ and $t_2 = t + h$ we have, by Taylor's expansion as $h \to 0^+$,

$$u(t, X_t) \sim \inf_{\alpha \in \mathcal{A}} \left(\int_t^{t+h} L(s, X_s, \alpha_s)\, \mathrm{d}s + u(t, X_t) + \Big(\partial_t u(t, X_t) + \partial_x u(t, X_t) \cdot \frac{\mathrm{d}X_t}{\mathrm{d}t} \Big) h \right);$$

that is, dividing by h and recalling the equation of X_t,

$$\inf_{\alpha \in \mathcal{A}} \left(\frac{1}{h} \int_t^{t+h} L(s, X_s, \alpha_s)\, \mathrm{d}s + \partial_t u(t, X_t) + \partial_x u(t,x) \cdot b(t, x, \alpha_t) \right) \xrightarrow[h \to 0^+]{} 0.$$

So, letting $h \to 0^+$ we obtain

$$-\partial_t u(t,x) - \inf_{\alpha \in \mathcal{A}} \big(L(t,x,\alpha_t) + \partial_x u(t,x) \cdot b(t,x,\alpha_t) \big) = 0, \tag{2.2}$$

where $\alpha_t \in A$ can be arbitrarily chosen. Then if one defines the *Hamiltonian* $H \colon [0,T] \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ as

$$H(t,x,p) \stackrel{\text{def}}{=} \sup_{a \in A} \big(-L(t,x,a) - p \cdot b(t,x,a) \big),$$

equality 2.2 reads as the following *Hamilton–Jacobi equation*:

$$-\partial_t u(t,x) + H(t,x,\partial_x u(t,x)) = 0, \quad (t,x) \in [0,T] \times \mathbb{R}.$$

Note that this partial differential equation is backward in time as the very definition of u gives us the terminal condition $u(T,x) = 0$ for all $x \in \mathbb{R}$.

At this point, we have associated a partial differential equation to our game, but why is this useful? Let $\alpha^*(t,x)$ be a maximum point in the definition of $H(t,x,\partial_x u(t,x))$; that is,

$$H(t,x,\partial_x u(t,x)) = -L(t,x,\alpha^*(t,x)) - \partial_x u(t,x)\cdot b(t,x,\alpha^*(t,x)).$$

By a so-called *verification theorem* it is possible to prove that if X^o solves (2.1) with $\alpha_s = \alpha_s^o \overset{\text{def}}{=} \alpha^*(s, X_s^o)$, then

$$u(t,x) = \int_t^T L(s, X_s^o, \alpha_s^o)\,\mathrm{d}t;$$

this means that the control α^o that we've built using α^* is optimal for the player, as it minimises its cost. Therefore, in order to determine the *optimal trajectory* X^o one basically needs to know the optimal drift $b(t,x,\alpha^*(t,x))$ (that is, the drift computed at $\alpha = \alpha^*$), and here the Hamilton–Jacobi equation comes into play: if $\alpha^*(t,x)$ is unique, then the optimal drift can be computed by knowing (the gradient of) the solution u, as the envelope theorem[2] tells us that

$$b(t,x,\alpha^*(t,x)) = -\partial_p H(t,x,\partial_x u(t,x)). \tag{2.3}$$

2.2 The N-player game

This deduction of an associated Hamilton–Jacobi equation can be done for the N-player game as well. Note that the definition of a Nash equilibrium tells us that, if $\alpha^o = (\alpha^{o,1},\dots,\alpha^{o,N})$ is a set of optimal strategies, then

$$J^i(\alpha^o) = \inf_{\alpha^i\in\mathcal{A}} J^i(\alpha^{o,-i},\alpha^i) \quad \forall\, i\in\{1,\dots,N\};$$

that is, $\alpha^{o,i}$ is optimal for the i-th player who plays a 1-player game with cost

$$\tilde{J}^i(\alpha^i) = J^i(\alpha^{o,-i},\alpha^i)$$

and state equation

$$\frac{\mathrm{d}X_t^i}{\mathrm{d}t} = b^i(t, X_t^{o,1},\dots,X_t^i,\dots,X^{o,N},\alpha_t^i),$$

where the trajectories of $X^{o,j}$ for $j\neq i$ are determined by the "frozen" controls $\alpha^{o,j}$. Then we can define the value function of the i-th player

$$u^i(t,x) \overset{\text{def}}{=} \inf_{\alpha^i\in\mathcal{A}} \int_t^T L^i(s, X_s^{o,-i}, X_s^i, \alpha_s^i)\,\mathrm{d}s$$

and proceed in analogy to what we did before. In the end, we obtain a system of N Hamilton–Jacobi equations

$$-\partial_t u^i(t,x) + \tilde{H}^i(t,x,D_x u^i(t,x)) = 0, \quad (t,x)\in[0,T]\times\mathbb{R}^N,$$

[2] Given $f = f(a,x)\colon A\times\mathbb{R}\to\mathbb{R}$ continuous with $\partial_x f$ continuous, the function $F(x)\overset{\text{def}}{=}\sup_{a\in A} f(a,x)$ is differentiable at each x such that $\arg\max_{a\in A} f(a,x)$ is a unique point $a^*(x)$, with $\partial_x F(x) = \partial_x f(a^*(x),x)$.

where $D_x \stackrel{\text{def}}{=} (\partial_{x^1}, \dots, \partial_{x^N})$ and $\tilde{H}^i \colon [0,T] \times \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ is given by

$$\tilde{H}^i(t,x,p) \stackrel{\text{def}}{=} \sup_{a \in A} \big(-L^i(t,x,a) - p^i \cdot b^i(t,x,a) \big) - \sum_{j \neq i} p^j \cdot b^j(t,x,\alpha_t^{o,j}).$$

Now we can define a *reduced* Hamiltonian

$$H^i(t,x,p^i) \stackrel{\text{def}}{=} \sup_{a \in A} \big(-L^i(t,x,a) - p^i \cdot b^i(t,x,a) \big)$$

and use relation (2.3) to obtain

$$b^i(t,x,\alpha^{*,i}(t,x)) = -\partial_p H^i(t,x,\partial_{x^i} u^i(t,x)); \tag{2.4}$$

but we have said that $\alpha^{*,i}$ gives the optimal controls, so $\alpha^{*,i}(t,x) = \alpha_t^{o,i}$, and this holds by symmetry also for $j \neq i$. Then the Hamilton–Jacobi equations take the following form, which we call the *Nash system*:

$$-\partial_t u^i(t,x) + H^i(t,x,\partial_{x^i} u^i(t,x)) + \sum_{j \neq i} \partial_p H^j(t,x,\partial_{x^j} u^j(t,x)) \cdot \partial_{x^j} u^i(t,x) = 0. \tag{2.5}$$

This is a system of backward partial differential equations, *strongly coupled* due to the presence of the sum over j, and it is what one has to study when considering closed-loop equilibria.

On the other hand, in the case of open-loop equilibria with b^i independent of x^j for $j \neq i$, the system can be simplified as we consider only controls that do not see the states of the other players. This translates in assuming that $u^i(t,x) = u^i(t,x^i)$, thus implying that each term in the "bad sum" vanishes; also, as the left-hand side of (2.4) can depend only on x^i, the Hamiltonian has to have the form $H^i(t,x,p^i) = \mathcal{H}^i(t,x^i,p^i) - F^i(t,x)$. Then, recalling that $x^j = X_t^{o,j}$, we can reduce the Nash system to

$$-\partial_t u^i(t,x) + \mathcal{H}^i(t,x,\partial_x u^i(t,x)) = F^i(t,x,X_t^{o,-i}), \quad (t,x) \in [0,T] \times \mathbb{R}, \tag{2.6}$$

where X^o solves

$$\frac{\mathrm{d}X_t^{o,i}}{\mathrm{d}t} = -\partial_p \mathcal{H}^i(t, X_t^{o,i}, \partial_x u^i(t, X_t^{o,i})).$$

In any case, however, one needs to deal with a system of N Hamilton–Jacobi equations. This becomes harder and harder as N grows, but in the limit, when "$N = \infty$", some effects can appear that reduce the number of equations ...

3 What if the number of players grows? (Mean-Field regime)

The key idea of J.-M. Lasry and P.-L. Lions and of M. Huang, R. P. Malhamé and P. E. Caines was to study differential games with many players in analogy to *mean-field* models in statistical mechanics, that were developed to analyse the behaviour of many particles interacting with each other through *aggregated* (averaged) *quantities*. Given this

analogy, the first two mathematicians introduced the terminology of *Mean-Field Games* to designate the newborn theory.

The two crucial features of the Mean-Field setting are that players are *indistinguishable* from one another and the impact of the behaviour of each single player on the others is *negligible* as $N \to \infty$ (so one also calls *small players* the infinitely many players that one gets in the limit).

This latter fact exemplifies with players seeing each other through aggregate quantities, such as the averaged state of the system; that is, for example, the drift of the i-th player can be of the form

$$b^i(t,x) = \frac{1}{N-1}\sum_{j\neq i} b^i(t,x^i,x^j),$$

so that the "impact" of player j is of order $1/N$, thus vanishing as $N \to \infty$. Here we've dropped the dependence on the control in order to discuss the simpler setting of a pure interacting particle system.

On the other hand, indistinguishability requires that players be *interchangeable*, so $b^i = b$ for all i, for some "common" drift b. This gives the game *a lot* of *symmetry*. So, if in addition all players start from the same position it should be clear that their trajectories will all be the same as that of a *representative player* whose state evolves according to the following equation:

$$\frac{\mathrm{d}X_t^1}{\mathrm{d}t} = \frac{1}{N-1}\sum_{j\neq i} b(t,X_t^1,X_t^j) = b(t,X_t^1,X_t^1), \tag{3.1}$$

where the latter equality follows from the fact that $X^j = X^1$ for all j.

This situation can seem rather trivial as it is in fact true for any fixed N; nevertheless, in differential games the states of the players are usually governed by *stochastic* differential equations, and not ordinary ones. We've decided to stick to ordinary (or *deterministic*, or "*noiseless*") differential equations in order to simplify our presentation so far. Let us now mention that in the stochastic framework one has (for example, in the case of the first controlled equation we've presented) an equation like

$$\mathrm{d}X_t = b(t,X_t,\alpha_t)\,\mathrm{d}t + \mathrm{d}B_t; \qquad \text{that is,} \quad X_t = X_0 + \int_0^t b(s,X_s,\alpha_s)\,\mathrm{d}s + B_t,$$

where B_t is a Brownian motion, which models the presence of a "noise" affecting the state X. As a consequence, X is a stochastic process (that is, X_t is a random variable at any time t). Then, "having the same initial position" translates into "having the same initial distribution" and, very roughly speaking, the second equality in (3.1) is recovered in an appropriate way for $N = \infty$ via the Law of Large Numbers.

More precisely, if the X_t^j are independent and identically distributed, then by the Law of Large Numbers,

$$\frac{1}{N-1}\sum_{j\neq i} b(t,X_t^i,X_t^j) \xrightarrow[N\to\infty]{} \tilde{\mathbb{E}}\big[b(t,X_t^i,\tilde{X}_t^1)\big],$$

where $\tilde{X}_t^i$ is an independent copy of X_t^i and $\tilde{\mathbb{E}}$ is the expectation with respect to such a copy; that is, the drift of the i-th player tends to

$$\int_{\mathbb{R}} b(t, X_t^i, y)\, \mathrm{d}m_t(y), \qquad \text{with} \quad m_t \overset{\text{def}}{=} \mathcal{L}(X_t^1) \quad (\text{the law of } X_t^1),$$

and so $X^i = X^1$ for all i as they all evolve according to the same equation. However, we must note that the X_t^j are identically distributed due to the assumption, yet they are not independent. The fact that this reduction to a single equation is actually possible is a consequence of a crucial phenomenon that happens, which goes under the name of *propagation of chaos*. By this we mean that, due to the structure of the interactions, the states of the players in fact become "*more and more independent*" as the number of players grows, so that our heuristic application of the Law of Large Numbers eventually produces a faithful result.

3.1 Open-loop case: the Mean-Field system

When the players become infinitely many, according to the foregoing arguments, in the open-loop case that we considered we obtain a sole "state equation" of the form

$$\mathrm{d}X_t = b(t, X_t, \mathcal{L}(X_t), \alpha_t)\, \mathrm{d}t + \mathrm{d}B_t,$$

which describes the evolution of the *equilibrium distribution* of the players, provided that α is chosen as the optimal control that we identified in the previous discussion. Also, in order to adopt a pure partial differential equation viewpoint, it can be shown that $m_t \overset{\text{def}}{=} \mathcal{L}(X_t)$ solves in the sense of distributions the following *Fokker–Planck equation*:

$$\partial_t m - \tfrac{1}{2}\, \partial_{xx}^2 m + \partial_x(m\, b(t, x, m, \alpha_t)) = 0. \tag{3.2}$$

On the other hand, if similar Mean-Field assumptions are made also on the dependence of the costs on the states of the players, system (2.6) "tends to" a single Hamilton–Jacobi equation of the form

$$-\partial_t u(t,x) - \tfrac{1}{2}\, \partial_{xx}^2 u(t,x) + \mathcal{H}(t, x, \partial_x u(t,x)) = f(t, x, m_t), \tag{3.3}$$

where the second-order derivative ∂_{xx}^2 (also in (3.2) above) appears not as a result of the limit but when deriving the equation starting from the stochastic differential equation instead of the ordinary one.

Finally, remember that the optimal drift and the value function are related by equality (2.4), so the Nash equilibrium of the Mean-Field game is described by the following system of a backward Hamilton–Jacobi equation and a forward Fokker–Planck equation on $[0,T] \times \mathbb{R}$:

$$\begin{cases} -\partial_t u - \tfrac{1}{2}\, \partial_{xx}^2 u + \mathcal{H}(t, x, \partial_x u) = f(t, x, m) \\ \partial_t m - \tfrac{1}{2}\, \partial_{xx}^2 m - \partial_x(m\, \partial_p \mathcal{H}(t, x, \partial_x u)) = 0, \end{cases}$$

with their respective terminal condition $u|_{t=T} = 0$ and initial condition $m|_{t=0} = m_0$. This system constitutes the main object in the study of (open-loop) Mean-Field games from an analytic point of view, and it is called the *Mean-Field system*.

Studying this system not only should be less complicated than studying the one of N Hamilton–Jacobi equations, coupled with the N stochastic differential equations for the states, but also allows to recover pieces of information on the N-player game when N is large (but finite). One important result that we mention is that if one solves the Mean-Field system and, using the value function u, builds up (open-loop) controls $\alpha^1, \dots, \alpha^N$ as we did before to be used by the players of the corresponding N-player game, then such controls are *almost optimal*. We say that they form an *ϵ-Nash equilibrium*, in the sense that, for some $\epsilon > 0$,

$$J^i(\alpha^{-i}, \alpha^i) \leq J^i(\alpha^{-i}, \beta) + \epsilon \qquad \forall\, i \in \{1, \dots, N\},\ \forall\, \beta \in \mathcal{A}.$$

This reads as follows: if the strategies α^i are applied, then *each player, by unilaterally changing its strategy, cannot have an advantage greater than ϵ* (that is, *cannot reduce its cost of more than ϵ*). Then, the "almost optimality" of the strategies built from the Mean-Field system consists in the fact that $\epsilon \to 0$ as $N \to \infty$.

3.2 Closed-loop case: the Master Equation

Things become more complicated when one tries to pass to the limit the Nash system for closed-loop strategies (system (2.5)) because of the presence of the sum over j, which gathers more and more terms as N grows.

Consider the following heuristic argument. In the Mean-Field framework, the impact of the j-th player on the i-th one ($i \neq j$) goes like $1/N$; this should be reflected in u^i depending on x^j in a weighted manner, and one expects $\partial_{x^j} u^i$ to go like $1/N$ as well. On the other hand, there's no reason why $\partial_{x^i} u^i$ should vanish as $N \to \infty$, so one can only say that it stays bounded. Then the sum over j consists in N terms that go like $1/N$: we can say it stays bounded as well, but does it converge? Do we have the *compactness* needed for a bounded sequence to have a convergent subsequence? If so, what does the sum converge to? This can be said to be the main question that arises when trying to identify a limit problem associated to the Nash system.

The correct guess turns out to be that u^i should converge to some function U (common to all players by symmetry) that is defined on a *space of measures*;[3] that is, for N large,

$$u^i(t,x) \approx U(t, x^i, m_x^{N,i}), \qquad \text{with} \quad m_x^{N,i} \overset{\text{def}}{=} \frac{1}{N-1} \sum_{\substack{1 \leq j \leq N \\ j \neq i}} \delta_{x^j},$$

δ_y being the Dirac delta distribution centred at y. With this idea, P. Cardaliaguet, F. Delarue, Lasry and Lions, exploiting a theory of *derivatives in the space of measures*, argued that in this situation

$$\partial_{x^j} u^i(t,x) \approx \frac{1}{N-1}\, D_m U(t, x^i, m_x^{N,i}, x^j),$$

[3] N.B. *Space of measures*, not *measure space*!

where $D_m U(t,x,m,y)$ is the so-called *intrinsic derivative* of $U(t,x,m)$ with respect to the measure m. So, considering that

$$\frac{1}{N-1}\sum_{\substack{1\le j\le N\\ j\neq i}} D_m U(t,x^i,m_x^{N,i},x^j) = \int_{\mathbb{R}} D_m U(t,x^i,m_x^{N,i},y)\,\mathrm{d}m_x^{N,i}(y),$$

one argues that

$$\sum_{j\neq i} \partial_p H(t,x,\partial_{x^j}u^j(t,x))\cdot\partial_{x^j}u^i(t,x) \\ \approx \int_{\mathbb{R}} \partial_p H(t,y,\partial_x U(t,y,m_x^{N,i}))\cdot D_m U(t,x^i,m_x^{N,i},y)\,\mathrm{d}m_x^{N,i}(y),$$

ignoring the difference between $m_x^{N,i}$ and $m_x^{N,j}$ which is expected to be negligible.

The conclusion at which we arrive is then that the Nash system should have for limit the following *Master Equation*:

$$-\partial_t U(t,x,m) + H(t,x,m,\partial_x U(t,x,m)) \\ + \int_{\mathbb{R}} \partial_p H(t,y,\partial_x U(t,y,m))\cdot D_m U(t,x,m,y)\,\mathrm{d}m(y) = 0,$$

for $(t,x,m)\in[0,T]\times\mathbb{R}\times\mathcal{P}(\mathbb{R})$. We mention that, as for the Mean-Field system, if one deduces the Master Equation starting from stochastic differential equations for the states rather than ordinary ones, then additional terms involving second-order derivatives appear. The unknown $U(t,x,m)$ can be interpreted as the minimal cost in the Mean-Field problem of a small player at time t in position x if the distribution of the other players is m.

This heuristic argument is not intended to simplify some technicalities that we wish to avoid in this presentation, rather making this heuristics rigorous is itself one of the intrinsic difficulties in this study of closed-loop Nash equilibria. As claimed by Cardialiaguet, Delarue, Lasry and Lions in the preface of [1],

> '*it seems especially difficult to get any a priori estimate that could be helpful for passing to the limit by means of a compactness argument.*'

So, the 'short cut' they use in order to 'bypass any detailed study of the Nash system' is to focus directly on the expected limit, the Master Equation. This can be considered as a "*top-down*" approach, as a proof of the convergence of the solution to the Nash system towards the solution to the Master Equation is then obtained *a posteriori*, once one knows to have a unique solution to the Master Equation that is sufficiently *smooth* in order to perform certain desired computations.

It is exactly starting from the displayed quotation that I wish to introduce a part of my work during the PhD, as my supervisor M. Cirant and I attempted to walk the road *bottom-up* from the Nash system to a limit form of it like the Master Equation. Therefore, the first step was to get those difficult *a priori estimates*, which essentially consist in showing that the Mean-Field interactions are reflected in the behaviour

$$\partial_{x^j}u^i \lesssim \frac{1}{N} \quad \forall\, j\neq i \tag{3.4}$$

of the derivatives of the value functions.

In addition, we tried to go beyond the classic setting of Mean-Field Games that I presented, by looking at what happens when one drops either one of the conditions that characterise the classic mean-field framework, namely *symmetry* and *negligibility of each single interaction*.

4 What if the number of players grows? (non-Mean-Field regime)

So, let me conclude with a basic discussion about some results that we've obtained.

4.1 Non-symmetric interactions

If the interactions are no longer symmetric, we cannot expect all the players to be described by a single representative one, not even in the limit. This in a sense corresponds to the lack of the "identically distributed" part in the requirements for the Law of Large Numbers to be applied, even if with some sort of propagation of chaos argument one can get the "independent" part in the limit.

In this case, under suitable structural assumptions, we are able to prove the aforementioned a priori estimates for a Nash system with Hamiltonians of the form

$$H^i(x, p^i) = \tfrac{1}{2}|p^i|^2 - f^i(x)$$

and then to pass to the limit via a compactness argument. We obtain the following limit equation:

$$U^\lambda(t, x, \mu_t) = \int_t^T \int_{\mathbb{R}} \left(\tfrac{1}{2}|\partial_x U^\lambda(s, y, \mu_s)|^2 + f^\lambda(y, \pi_{\mathbb{R}\sharp}\mu_s) \right) \mathrm{d}\bar{m}_s^\lambda(y)\, \mathrm{d}s.$$

The parameter λ varies in a compact Λ and takes account of the lack of symmetry, in the sense that it basically parametrises on Λ the infinite population at the limit; μ is a flow of probability measures that evolves according to a partial differential equation of Fokker–Planck type, but it is a measure on $\Lambda \times \mathbb{R}$ and not just on $\mathbb{R}$ as in the symmetric Mean-Field case; finally, $\bar{m}^\lambda$ is another flow of probability measures, governed by another Fokker–Planck equation, that can be interpreted as the equilibrium distribution of the limit player identified by λ, provided that it starts the game in position x.

This equation for the limit value function U can be considered as a *weak formulation of a Master Equation* associated to a non-symmetric Mean-Field game. Essentially, it is a Master Equation with an additional parameter λ, a measure supported on an "enlarged" space, and that's been integrated along an optimal trajectory for the limit game. Naively speaking, if U was sufficiently smooth then one could differentiate our equation to remove such an integration, thus obtaining exactly the Master Equation one expects in its "differential form".[(4)]

[(4)] N.B. I'm not meaning the mathematical object called *differential form*. ☺

4.2 Non-negligible interactions

A completely different situation is instead that of interactions which are possibly symmetric but which we call of *non-Mean-Field* type in that the impact of each player on the others doesn't go like $1/N$ and can remain *non-negligible* (for example, of order 1) also when the players become infinitely many.

We studied this case when there is an underling graph structure that establishes how players are "connected" in such a way that the weight of each player's behaviour on the strategies of the others depends on its *graph distance* from them; that is, in this setting, the farther a player is "from me" the less I will be affected by its behaviour. So, given a player, there will always be players that strongly affect its strategy, no matter the total number of players.

With suitable hypotheses on how the strength of the interactions decays with the distance, we've been able to prove some results that are related to what we named *unimportance of distant players*. By this we mean that players that are far from a given one are less important when it comes to determine the optimal behaviour of that given player, and this fact is reflected in basically two implications.

The first one, that we've obtained in a framework of open-loop equilibria for a finite number of players, is that if one's interested in the equilibrium distribution of a player, then this is well approximated by the equilibrium distribution of that player when one considers instead a "truncated" game constructed by completely ignoring players that are sufficiently far from it. So, this can be viewed as a way to reduce the complexity of the system one has to study by going in the opposite direction of Mean-Field Games; that is, instead of exploiting the presence of some aggregate quantity that affects the strategies and describes a limit system, we consider that one can "cut out" from the game distant players, with just a small error.

Let me mention that though this could seem a trivial fact, actually it isn't. Even if "I" don't see players that are too far from me, I see those who are nearer, who in turn see those who are close to them and so on, so that I "indirectly" see all other players; hence, in order for the far players to be negligible, we must admit that there's some kind of "finite propagation speed of pieces of information". This is a nontrivial fact that strongly depends on additional structural hypotheses of the interactions; indeed, there are examples when, if the duration of the game is long enough, then even the impact of distant players can't be ignored, and, on the other hand, we also have a result showing that distant players can have little weight for an arbitrarily long duration of a game.

This last fact is actually shown by means of the second implication of the unimportance of distant players, which is the following. Consider a closed-loop game; suppose that players are index by $i \in \mathbb{Z}$ and the graph distance of two players i and j is given by $|i-j|$. Then, under appropriate assumptions, the sup-norms of the derivatives $\partial_{x^j} u^i$ of the value functions vanish as $|i-j| \to \infty$, so rapidly that

$$\sum_j \left\| \partial_{x^j} u^i \right\|_\infty \quad \text{converges.}$$

This is the counterpart of (3.4) that reflects the negligibility of distant players only.

In addition, since we have summability of that series, we can write an *infinite-dimensional Nash system*. It will have the form of the standard Nash system, but the sum over j will actually be a series and the variable x will be a vector with infinitely many coordinates. So, instead of having "$\mathbb{R}^N \to \mathcal{P}(\mathbb{R})$" as $N \to \infty$ like for the Master Equation, we have "$\mathbb{R}^N \to \mathbb{R}^\infty$".

This second kind of non-Mean-Field games seems the hardest to deal with in general. A reason one could think about is that there seems to be no way not to study the Nash system: even if taking the limit for infinitely many players can in certain situations give rise to some additional exploitable structure (which I won't say more about), one still needs to work with equations in $\mathbb{R}^\infty$, to which finite-dimensional results can be adapted only if they are independent of the dimension. This in the end essentially begs for estimates on the N-player Nash system not dissimilar from the 'especially difficult' ones we've talked about and possibly even more precise...

References

[1] P. Cardaliaguet, F. Delarue, J.-M. Lasry, and P.-L. Lions, "The master equation and the convergence problem in mean field games". Volume 201 of Annals of Mathematics Studies. Princeton University Press, Princeton, NJ, 2019.

[2] P. Cardaliaguet and A. Porretta, *An introduction to mean field game theory.* In "Mean field games", volume 2281 of Lecture Notes in Math. Springer, Cham (2020), 1–158.

[3] R. Carmona and F. Delarue, "Probabilistic theory of mean field games with applications. I". Volume 83 of Probability Theory and Stochastic Modelling. Springer, Cham, 2018. Mean field FBSDEs, control, and games.

[4] M. Cirant and D.F. Redaelli, *A priori estimates and large population limits for some non-symmetric Nash systems with semimonotonicity.* In preparation (2024).

[5] M. Cirant and D.F. Redaelli, *Some remarks on linear-quadratic closed-loop games with many players.* Submitted (2024).

[6] M. Huang, R.P. Malhamé, and P.E. Caines, *Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle.* Commun. Inf. Syst. 6/3 (2006), 221–251.

[7] J.-M. Lasry and P.-L. Lions, *Mean field games.* Jpn. J. Math. 2/1 (2007), 229–260.

[8] D.F. Redaelli, *Short-time well-posedness of an infinite-dimensional Nash system arising from differential games with nearsighted interactions.* In preparation (2024).

Numerical Solution of Wave Propagation Phenomena in Viscoelastic Materials

Nicolò Crescenzio (*)

Abstract. Many materials, such as plastics, wood, concrete and metals at high temperatures, exhibit a mechanical behaviour that is intermediate between the elastic and the viscous one. Consequently, these materials cannot be adequately described using the well-known classical theories of elasticity and viscosity and it is therefore necessary to consider a more general theory that is capable of modelling the behaviour of these materials, also known as viscoelastic materials. In the first part of the talk, we will provide a brief overview of the theory of linear viscoelasticity, with a particular focus on the so-called Kelvin-Voigt rheology. Then, we will discuss the problem of viscoelastic wave propagation phenomena in a Kelvin-Voigt material and show numerical results obtained by means of a Galerkin spectral approach.

1 Introduction

Many materials are characterized by the fact they exhibit a mechanical behavior that cannot be adequately described using the well-known classical theories of elasticity and viscosity. Consequently, it is necessary to develop a more general theory that is capable of modelling the behavior of these materials. In particular, we consider the class of the so-called *viscoelastic materials*, which possess both an elastic component and a viscous component and therefore exhibit a behavior that is itermediate between the elastic and viscous one. Among the materials showing viscoelastic behavior there are plastics, wood, natural and synthetic fibers, concrete and metals at elevated temperatures. The theory of viscoelasticity has been developed starting from the nineteenth century with the contribution of many eminent scientists, such as Maxwell [1], Voigt [2], Boltzmann [3] and Volterra [4,5]. The reasearch in this field has further increased in the twentieth century, when synthetic polymers have started to be engineered and used in large scale in a variety of applications.

The viscoelasticity theory that we will discuss here is mainly based on the comprehensive surveys written by [6-10] and it holds only for infinitesimal displacements and deformations. In this framework, the constitutive equation to be sought is the one relating

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: nicolo.crescenzio@math.unipd.it . Seminar held on 23 May 2024.

the stress $\boldsymbol{\sigma}$ and the strain $\boldsymbol{\varepsilon}$. In particular, this relationship is assumed to be linear, i.e., the stress will be given as a linear functional of the strain. For this reason, the theory that we shall present is also referred to as *linear viscoelasticity theory*. This theory affirms that the current value of the stess depends not only upon the present value of the strain, but also on their complete past history. More formally, this can be expressed through the following integral expression

$$\boldsymbol{\sigma}(t) = G(0)\boldsymbol{\varepsilon}(t) + \int_0^t \boldsymbol{\varepsilon}(t-s)\frac{dG(s)}{ds}\,ds. \tag{1}$$

The function $G(t)$, which describes the viscoelastic properties of the material, is called *relaxation function*. Integrals over the history of strain, as the one in (1), are sometimes referred to as hereditary integrals and materials whose contitutive relations contain such integrals are described as having memory. Moreover, notice that, if the hereditary integral is not present in Equation (1), then the constitutive relation reduces to the well known Hooke's Law defining elastic materials. An alternative form of the stress-strain relationship can be obtained by reversing the role of the stress and the strain, namely, the current strain is now determined by the current value and past history of stress:

$$\boldsymbol{\varepsilon}(t) = J(t)\boldsymbol{\sigma}(0) + \int_0^t J(t-\tau)\frac{d\boldsymbol{\sigma}(t)}{d\tau}\,d\tau$$

where the function $J(t)$ is called *creep function*.

It is worth highlighting that the theory of linear viscoelasticity discussed here represents a simplified model based on specific assumptions and, as such, it cannot be always employed. Consequently, in situations where the linear constitutive equations will yield only a poor approximation of the actual behaviour of the material, a more accurate description for viscoelastic materials may be obtained by considering a nonlinear viscoelastic theory. Furthermore, in the current discussion we also neglect the thermal effects, but it is worth noting that these aspectes might be necessary to be taken into account as well depending on the problem. For the presentation of a far more complex theory of viscoelasticity, which considers both the nonlinear and thermal contributions, we refer to, e.g., [9, 11-16].

2 Viscoelastic Fluids and Solids

So far, our discussion has been confined to the realm of linear viscoelastic materials, without specifying whether they are solids or fluids. This is because the viscoelastic behaviour can manifest in both solid and fluid materials. The distinction between viscoelastic solids and viscoelastic fluids can be seen by performing two simple laboratory tests, that are *relaxation of stress* and *creep*.

2.1 Relaxation of Stress

Let us consider a sudden constant shear strain $\boldsymbol{\varepsilon}$ applied starting from time $t = 0$ to an undeformed body, i.e., $\boldsymbol{\varepsilon}(t) = \boldsymbol{\varepsilon}_0 H(t)$, where $H(t)$ is the Heaviside step function. In this

case it can be shown that the resulting stress is given by

$$\boldsymbol{\sigma}(t) = G(t)\boldsymbol{\varepsilon}_0.$$

Experimental observations have shown that the initial stress is high and then it relaxes over time to some final value. Specifically, the initial value is due the elastic reponse of the material, whereas the relaxation is due to the viscous effects. From a physical viewpoint, the high initial resistance to deformation is due to frictional forces opposing the reorganization of molecules required by the change of shape. These internal forces are gradually overcome leading to what is called *relaxation* of stress. Concerning the behaviour of the relaxation function at large times, i.e. $\lim_{t\to\infty} G(t)$, it has been observed that for a viscoelastic solid material

$$\lim_{t\to\infty} G(t) = c > 0,$$

while for a viscoelastic fluid material

$$\lim_{t\to\infty} G(t) = 0.$$

The qualitative behavior of $G(t)$ for both viscoelastic solids and fluids is shown in Figure 1.

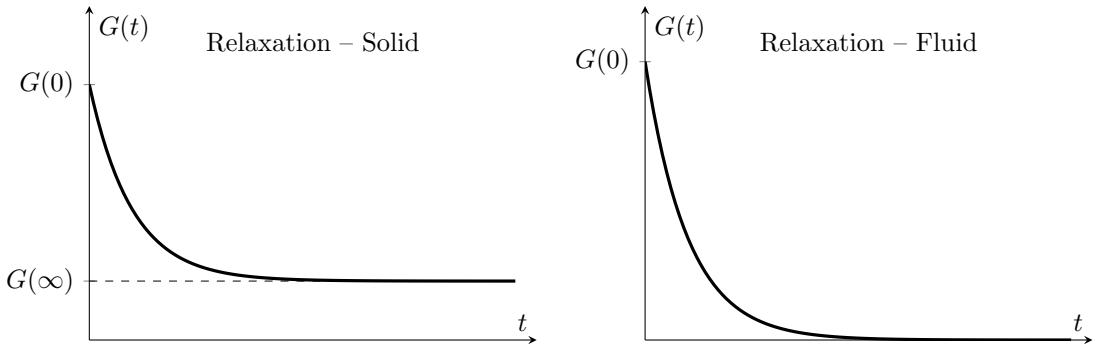


Figure 1: Qualitative behavior of the relaxation function $G(t)$ for viscoelastic solid materials (left panel) and viscoelastic fluid materials (right panel). In both cases the function shows a decreasing behavior over time, approaching a final value that is non-zero for solids and zero for fluids.

2.2 Creep

Let us now consider the behavior of a material subject to a constant shear stress $\boldsymbol{\sigma}$ starting from time $t = 0$, i.e., $\boldsymbol{\sigma}(t) = H(t)\boldsymbol{\sigma}_0$, where $H(t)$ is the Heaviside step function. In this case, the strain-stress relation has the form

$$\boldsymbol{\varepsilon}(t) = J(t)\boldsymbol{\sigma}_0,$$

which shows that the resultant strain is not instantaneous but develops over time. It has been observed experimentally that a suddenly imposed stress causes a certain instantaneous

strain, which is due to the elastic response of the material and is given by $J(0)\boldsymbol{\sigma}$ in the equation above. Subsequently, the strain increases over time to some final value (for a solid) or indefinitely (for a liquid). This material behavior is named *creep*. Having defined

$$J(\infty) := \lim_{t\to\infty} J(t), \qquad \dot{J}(\infty) := \lim_{t\to\infty} \dot{J}(t)$$

we see that, for a solid material, saying that creep ceases is equivalent to $J(\infty)$ being finite (see Figure 2, top left panel) and in this case $J(\infty)$ may be regarded as a natural generalization of the inverse of the elastic modulus (i.e., the compliance). Moreover, for viscoelastic solid materials $\dot{J}(\infty)$ is zero. On the other hand, for viscoelastic fluids creep continues indefinitely and $\dot{J}(\infty)$ is a non-zero constant (see Figure 2, top right panel).

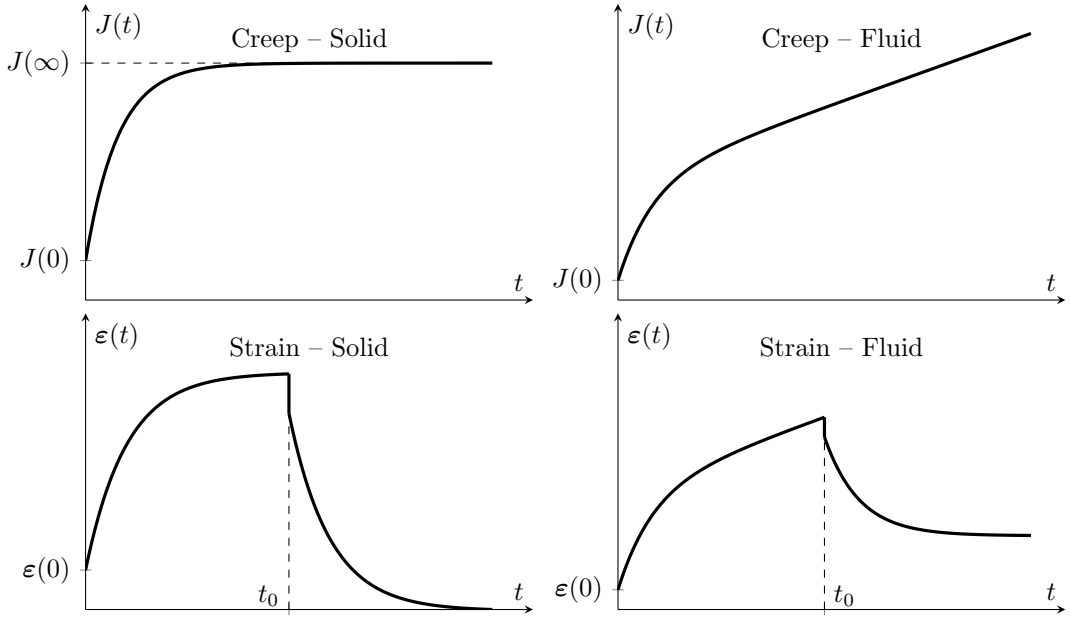


Figure 2: Top row shows the qualitative behavior of the creep function. In viscoelastic solid materials (left) the creep function increases over time and then reaches a finite value, whereas in viscoelastic fluid materials (right) it increases indefinitely. Bottom row shows the strain response to a constant stress applied for time values in the interval $[0, t_0]$: solid materials recover completely (left), while fluids are characterized by a permanent deformation (right).

Now consider a material that is subject to a constant shear stress $\boldsymbol{\sigma}_0$ only for a limited amount of time, say for $t \in [0, t_0]$, after which it is released. In this case, experiments have shown that, after $t = t_0$, the strain immediately decreases (elastic recovery) and it then continuous to gradually decrease over time (anelastic recovery) until a residual strain is reached. Specifically, the strain, at large values of time, can be approximated as follows

$$\boldsymbol{\varepsilon}(t) \approx \dot{J}(t) t_0 \boldsymbol{\sigma}_0.$$

This relation clearly shows that if $\dot{J}(t)$ tends to zero, then $\boldsymbol{\varepsilon}(t)$ vanishes. This is the expected behavior of a viscoelatic solid materials that has the capability to recover completely from applied stresses, like an elastic body does, but with a time delay due to internal losses (see Figure 2, bottom left panel). On the other hand, if $\dot{J}(t)$ is finite, then the permanent deformation occurs as a result of the application and removal of stress and this justifies the term viscoelastic fluid (see Figure 2, bottom right panel). However, even viscoelastic fluid materials partially exhibit recovery, since the creep function has the form

$$J(t) = \tilde{J}(t) + \frac{t}{2\eta}$$

where $\tilde{J}(t)$ approaches a finite positive asymptote for large values of time and η is a positive constant called *zero shear rate viscosity*.

3 Dynamic Loading and Complex Modulus

In addition to the creep and stress relaxation tests, a dynamic test might be useful when studying the behavior of viscoelastic materials. The dynamic test consists in prescribing a cyclic history of strain given by

$$\boldsymbol{\varepsilon}(t) = \boldsymbol{\varepsilon}_0 \cos(\omega t), \tag{2}$$

where $\boldsymbol{\varepsilon}_0$ is the amplitude and ω is the frequency of oscillation. The stress response as a function of time t depends on the characteristics of the material. Specifically, for an elastic solid material, the stress is proportional to the strain, i.e., $\boldsymbol{\sigma}(t) = E\boldsymbol{\varepsilon}(t)$. Therefore, when the strain is defined as in (2), the stress response of an elastic material it is given by

$$\boldsymbol{\sigma}(t) = E\boldsymbol{\varepsilon}_0 \cos(\omega t),$$

which implies that the stress response caused by the strain is immediate. In this case, we also say that the stress is in phase with the strain. For a viscous material, the stress is proportional to the strain rate, i.e., $\boldsymbol{\sigma}(t) = \eta\dot{\boldsymbol{\varepsilon}}(t)$. Thus, for a strain history defined as in (2), the stress response has the form:

$$\boldsymbol{\sigma}(t) = -\eta\boldsymbol{\varepsilon}_0\omega \sin(\omega t) = \eta\boldsymbol{\varepsilon}_0\omega \cos\left(\omega t + \frac{\pi}{2}\right),$$

which shows that the stress is out of phase with the strain and that, in particular, the strain is behind the stress by a 90 degree phase lag. In a linear viscoelastic material, laboratory experiments have shown that the stress as a function of time appears complicated in the very first few cycles but it eventually reaches a steady-state condition in which the resulting stress is also sinusoidal, having the same angular frequency ω but retarded in phase by an angle δ. Moreover, the same result has been observed even when the controlled variable was the stress and not the strain. So, we write the stress response function as:

$$\boldsymbol{\sigma}(t) = \boldsymbol{\sigma}_0 \sin(\omega t + \delta), \tag{3}$$

where $\delta \in [0, \pi/2]$ and $\boldsymbol{\sigma}_0 = \boldsymbol{\sigma}_0(\omega)$. Writing equation (3) in the equivalent form

$$\boldsymbol{\sigma}(t) = \boldsymbol{\sigma}_0 \cos\delta \cos(\omega t) - \boldsymbol{\sigma}_0 \sin\delta \sin(\omega t),$$

it is clear that the stress response is the sum of an in-phase and an out-of-pase response. It is often more useful to consider a strain history specified by the following complex function of time

$$\boldsymbol{\varepsilon}(t) = \boldsymbol{\varepsilon}_0 \exp(i\omega t)$$

and a complex response stress function having the form

$$\boldsymbol{\sigma}(t) = \boldsymbol{\sigma}_0 \exp(i\omega t + i\delta),$$

which can also be written as

$$\boldsymbol{\sigma}(t) = M(\omega)\boldsymbol{\varepsilon}_0 \exp(i\omega t), \qquad M(\omega) = \frac{\boldsymbol{\sigma}_0}{\boldsymbol{\varepsilon}_0} \exp(i\delta).$$

The function $M(\omega)$ is called *complex modulus* and it is a complex function of the frequency whose real and imaginary part are often referred to as the *storage* and *loss* moduli, respectively. The names are motivated by the fact that the storage modulus is associated with energy storage and release during periodic deformation, whereas the loss modulus is associated with the dissipation of energy and its transformation into heat. Finally, the phase angle δ is also called *loss angle*.

4 The Kelvin-Voigt Constitutive Model

Hereafter we will focus on a particular type of linear viscoelastic consistitutive model, the so-called *Kelvin-Voigt model*. The aim of this section is to introduce this model, starting with the one-dimensional case and then presenting the extension to the multi-dimensional case.

4.1 One-dimensional case

In the one-dimensional case, linear viscoelastic constitutive equations can be defined in terms of mechanical models consisting of two basic elements, weightless springs and dashpots, that are connected in series and parallel as in electrical circuits. In particular, the *spring*, whose stress-strain relation is given by $\boldsymbol{\sigma} = E\boldsymbol{\varepsilon}$, clearly describes the elastic property of the masterial, as it is able to undergo an instantaneous elastic strain when loaded, to mantain that strain as long as the load is applied and then to undergo an instantaneous de-straining when the load is removed. A *dashpot* is a positon-cylinder mechanism filled with a viscous liquid and characterized by a stress-strain relation of the form $\boldsymbol{\sigma} = \eta\dot{\boldsymbol{\varepsilon}}$, where η is the dynamic viscosity of the fluid in the cylinder. A strain in a dashpot is achieved by dragging the piston through the fluid. When suddenly applying a constant load, the strain is seen to increase linearly as long as the stress is applied. There is no instantaneous movement of the dashpot at the onset of load, as it takes time for the strain to build up. Then, when the load is removed, there is no stress to move the piston back throug the

fluid and therefore any strain built up is permanent. On the other hand, when a constant strain $\boldsymbol{\varepsilon}_0$ is imposed on the dashpot, the stress is given by $\boldsymbol{\sigma} = \eta\boldsymbol{\varepsilon}_0\delta(t)$. However, since an infinite stress is impossible in reality, it is therefore impossible to impose an instantaneous finite deformation on the dashpot.

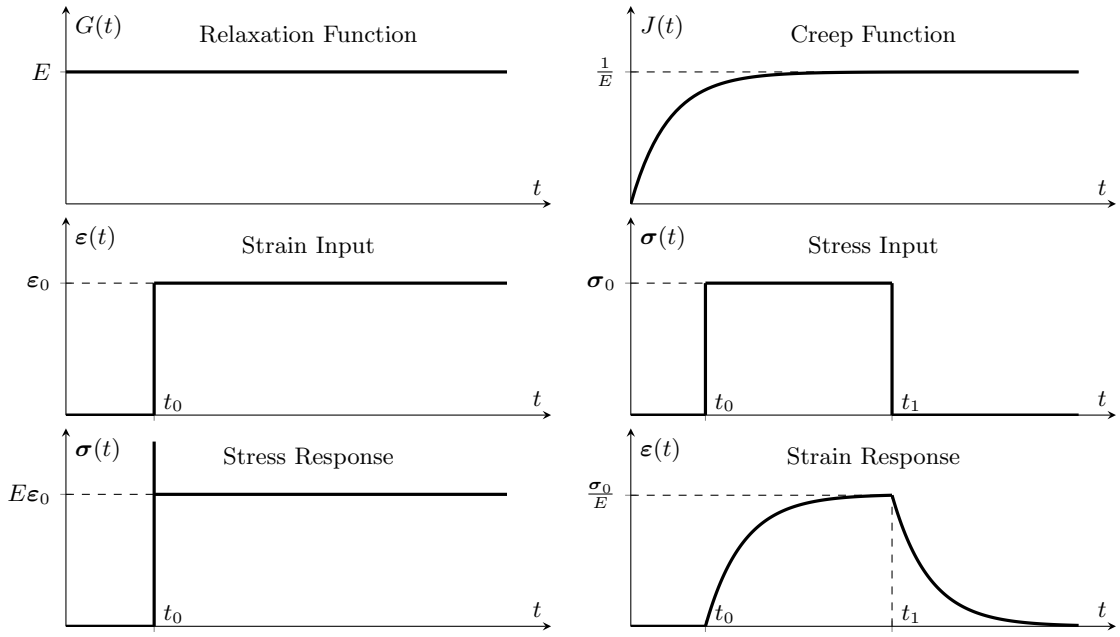


Figure 3: Kelvin-Voigt model for viscoelastic materials. Left column shows the behavior of the relaxation function (top row), the constant strain applied starting from time $t = t_0$ in a relaxation experiment (middle row) and the stress response to the applied constant strain (bottom row). Right column shows the creep function (top row), the constant stress applied in a creep experiment starting from time $t = t_0$ and mantained until $t = t_1$ (middle row) and the strain response to the constant applied stress (bottom row).

One of the simplest and most-known mechanical models capable of describing the mechanical properties of viscoelastic materials is the Kelvin-Voigt model, which consists of a spring and a dashpot connected in parallel. In this model the total stress $\boldsymbol{\sigma}$ is composed by an elastic stress

$$\boldsymbol{\sigma}_1 = E\boldsymbol{\varepsilon}$$

and a viscous stress

$$\boldsymbol{\sigma}_2 = \eta\dot{\boldsymbol{\varepsilon}}.$$

In this case, the stress-strain constitutive relation is

$$\boldsymbol{\sigma} = \boldsymbol{\sigma}_1 + \boldsymbol{\sigma}_2 = E\boldsymbol{\varepsilon} + \eta\frac{d\boldsymbol{\varepsilon}}{dt}.$$

and it can be shown that the relaxation function is given by

$$G(t) = EH(t) + \eta\delta(t)$$

whereas the creep function has the form

$$J(t) = \frac{1}{E}\left(1 - \exp\left(-\frac{t}{\tau}\right)\right) H(t),$$

where $\tau = \eta/E$ is also called *retardation time*. The relaxation and creep functions are represented in Figure 3, top row, and they clearly suggest that the Kelvin-Voigt model is more appropriate to represent viscoelastic solid materials. In particular, the relaxation function does not show any time dependence, as in the case of pure elastic solids. Moreover, the presence of the delta function implies that, in practice, it is impossible to impose an instantaneous strain on the medium, as it would require an infinite stress to be applied (see also Figure 3, bottom left). In other words this means that there is no instantaneous response to a suddenly applied finite stress. In the creep experiment, the spring would want to stretch, but it is held back by the dashpot, which cannot react instantaneously and then takes all the stress. For this reason, the creep function does not present an instantaneous strain. Subsequently, the dashpot extends and begins to transfer the stress to the spring which starts to strain at decreasing rate. At infinite time, the entire stress is on the spring and the strain approaches an asymptotical value. When the Kelvin-Voigt model is unloaded, the spring will want to contract but again the dashpot hold it back. However, the spring eventually gradually relaxes to its undeformed state and full recovery occurs (see Figure 3, bottom right).

4.2 Multi-dimensional case

The multi-dimensional generalization of the Kelvin-Voigt viscoelastic constitutive model has been analysed for the first time by Carcione and co-workers in [17]. As in the one-dimensional case, the stress tensor is expressed by the sum of an elastic and a viscous contribution, i.e.,

$$\boldsymbol{\sigma}(\boldsymbol{u}, \dot{\boldsymbol{u}}) = \boldsymbol{\sigma}_{el}(\boldsymbol{u}) + \boldsymbol{\sigma}_{vi}(\dot{\boldsymbol{u}})$$

where the elastic and viscous stress tensors are defined as

$$\boldsymbol{\sigma}_{el}(\boldsymbol{u}) = \lambda_{el}\mathrm{Tr}(\boldsymbol{\varepsilon}(\boldsymbol{u}))\mathbb{I} + 2\mu_{el}\boldsymbol{\varepsilon}(\boldsymbol{u}), \qquad \boldsymbol{\sigma}_{vi}(\dot{\boldsymbol{u}}) = \lambda_{vi}\mathrm{Tr}(\boldsymbol{\varepsilon}(\dot{\boldsymbol{u}}))\mathbb{I} + 2\mu_{vi}\boldsymbol{\varepsilon}(\dot{\boldsymbol{u}}).$$

The pairs of paramters λ_{el}, μ_{el} and λ_{vi}, μ_{vi} are called the elastic and viscous Lamè parameters, respectively, and they define the mechanical properties of the materials. Hereafter, we will assume these parameters to be constants.

5 Wave Propagation on Viscoelastic Materials

The propagation of waves in both elastic and viscoelastic materials is governed by the so-called Cauchy equilibrium equations:

$$\rho\ddot{\boldsymbol{u}} = \mathrm{div}\boldsymbol{\sigma}(\boldsymbol{u})$$

where ρ is the density of the material, $\boldsymbol{u}$ is the displacement field and $\boldsymbol{\sigma}$ is the stress tensor. This vector Partial Differential Equation describes the conservation of linear momentum in any continuum, either fluid or solid. The displacement equations of motion associated to a particular material are then obtained by specifing the constitutive law that relates the stress $\boldsymbol{\sigma}$ to the displacement field $\boldsymbol{u}$. We now briefly discuss the main aspects of wave propagation in a homogeneous and isotropic viscoelastic material and, in particular, we start again by considering the one-dimensional case, since it is easier to introduce and its generalization to the multi-dimensional case is straightforward.

Let us consider a one-dimensional displacement wave having the following expression

$$u(t,x) = u_0 \exp\left(i\omega t - ikx\right)$$

where $k \in \mathbb{C}$ is the complex wavenumber and $\omega \in \mathbb{R}$ is the angular frequency. Substituting this expression into Cauchy equation and using the definition of complex modulus $M(\omega)$, it can be shown that the following dispersion relation holds:

$$M(\omega)k^2 = \rho\omega^2.$$

This relation provides the following expression of the complex velocity

$$v_c(\omega) = \frac{\omega}{k} = \sqrt{\frac{M(\omega)}{\rho}}.$$

Now, writing the complex wavenumber as $k = \kappa - i\alpha$, where $\kappa, \alpha \in \mathbb{R}_+$, we can rewrite the expression for the plane displacement wave as follows

$$u(t,x) = u_0 \exp(-\alpha x) \exp(i\omega t - i\kappa x).$$

This shows that α is responsible for the change in amplitude of the wave as it propagates in a viscoelastic material. For this reason α is also called *attenuation factor* and it can be computed as follows:

$$\alpha(\omega) = -\omega \operatorname{Im}\left(\frac{1}{v_c}\right). \tag{4}$$

Moreover, the real wavenumber κ allows to define the *phase velocity* as

$$v_p(\omega) = \frac{\omega}{\kappa} = \left[\operatorname{Re}\left(\frac{1}{v_c}\right)\right]^{-1}. \tag{5}$$

Finally, another parameter that allows to quantify dissipation is the *quality factor* Q, as its inverse, Q^{-1}, is the *dissipation factor*. The quality factor is defined as twice the time-averaged strain-energy density divided by the time-averaged dissipated-energy density and it can be computed as

$$Q = \frac{\operatorname{Re}(v_c^2)}{\operatorname{Im}(v_c^2)} \tag{6}$$

We now consider the specific case of a one-dimensional Kelvin-Voigt viscoelastic material and derive the expressions of some of the quantities defined above. It can be shown that the complex modulus $M(\omega)$ is given by

$$M(\omega) = E + i\omega\eta$$

where it is easy to see that the real and imaginary part are only related to the elastic and viscous parameters, respectively. Then, using Equation (5), we obatin the following expression of the phase velocity

$$v_p(\omega) = \sqrt{\frac{E}{\rho}} \frac{\sqrt{1 + (\omega\eta/E)^2}}{\sqrt{\frac{1}{2}\left(\sqrt{1 + (\omega\eta/E)^2} + 1\right)}}.$$

This shows that $v_p \to \sqrt{E/\rho}$ for $\omega \to 0$ and that $v_p \to \infty$ for $\omega \to \infty$, which means that the elastic (lossless) velocity is obtained at the low-frequency limit whereas high frequencies propagate with infinite velocity. The qualitative behaviour of the phase velocity for a Kelvin-Voigt material is shown in Figure 4 left panel. The attenuation factor can be computed using Equation (4) and it is given by

$$\alpha(\omega) = \omega \frac{\sqrt{\frac{1}{2}\left(\sqrt{(E/\rho)^2 + (\omega\eta/\rho)^2} - E/\rho\right)}}{\sqrt{(E/\rho)^2 + (\omega\eta/\rho)^2}}$$

which shows that, in a Kelvin-Voigt material, the higher the angular frequency of the propagating wave, the greater the attenuation of its amplitude in space (see Figure 4, right panel). Finally, using Equation (6), it is possible to see that the quality factor is given by

$$Q(\omega) = \frac{E}{\omega\eta}.$$

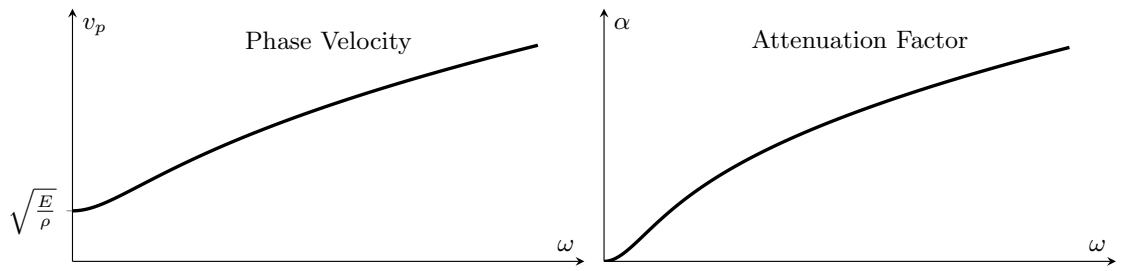


Figure 4: Qualitative behaviour of the phase velocity v_p (left panel) and the attenuation factor α (right panel) in a one-dimensional Kelvin-Voigt viscoelastic material.

All the definitions given so far can easily be extended to the multidimensional case, the only difference being that in the latter case there are two different types of waves propagating in a (visco)elastic material: the P-wave and the S-wave. In the case of P-waves, the deformation consists only of volume change with no rotations or change in shape and therefore the motion is characterized by compressions and dilatations along the direction of propagation. On the other hand, S-waves are characterized by a deformation consisting of shear and rotation only, with the dilatation or volume change being equal to zero. For a Kelvin-Voigt viscoelastic material, the P- and S-wave complex velocities are given by

$$Q_p = \frac{\lambda_{el} + 2\mu_{el}}{\omega(\lambda_{vi} + 2\mu_{vi})}, \qquad Q_s = \frac{\mu_{el}}{\omega\mu_{vi}}$$

Moreover, the P- and S-wave quality factors are defined as

$$Q_p = \frac{\lambda_{el} + 2\mu_{el}}{\omega(\lambda_{vi} + 2\mu_{vi})}, \qquad Q_s = \frac{\mu_{el}}{\omega\mu_{vi}}.$$

6 Numerical Simulations

In this section we present a Galerkin spectral approach for the computation of the numerical solution of wawe propagation phenomena in Kelvin-Voigt materials. We first introduce the initial boundary value problem and then discuss its weak formulation. Subsequently, we describe the Galerkin discretisation in space of the weak formulation and we conclude by showing the results obtained by using this numerical approach to solve a variant of Lamb's Problem.

6.1 Initial-Boundary Value Problem

Let us consider a domain $\Omega \subset \mathbb{R}^d$, where $d = 1, 2, 3$, with boundary $\partial\Omega \equiv \Gamma := \Gamma_D \cup \Gamma_N$, $\Gamma_N \cap \Gamma_D = \emptyset$, and denote with $\boldsymbol{n}$ the outward unit normal to Γ. We then consider homogeneous Neumann boundary conditions imposed on the Neumann boundary Γ_N and time-dependent Dirichlet boundary conditions on the Dirichlet boundary Γ_D. Therefore, the Initial-Boundary Value Problem can be written as:

$$\ddot{\boldsymbol{u}} = \operatorname{div}(\boldsymbol{\sigma}_{\text{el}}(\boldsymbol{u}) + \boldsymbol{\sigma}_{\text{vi}}(\dot{\boldsymbol{u}})) \qquad \text{in } (0,T] \times \Omega, \tag{7a}$$

$$\boldsymbol{u}(x,0) = \boldsymbol{u}_0(x)\ ; \quad \dot{\boldsymbol{u}}(x,0) = \boldsymbol{v}_0(x) \qquad \text{in } \Omega \text{ (initial conditions)}, \tag{7b}$$

$$[(\boldsymbol{\sigma}_{\text{el}} + \boldsymbol{\sigma}_{\text{vi}})\boldsymbol{n}](x,t) = 0 \qquad \text{in } (0,T] \times \Gamma_N \text{ (Neumann BCs)}, \tag{7c}$$

$$\boldsymbol{u}(x,t) = \boldsymbol{u}_D(x,t)\ ; \quad \dot{\boldsymbol{u}}(x,t) = \dot{\boldsymbol{u}}_D(x,t) \qquad \text{in } (0,T] \times \Gamma_D \text{ (Dirichlet BCs)}. \tag{7d}$$

6.2 Weak Formulation

Notice that the system (7) is complemented by non-homogeneous time-dependent Dirichlet boundary conditions. To handle this type of boundary conditions, we consider a linear and bounded operator, called *lifting operator* $\mathcal{L}$, that allows to construct a curve $t \mapsto \mathcal{L}\boldsymbol{u}_D$ such that the function $\mathcal{L}\boldsymbol{u}_D$, defined on the whole Ω, meets the boundary conditions at any time $t > 0$ and enjoys some regularity properties both in space and time. Then, we rewrite system (7) with respect to $\tilde{\boldsymbol{u}} := \boldsymbol{u} - \mathcal{L}\boldsymbol{u}_D$:

$$
(8) \quad \begin{cases} \ddot{\tilde{\boldsymbol{u}}} - \operatorname{div}(\boldsymbol{\sigma}_{\mathrm{el}}(\tilde{\boldsymbol{u}}) + \boldsymbol{\sigma}_{\mathrm{vi}}(\dot{\tilde{\boldsymbol{u}}})) = \operatorname{div}(\boldsymbol{\sigma}_{\mathrm{el}}(\mathcal{L}\boldsymbol{u}_D) + \boldsymbol{\sigma}_{\mathrm{vi}}(\mathcal{L}\dot{\boldsymbol{u}}_D)) - \mathcal{L}\ddot{\boldsymbol{u}}_D & \text{in } (0,T]\times\Omega \\ [(\boldsymbol{\sigma}_{\mathrm{el}}(\tilde{\boldsymbol{u}}) + \boldsymbol{\sigma}_{\mathrm{vi}}(\dot{\tilde{\boldsymbol{u}}}))\boldsymbol{n}](x,t) = 0 & \text{in } (0,T]\times\Gamma_N \\ \tilde{\boldsymbol{u}}(x,t) = 0, \quad \dot{\tilde{\boldsymbol{u}}}(x,t) = 0 & \text{in } (0,T]\times\Gamma_D \\ \tilde{\boldsymbol{u}}(x,0) = \boldsymbol{V}_0(x), \quad \dot{\tilde{\boldsymbol{u}}}(x,0) = \boldsymbol{U}_0(x) & \text{in } \Omega, \end{cases}
$$

where the initial conditions $\boldsymbol{U}_0$ and $\boldsymbol{V}_0$ are defined as follows:

$$
\boldsymbol{V}_0(x) := \boldsymbol{v}_0(x) - \mathcal{L}\dot{\boldsymbol{u}}_D(x,0), \qquad \boldsymbol{U}_0(x) := \boldsymbol{u}_0(x) - \mathcal{L}\boldsymbol{u}_D(x,0).
$$

Note that the right hand side of the first equation provides the following forcing term

$$
\boldsymbol{f}(x,t) := \operatorname{div}(\boldsymbol{\sigma}_{\mathrm{el}}(\mathcal{L}\boldsymbol{u}_D) + \boldsymbol{\sigma}_{\mathrm{vi}}(\mathcal{L}\dot{\boldsymbol{u}}_D)) - \mathcal{L}\ddot{\boldsymbol{u}}_D.
$$

It is clear that, if $\tilde{\boldsymbol{u}}$ solves (8), then $\boldsymbol{u} := \tilde{\boldsymbol{u}} + \mathcal{L}\boldsymbol{u}_D$ solves our original problem, i.e., the system of equations (7). To simplify the notation, hereafter we drop the tilda from system (8).

As it is customary in the treatment of second order linear equations such as (8), we can reduce our problem to a first order system of linear evolution equations by doubling the variables. More specifically, we introduce the variable $\boldsymbol{v}$ and impose the equality $\boldsymbol{v} = \dot{\boldsymbol{u}}$ in a weaker sense by exploiting the duality of $[H^1_D(\Omega)]^d$. To this aim, the system is rewritten as:

$$
(9) \quad \begin{cases} \dot{\boldsymbol{v}} - \operatorname{div}(\boldsymbol{\sigma}_{\mathrm{el}}(\boldsymbol{u}) + \boldsymbol{\sigma}_{\mathrm{vi}}(\boldsymbol{v})) = \boldsymbol{f} & \text{in } (0,T]\times\Omega \\ -\operatorname{div}(\boldsymbol{\sigma}_{\mathrm{el}}(\dot{\boldsymbol{u}})) = -\operatorname{div}(\boldsymbol{\sigma}_{\mathrm{el}}(\boldsymbol{v})) & \text{in } (0,T]\times\Omega \\ [(\boldsymbol{\sigma}_{\mathrm{el}}(\boldsymbol{u}) + \boldsymbol{\sigma}_{\mathrm{vi}}(\boldsymbol{v}))\boldsymbol{n}](x,t) = 0 & \text{in } (0,T]\times\Gamma_N \\ \boldsymbol{u}(x,t) = 0, \quad \boldsymbol{v}(x,t) = 0, & \text{in } (0,T]\times\Gamma_D \\ \boldsymbol{u}(x,0) = \boldsymbol{U}_0(x), \quad \boldsymbol{v}(x,0) = \boldsymbol{V}_0(x), & \text{in } \Omega \end{cases}.
$$

Note that the second equation now includes the operator $\operatorname{div}\boldsymbol{\sigma}_{\mathrm{el}}$ with a negative sign. After testing against $\boldsymbol{\phi}^v, \boldsymbol{\phi}^u \in [H^1_D(\Omega)]^d$ and integrating by parts we obtain the following weak formulation:

$$
(10) \quad \begin{aligned} &\int_\Omega (\dot{\boldsymbol{v}}\cdot\boldsymbol{\phi}^v + [\boldsymbol{\sigma}_{\mathrm{el}}(\boldsymbol{u}) + \boldsymbol{\sigma}_{\mathrm{vi}}(\dot{\boldsymbol{u}})] : \nabla^{\mathrm{S}}\boldsymbol{\phi}^v)\, d\boldsymbol{x} = \int_\Omega \boldsymbol{f}\cdot\boldsymbol{\phi}^v\, d\boldsymbol{x}, && \forall \boldsymbol{\phi}^v \in [H^1_D(\Omega)]^d, \\ &\int_\Omega \boldsymbol{\sigma}_{\mathrm{el}}(\dot{\boldsymbol{u}}) : \nabla^{\mathrm{S}}\boldsymbol{\phi}^u\, d\boldsymbol{x} = \int_\Omega \boldsymbol{\sigma}_{\mathrm{el}}(\boldsymbol{v}) : \nabla^{\mathrm{S}}\boldsymbol{\phi}^u\, d\boldsymbol{x}, && \forall \boldsymbol{\phi}^u \in [H^1_D(\Omega)]^d. \end{aligned}
$$

6.3 Galerkin Space Discretisation

The numerical scheme starts by defining a finite dimensional subspace $\mathcal{V}_{N_0}$ of $[H^1_D(\Omega)]^{2d}$, where we look for our candidate approximate solution $(\boldsymbol{v}_N, \boldsymbol{u}_N)$ of problem (10), and a finite dimensional subspace $\mathcal{V}_N \geq \mathcal{V}_{N_0}$ of $[H^1(\Omega)]^{2d}$. We then need to define a set of basis functions such that $\mathcal{V}_{N_0} = \text{span}(\boldsymbol{\Phi}_1, \dots, \boldsymbol{\Phi}_{N_0})$ and $\mathcal{V}_N = \text{span}(\hat{\boldsymbol{\Phi}}_1, \dots, \hat{\boldsymbol{\Phi}}_N)$. We can think of the vector basis functions $\boldsymbol{\Phi}_i$ and $\hat{\boldsymbol{\Phi}}_i$ as subdivided into two blocks, namely $\boldsymbol{\Phi}_i = (\boldsymbol{\Phi}_i^v, \boldsymbol{\Phi}_i^u)^\mathsf{T}$ and $\hat{\boldsymbol{\Phi}}_i = (\hat{\boldsymbol{\Phi}}_i^v, \hat{\boldsymbol{\Phi}}_i^u)^\mathsf{T}$, where the former refers to $\boldsymbol{v}_N$ and the latter to $\boldsymbol{u}_N$. We choose our basis functions by picking *continuous* and piecewise smooth functions: this, in particular, enforces the continuity of $\boldsymbol{u}_N$ and $\boldsymbol{v}_N$. We can then write:

$$\begin{pmatrix} \boldsymbol{v}_N \\ \boldsymbol{u}_N \end{pmatrix} = \sum_{j=1}^{N_0} \begin{pmatrix} \eta_j \\ \theta_j \end{pmatrix} \boldsymbol{\Phi}_j = \begin{pmatrix} \sum_{j=1}^{N_0} \eta_j \boldsymbol{\Phi}_j^v \\ \sum_{j=1}^{N_0} \theta_j \boldsymbol{\Phi}_j^u \end{pmatrix}, \tag{11}$$

$$\begin{pmatrix} \mathcal{L}\dot{\boldsymbol{u}}_D \\ \mathcal{L}\boldsymbol{u}_D \end{pmatrix} = \sum_{j=1}^{N} \begin{pmatrix} \eta_{j,D} \\ \theta_{j,D} \end{pmatrix} \hat{\boldsymbol{\Phi}}_j = \begin{pmatrix} \sum_{j=1}^{N} \eta_{j,D} \hat{\boldsymbol{\Phi}}_j^v \\ \sum_{j=1}^{N} \theta_{j,D} \hat{\boldsymbol{\Phi}}_j^u \end{pmatrix}, \tag{12}$$

where $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ are the unknowns of our problem and $\boldsymbol{\eta}_D$ and $\boldsymbol{\theta}_D$ are given. Note that the spatial gradient of the expansion applies to the basis functions $\boldsymbol{\Phi}$ while the time derivative to the coefficients $(\boldsymbol{\eta}^\mathsf{T}, \boldsymbol{\theta}^\mathsf{T})^\mathsf{T}$. Substituting the expressions defined in (11)-(12) into the weak formulation (10) and taking test functions $\boldsymbol{\Phi}_i$, with $i = 1, 2, \dots, N_0$, we obtain a system of first order Ordinary Differential Equations for the coefficients that can be written as:

$$\begin{aligned} &\mathbf{M}\dot{\boldsymbol{\xi}} + \mathbf{H}\boldsymbol{\xi} = \boldsymbol{F} \\ &\boldsymbol{\xi}(0) = \boldsymbol{\xi}_0, \end{aligned} \tag{13}$$

where $\boldsymbol{\xi} = (\boldsymbol{\eta}^\mathsf{T}, \boldsymbol{\theta}^\mathsf{T})^\mathsf{T}$ and $\boldsymbol{\xi}_0$ is the given initial condition. Here, matrices $\mathbf{M}$ and $\mathbf{H}$ are given by

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}^v & \mathbf{0} \\ \mathbf{0} & \mathbf{M}^u \end{pmatrix} \qquad \mathbf{H} = \begin{pmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{0} \end{pmatrix}$$

and their blocks have dimension $N_0 \times N_0$ with components (i, j) defined by

$$(\mathbf{M}^v)_{i,j} = \int_\Omega \boldsymbol{\Phi}_i^v \cdot \boldsymbol{\Phi}_j^v \, d\boldsymbol{x}, \qquad (\mathbf{M}^u)_{i,j} = \int_\Omega \boldsymbol{\sigma}_{\text{el}}(\boldsymbol{\Phi}_j^u) : \nabla \boldsymbol{\Phi}_i^u \, d\boldsymbol{x}$$

and

$$(\mathbf{H}_{11})_{i,j} = \int_\Omega \boldsymbol{\sigma}_{\text{vi}}(\boldsymbol{\Phi}_j^v) : \nabla \boldsymbol{\Phi}_i^v \, d\boldsymbol{x}, \quad (\mathbf{H}_{12})_{i,j} = \int_\Omega \boldsymbol{\sigma}_{\text{el}}(\boldsymbol{\Phi}_j^u) : \nabla \boldsymbol{\Phi}_i^v \, d\boldsymbol{x}, \quad (\mathbf{H}_{21})_{i,j} = -\int_\Omega \boldsymbol{\sigma}_{\text{el}}(\boldsymbol{\Phi}_j^v) : \nabla \boldsymbol{\Phi}_i^u \, d\boldsymbol{x}.$$

Then, the vector $\boldsymbol{F}$ in (13) is defined by setting

$$\boldsymbol{F}_i := \begin{cases} -\int_\Omega (\boldsymbol{\sigma}_{\text{el}}(\mathcal{L}\boldsymbol{u}_D) + \boldsymbol{\sigma}_{\text{vi}}(\mathcal{L}\dot{\boldsymbol{u}}_D)) : \nabla \boldsymbol{\Phi}_i^v - \mathcal{L}\ddot{\boldsymbol{u}}_D \cdot \boldsymbol{\Phi}_i^v \, d\boldsymbol{x} & i = 1, 2, \dots, N_0, \\ 0 & i = N_0 + 1, \dots, 2N_0. \end{cases}$$

6.4 A Numerical Example: (a variant of) Lamb's Problem

In this final section, we show how the numerical approach previously discussed can be applied to compute the solution of a variant of the so-called Lamb's Problem [18]. The original problem consists of a uniform (with respect to y) normal line load that is suddenly applied to the surface of an homogeneous, isotropic and elastic half space, $z = 0$, at time $t = 0$. The uniformity of the loading in the y direction makes this a plane strain problem (i.e., $u_y = 0$ and $\partial/\partial y = 0$). This problem has been proposed and studied for the first time at the beginning of the twentieth century by Lamb, who was investigating the propagation of surface waves. The problem we solve numerically here is slightly different in the sense that we consider a computational bounded domain $\Omega = [0, 1]^2$ instead of the half-plane (x, z) (see Figure 5, left panel) and we impose zero Neumann boundary conditions on $\partial\Omega$. Moreover, we consider a forcing function having the form $F(x, z, t) = \delta(x - 1/2)\delta(z)f(t)$, where the function $f(t)$ defining the time-dependence behaviour of the applied force is given by the so-called Ricker wavelet

$$f(t) = A(1 - 2(\pi f_0(t - t_0))^2) \exp\left(-(\pi f_0(t - t_0))^2\right)$$

with parameters $A = 1$, $f_0 = 2$ and $t_0 = 1$ (see Figure 5, right panel). Figure 6 shows a comparison between the elastic and viscoelastic responses (horizontal and vertical components of both displacement and velocity) computed at the point having coordinates $(0.6, 0)$, i.e., the blue point in Figure 5. In particular, in the viscoelastic case, different values of the quality factors Q_p and Q_s have been considered. The plots clearly show that a lower value of the quality factor is associated to a higher attenuation of the displacement and valocity amplitude. Finally, Figure 7 shows the displacement magnitude in the elastic and viscoelastic case for six time instants. Owing to the zero Neumann boundary conditions imposed on $\partial\Omega$, we observe that, both in the elastic and viscoelastic cases, the solution is characterized by reflections at the boundaries. Furthermore, in the viscoelastic case we can see again that the magnitude of the displacement is smaller than the one obtained in the purely elastic case and that it descreases as the waves propagate in space.

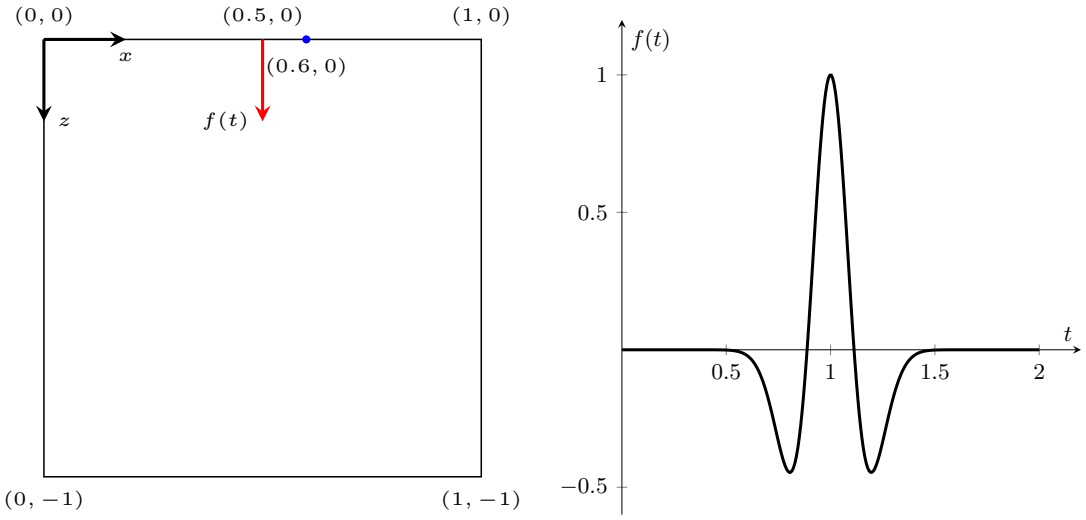


Figure 5: Left panel: computational domain used to compute the numerical solution of a variant of Lamb's Problem. The red arrow denotes the vertical point load, which is applied in the midpoint of the top boundary. Right panel: Ricker wavelet defining the time dependence of the concentrated load.

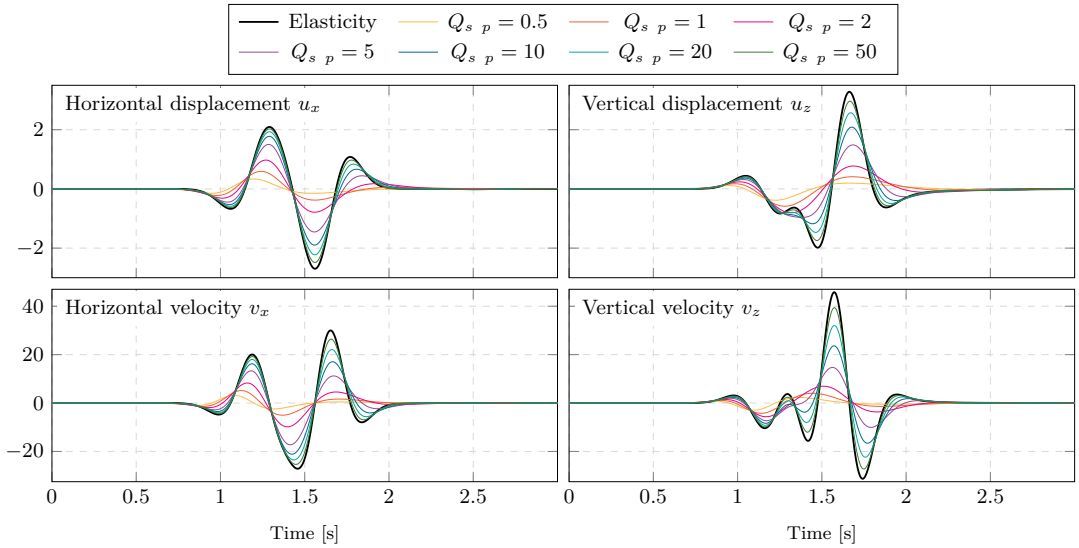


Figure 6: Comparison between the elastic and viscoelstic response for the variant of Lamb's Problem considered in Figure 5. The viscoelastic solutions have been obtained by considering different values of the quality factors $Q_s = Q_p$. The responses are computed at the point of coordinate $(0.6, 0)$, i.e., the blue point in Figure 5 (left panel). All the panels show that a lower value of the quality factor results in an higher attenuation of the amplitude wave.

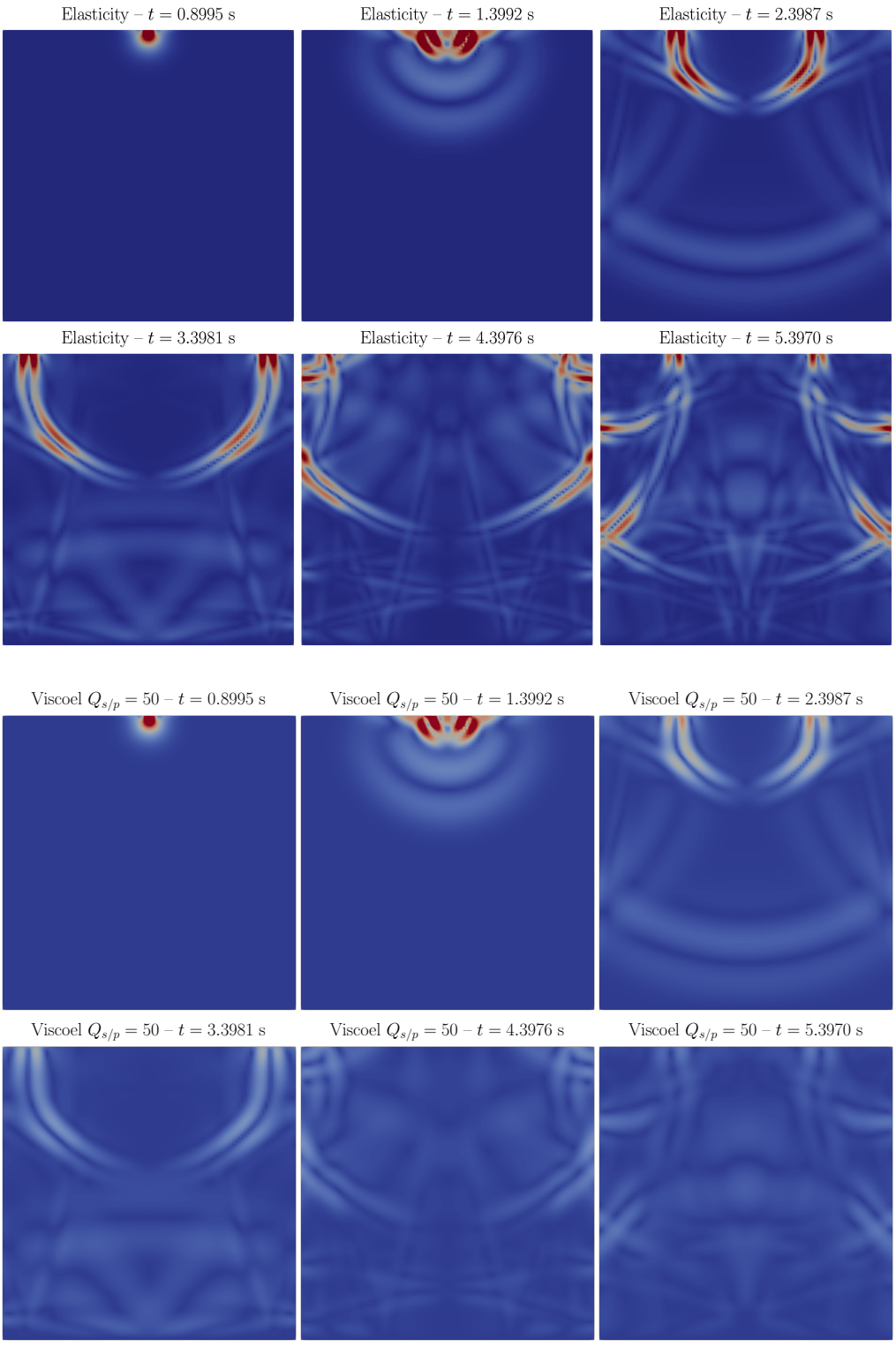


Figure 7: Displacement magnitude for six different time instants. Comparison between the solution in the elastic case (top) and the solution in the Kelvin-Voigt case with $Q_p = Q_s = 50$ (bottom).

References

[1] J.C. Maxwell, *On the dynamical theory of gases.* In *The kinetic theory of gases: an anthology of classic papers with historical commentary*, pages 197–261. World Scientific, 1867.

[2] W. Voigt, *Über innere Reibung fester Körper, insbesondere der Metalle.* Annalen der Physik, 283(12): 671–693, 1892.

[3] L. Boltzmann, *Zur Theorie der elastischen Nachwirkung.* Annalen der Physik, 241(11): 430–432, 1878.

[4] V. Volterra, "Equazioni integro-differenziali della elasticità nel caso della isotropia.". Tipografia della Regia Accademia del Lincei, 1910.

[5] V. Volterra, "Energia nei fenomeni elastici ereditari". Pontificia Academia Scientiarum, 1940.

[6] H.T. Banks, S. Hu, and Z.R. Kenz, *A brief review of elasticity and viscoelasticity for solids.* Advances in Applied Mathematics and Mechanics, 3(1): 1–51, 2011.

[7] J.M. Carcione, "Wave fields in real media: Wave propagation in anisotropic, anelastic, porous and electromagnetic media". Elsevier, 2007.

[8] R. Christensen, "Theory of viscoelasticity: an introduction". Elsevier, 2012.

[9] M.E. Gurtin and E. Sternberg, *On the linear theory of viscoelasticity.* Archive for Rational Mechanics and Analysis, 11(1): 291–356, 1962.

[10] J.M. Golden and G.A. Graham, "Boundary value problems in linear viscoelasticity". Springer Science & Business Media, 2013.

[11] W.N. Findley and F.A. Davis, "Creep and relaxation of nonlinear viscoelastic materials". Courier corporation, 2013.

[12] A.E. Green and R.S. Rivlin, *The mechanics of non-linear materials with memory.* Archive for Rational Mechanics and Analysis, 1(1): 1–21, 1957.

[13] C.S. Drapaca, S. Sivaloganathan, and G. Tenti, *Nonlinear constitutive laws in viscoelasticity.* Mathematics and mechanics of solids, 12(5): 475–501, 2007.

[14] A.C. Pipkin and T.G. Rogers, *A non-linear integral representation for viscoelastic behaviour.* Journal of the Mechanics and Physics of Solids, 16(1): 59–72, 1968.

[15] R. Schapery., *Nonlinear viscoelastic solids.* International journal of solids and structures, 37(1-2): 359–366, 2000.

[16] A. Wineman, *Nonlinear viscoelastic solids – a review.* Mathematics and mechanics of solids, 14(3): 300–366, 2009.

[17] J.M. Carcione, F. Poletto, and D. Gei, *3-d wave simulation in anelastic media using the Kelvin-Voigt constitutive equation.* Journal of Computational Physics, 196(1): 282–297, 2004.

[18] H. Lamb, *On the propagation of tremors over the surface of an elastic solid.* Proceedings of the Royal Society of London, 72: 128–130, 1903.

Strichartz estimates for the Dirac equation in different settings

Elena Danesi (*)

Abstract. The Dirac equation is a first order partial differential equation that comes for quantum mechanics and general relativity. From the mathematical side, it can be listed within the class of dispersive equations, together with, e.g., the Schrödinger, wave and Klein-Gordon equations. In the years, because of the study of nonlinear systems, a lot of effort has been devoted to developing tools to quantify the dispersion of a system. Among these tools we find a priori estimates on the solutions, such as decay or Strichartz estimates. In the first part of this essay we will describe the class of dispersive equations. Special focus will be posed on the Schrödinger and wave equations, in order to present these kind of estimates as well as some classical tools to prove them. In the second part, we will first introduce the Dirac equation on $\mathbb{R} \times \mathbb{R}^3$ and describe its connection with the above mentioned equations. We will then describe the equation in curved spaces. To conclude, a survey of some recent results concerning the validity of Strichartz estimates for the "curved" Dirac equation in specific settings, in particular compact or asymptotically flat manifolds, is presented.

1 What is a dispersive equation?

Let us consider the following PDEs:

- *transport eq.*
$$\partial_t u + v \cdot \nabla_x u = 0, \quad v \in \mathbb{R}^n,$$
- *phase rotation eq.*
$$i\partial_t u + \omega_0 u = 0, \quad \omega_0 \in \mathbb{R},$$
- *Schrödinger eq.*
$$ih\partial_t u - \frac{h^2}{2m}\Delta u = 0, \quad m > 0,$$
- *wave eq.*
$$\partial_{tt}^2 u - c^2 \Delta u = 0,$$

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: edanesi@math.unipd.it . Seminar held on 5 June 2024.

- *Klein-Gordon eq.*
$$\partial_{tt}^2 u - c^2 \Delta u + \frac{m^2 c^4}{h^2} u = 0, \quad m > 0.$$

Here $(t, x) \in \mathbb{R} \times \mathbb{R}^n$, $n \geq 1$, h is the Planck's constant and c is the speed of light. To each of these equations we can associate a function $\omega \colon \mathbb{R}^n \to \mathbb{R}$ which is defined as follows: $\omega(k)$ is such that the plane wave
$$\phi(t, x) \coloneqq e^{ik\cdot x + i\omega(k)t}$$
is a solution of the corresponding equation. If we compute the value of ω in the various cases we get

- *transport eq.*
$$\partial_t u + v \cdot \nabla_x u = 0 \quad \to \quad \omega(k) = -k \cdot v,$$

- *phase rotation eq.*
$$i\partial_t u + \omega_0 u = 0 \quad \to \quad \omega(k) = \omega_0,$$

- *Schrödinger eq.*
$$ih\partial_t u - \frac{h^2}{2m} \Delta u = 0 \quad \to \quad \omega(k) = \frac{h}{2m} |k|^2,$$

- *wave eq.*
$$\partial_{tt}^2 u - c^2 \Delta u = 0 \quad \to \quad \omega(k) = \pm c|k|,$$

- *Klein-Gordon eq.*
$$\partial_{tt}^2 u - c^2 \Delta u + \frac{m^2 c^4}{h^2} u = 0 \quad \to \quad \omega(k) = \pm \sqrt{c^2 |k|^2 + \frac{m^2 c^4}{h^2}}.$$

Moreover, we define the *group velocity* as $v_g \coloneqq \nabla_k \omega(k)$. We observe that we can identify two different behaviors; v_g can depend on k or not. In the first case, we call the corresponding equation *dispersive*. Among this group we can further distinguish the case where the group velocity is uniformly bounded in k and when it is not. In the former case the equation is called dispersive with infinite speed of propagation. This is the case for example of the Schrödinger equation. In the latter the equation is said to be dispersive with finite speed of propagation. Examples are given by the wave and Klein-Gordon equations. To explain why it is helpful to make this distinction when studying the behavior of the solutions of the above mentioned equations, let us do some heuristics. Without loss of generality we restrict to $n = 1$. We consider a wave packet, i.e. a superposition of waves, given by
$$\Psi(t, x) = \int_{\mathbb{R}} \phi(k) e^{ikx + i\omega(k)t} dk,$$
where ϕ is a smooth function with support focused around $k_0 \in \mathbb{R}$. We can then expand $\omega(k) \simeq \omega(k_0) + \omega'(k_0)(k - k_0)$ and rewrite Ψ as
$$\Psi(t, x) = e^{ik_0 \left(x + \frac{\omega(k_0)}{k_0}\right) t} \int_{\mathbb{R}} \phi(k) e^{i(k-k_0)(x+\omega'(k_0)t)} dk.$$

The plane wave in front of the integral is moving with speed $\frac{\omega(k_0)}{k_0}$ and tells us how ripples are moving. The integral instead modulates the envelope of the wave packet and it is moving with speed $\omega'(k_0)$. Therefore, if $\omega'(k_0)$ depends on k_0, different frequencies are propagating at different velocities, dispersing the solution over time. If, instead, the group velocity is constant, the envelope will travel maintaining its shape. The interested reader can look at `https://en.wikipedia.org/wiki/Group_velocity` at the animated plots to better visualize these phenomena.

2 How to capture dispersion

In the years, thanks to the research on nonlinear models, it has been understood that dispersion plays a fundamental role in the dynamics of a system. Consequently, a great effort has been devoted to studying the tools which permit to quantify dispersive phenomena in terms of a priori estimates for the free flows. In this section we describe two types of a priori estimates, the *time-decay* and *Strichartz estimates*, which as we will see are not unrelated. We will focus on the Schrödinger and wave equations in order to present some classical tools that are widely used in this field. We will not enter deep in the details, no proofs of the results are presented. We refer for them to [2] (chapter 8).

2.1 Time-decay estimates

i) *Schrödinger equation:* Let us consider the Cauchy problem

$$\begin{cases} i\partial_t u - \Delta u = 0, & (t,x) \in \mathbb{R} \times \mathbb{R}^n, \\ u(0,x) = u_0(x), \end{cases} \tag{1}$$

where $u\colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{C}$. The goal is now to find a good representation of the solution. In order to avoid technical issues, we restrict our attention to the Schwartz space $\mathcal{S}(\mathbb{R}^n)$, the vector space of smooth rapidly decreasing functions. Moreover, we adopt the following notation for the *Fourier transform* with respect to the space variable:

$$\mathcal{F}u(\xi) = \hat{u}(\xi) \coloneqq \int u(x) e^{-ix\cdot\xi} dx, \quad \forall \xi \in \mathbb{R}^n.$$

We recall the following useful property of the Fourier transform:

$$\mathcal{F}(\partial^\alpha u)(\xi) = i^{|\alpha|} \xi^\alpha \mathcal{F}u(\xi), \quad \xi \in \mathbb{R}^n,$$

for any multiindex $\alpha = (\alpha_1, \dots, \alpha_n)$ with length $|\alpha|$. Therefore, by taking the Fourier transform of (1) we have

$$\begin{cases} i\partial_t \hat{u}(t,\xi) - i^2|\xi|^2 \hat{u}(t,\xi) = 0, \\ \hat{u}(0,\xi) = \hat{u}_0(\xi), \end{cases}$$

which has solution

$$\hat{u}(t,\xi) = e^{it|\xi|^2} \hat{u}_0(\xi).$$

Taking the inverse Fourier transform we finally get

$$u(t,x) = \mathcal{F}^{-1}\Big(e^{it|\xi|^2}\hat{u}_0(\xi)\Big)(x) =: e^{-it\Delta}u_0(x).$$

Moreover, by explicit computations it is not difficult to get that

$$u(t,x) = \frac{1}{(4\pi it)^{\frac{n}{2}}}\int_{\mathbb{R}^n} e^{i\frac{|x-y|^2}{4t}}u_0(y)dy.$$

Therefore, it is straightforward to obtain the following time-decay estimate for the solution of the Schrödinger equation

$$\|e^{-it\Delta}u_0\|_{L^\infty_x} \leq C|t|^{-\frac{n}{2}}\|u_0\|_{L^1_x}, \tag{2}$$

for some constant $C > 0$ which does not depend on t nor on the initial datum.

ii) *Wave equation:* Let us now focus on the Cauchy problem associated with the wave equation

$$\begin{cases} \partial^2_{tt}u - \Delta u = 0, \quad (t,x) \in \mathbb{R}\times\mathbb{R}^n, \\ u(0,x) = u_0(x), \\ \partial_t u(0,x) = u_1(x), \end{cases} \tag{3}$$

where as before $u\colon \mathbb{R}\times\mathbb{R}^n \to \mathbb{C}$. We would like to investigate the validity of a polynomial time-decay estimate as in the case of Schrödinger. To do so, we play the same game as before. We pass to the Fourier transform in (3). We obtain

$$\begin{aligned} u(t,x) &= \mathcal{F}^{-1}\big(\cos(t|\xi|)\hat{u}_0\big)(x) + \mathcal{F}^{-1}\left(\frac{\sin(t|\xi|)}{|\xi|}\hat{u}_1(\xi)\right)(x) \\ &=: \cos(t\sqrt{-\Delta})u_0(x) + \frac{\sin(t\sqrt{-\Delta})}{\sqrt{-\Delta}}u_1(x). \end{aligned} \tag{4}$$

This representation suggests to focus on the study of

$$g := e^{it\sqrt{-\Delta}}f = \mathcal{F}^{-1}\big(e^{it|\xi|}\hat{f}\big)(x)$$

since then

$$\begin{aligned} \cos(t\sqrt{-\Delta}) &= \frac{e^{it\sqrt{-\Delta}} + e^{-it\sqrt{-\Delta}}}{2} = Re(e^{it\sqrt{-\Delta}}), \\ \sin(t\sqrt{-\Delta}) &= \frac{e^{it\sqrt{-\Delta}} - e^{-it\sqrt{-\Delta}}}{2i} = Im(e^{it\sqrt{-\Delta}}). \end{aligned}$$

The analysis of the decay of this propagator is not as simple as the one of Schrödinger; in fact, it involves the study of oscillatory integrals via the method of stationary/non-stationary phase. An extensive treatment of this topic can be found in [9]. Nevertheless, it has been proved that the following time-decay holds

$$\|e^{it\sqrt{-\Delta}}u_0\|_{L^\infty_x} \leq C|t|^{-\frac{n-1}{2}}\|f\|_{L^1_x} \tag{5}$$

where $C > 0$ if f is *frequency localized*, that is, if

$$\operatorname{supp}\hat{f} \subseteq \{r \leq |\xi| \leq R\} \quad \text{for some} \quad 0 < r < R < +\infty. \tag{6}$$

2.2 Strichartz estimates

We observe that, by Plancherel's theorem, we have "for free" an identity for both Schrödinger and wave propagator. Indeed

$$\|e^{-it\Delta}u_0\|_{L^2_x} = \|e^{it|\xi|}\hat{u}_0\|_{L^2_\xi} = \|\hat{u}_0\|_{L^2_\xi} = \|u_0\|_{L^2_x}$$

and similarly

$$\|e^{it\sqrt{-\Delta}}f\|_{L^2_x} = \|f\|_{L^2_x}.$$

These estimates can be combined with the decay estimates found in the previous section to obtain a number of inequalities involving space-time Lebesgue norms, called Strichartz estimates. The classical method is based on real intepolation and duality arguments. It is summarized in this result of Keel and Tao (see [10], Theorem 1.2) that we present here in a simplified version.

Proposition 1 *Let us assume that for each $t \in \mathbb{R}$ we have an operator $U(t)\colon L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$ such that*

i) *for some $c_1 > 0$*

$$\|U(t)f\|_{L^2} \le c_1\|f\|_{L^2},$$

ii) *for some $\sigma > 0$ and $c_2 > 0$*

$$\|U(s)U(t)^*f\|_{L^\infty} \le c_2|t-s|^{-\sigma}\|f\|_{L^1}.$$

Then the estimate

$$\|U(t)f\|_{L^p_t L^q_x} \le c\|f\|_{L^2_x}$$

holds for any $(p,q) \in [2,+\infty]^2$ $\sigma-$admissible, i. e. such that

$$\frac{1}{p} + \frac{\sigma}{q} = \frac{\sigma}{2}, \qquad (p,q,\sigma) \neq (2,\infty,1).$$

All the constants that appear do not depend on t and on f.

The mixed Lebesgue norms are defined as

$$\|f\|_{L^p_t(\mathbb{R};L^q_x(\mathbb{R}^n))} = \left(\int_{\mathbb{R}}\left(\int_{\mathbb{R}^n}|f|^q dx\right)^{\frac{p}{q}} dt\right)^{\frac{1}{q}} \quad \forall p,q \in [1,+\infty),$$

with natural modifications if $p, q = +\infty$. Therefore, from (2) and Proposition 1, we have the following family of Strichartz estimates for the Schrödinger propagator

$$\|e^{-it\Delta}u_0\|_{L^p_t L^q_x} \le c\|u_0\|_{L^2}$$

for any $p, q \ge 2$ such that

$$\frac{2}{p} + \frac{n}{q} = \frac{n}{2}, \quad q < \frac{2n}{n-2}. \tag{7}$$

A couple of indexes satisfying (7) is said to be *Schrödinger admissible*. Let us stress that, if $n > 2$, the endpoint $q = \frac{2n}{n-2}$ is admissible. Here and in the following we use c to denote a positive constant which depends only on p, q, n.
On the other side, for the propagator associated to the wave equation we have that if f is frequency localized (as in (6)) then

$$\|e^{it\sqrt{-\Delta}} f\|_{L_t^p L_x^q} \leq c\|f\|_{L^2}$$

for any $p, q \geq 2$ such that

$$\frac{2}{p} + \frac{n-1}{q} = \frac{n-1}{2}, \quad q < \frac{2(n-1)}{n-3}. \tag{8}$$

A couple of indexes satisfying (8) is said to be *wave admissible*. In this case, the endpoint $q = \frac{2(n-1)}{n-3}$ is admissible if $n > 3$. It seems however too restrictive to consider only frequency localized initial data. To recover the general case, one relies on the *Paley-Littlewood decomposition*. We refer to [1] for more details and applications. Roughly speaking, the main idea of this procedure consists in sampling the frequencies by means of a decomposition in the frequency space in annuli of size 2^j, $j \in \mathbb{Z}$. In this way, one obtains a decomposition of the function into a sum of a countable number of functions whose Fourier transform is supported in an annulus. In this way it is possibile to prove the following family of Strichartz estimates for a function f without any assumption of localization:

$$\|e^{it\sqrt{-\Delta}} f\|_{L_t^p L_x^q} \leq c\|f\|_{\dot{H}^s}$$

where p, q are as in (8), $s = n\left(\frac{1}{2} - \frac{1}{q}\right) - \frac{1}{p}$ and the norm on the RHS is the homogeneous Sobolev norm which can be defined for any $\gamma \in \mathbb{R}$ as

$$\|f\|_{\dot{H}^\gamma} = \||\xi|^\gamma \hat{f}(\xi)\|_{L^2_\xi} = \|(\sqrt{-\Delta})^\gamma f\|_{L^2}.$$

Finally, we can combine these estimates with the decomposition (4) to get the Strichartz estimates for the wave equation; let u be a solution of (3), then

$$\|u\|_{L_t^p L_x^q} \leq c(\|u_0\|_{\dot{H}^s} + \|u_1\|_{\dot{H}^{s-1}})$$

where p, q satisfy conditions (8) and $s = n\left(\frac{1}{2} - \frac{1}{q}\right) - \frac{1}{p}$.
To conclude, let us discuss what kind of information it is possible to interfere from the Strichartz estimates; locally in time, they describe a type of smoothing effect, but reflected in a gain of integrability rather than regularity (if the datum is in L^2_x, the solution $u(t)$ is in L^q_x with $q > 2$ for most of the time), and only if one averages in time. For fixed time, no gain in integrability is possible (see Exercise 2.35 in [13]). Globally in time, they describe a decay effect: the L^q_x norm of a solution $u(t)$ must decay to zero as $t \to \infty$, at least in some L^p_t-averaged sense. Both effects of the Strichartz estimates reflect the dispersive nature of the equation (i.e. that different frequencies propagate in different directions); it is easy to verify that no such estimates are available for the dispersionless equations (e.g. transport equation), except for the trivial pair of exponents $(p, q) = (\infty, 2)$.
Even if we will not discuss the applications in this account, we remark that these estimates are widely used in the study of local and global well-posedness and scattering results for nonlinear systems. An extensive treatment of this topic can be found in [11].

3 The Dirac equation

The first goal of this section is to present the classical derivation of the Dirac equation on $\mathbb{R}^3$. For the sake of brevity we will omit some details for which we refer the reader to Dirac's paper [8] and to [14]. We then describe the Strichartz estimates that hold for the Dirac equation in a flat setting. In the second subsection we switch to curved backgrounds, describing some very recent results regarding the dispersion, in terms of the validity of Strichartz estimates, of these systems.

3.1 The Dirac equation on $\mathbb{R}^n$

The Dirac equation was introduced by Paul Dirac in 1928 to describe the motion of fermions, such as electrons, which move freely in $\mathbb{R}^3$. In order to explain how it was derived, we start from the relativistic energy-momentum relation:

$$E^2 = p^2c^2 + m^2c^4 \tag{9}$$

where $p = (p_1, p_2, p_3)$ and m are respectively the momentum and the mass of the particle and c is the speed of light. Then, formally, the transition from classical to quantum mechanics can be accomplished by substituting appropriate operators for the classical quantities. In particular,

$$E = p_0 \to ih\partial_t, \qquad p_j = -ih\partial_{x_j}, \quad j = 1, 2, 3, \tag{10}$$

where h is the Planck's constant. Therefore, one obtains the Klein-Gordon equation

$$\partial_{tt}^2\psi - c^2\Delta\psi + \frac{m^2c^4}{h^2}\psi = 0,$$

with $\psi(t, x)$ a scalar function. The resulting equation is Lorentz covariant, but it does not allow to describe the internal structure of the electrons, namely the spin. Moreover, if one tries to construct a conserved current as for the Schrödinger equation, one obtains

$$\psi^*\partial_t\psi - \psi\partial_t\psi^* = 0,$$

but the quantity defined on the LHS is not positive definite, so it is impossible to interpret it as a probability density. The goal is then to find a first order in time equation which admits a straightforward interpretation as in the Schrödinger equation. The first idea would be to take the square-root of (9)

$$E = \sqrt{p^2c^2 + m^2c^4}$$

and quantized as before. In this way one gets the following equation

$$ih\partial_t\psi = \sqrt{-c^2h^2\Delta + m^2c^4}\psi.$$

This forces to face the problem of interpreting the square-root operator on the RHS. In order to solve this problem, Dirac's idea was to look for a linearized equation of the form

$$(p_0 + \alpha_1 p_1 + \alpha_2 p_2 + \alpha_3 p_3 + m\alpha_4)\psi = 0,$$

where $\{\alpha_i\}_{i=1}^4$ are some dynamical variables or operators that are independent of t, x_1, x_2, x_3. That is to say, by "squaring" the equation we should obtain the energy-momentum relation. Recalling (10) we formally impose

$$(i\partial_t - i\sum_{j=1}^{3}\alpha_j\partial_{x_j} + m\alpha_4)(-i\partial_t - i\sum_{j=1}^{3}\alpha_j\partial_{x_j} + m\alpha_4) = \left(\partial_{tt}^2 - \Delta + m^2\right)\otimes \mathbb{1}. \tag{11}$$

Here and in the following, to lighten the notation, we set $c = h = 1$. We compute the LHS and get

$$\partial_{tt}^2 - \sum_{j=1}^{3}\alpha_j^2\partial_{x_jx_j}^2 + m^2\alpha_4^2 - \sum_{i,j=1,i<j}^{3}(\alpha_i\alpha_j + \alpha_j\alpha_i)\partial_{x_i}\partial_{x_j} - im\sum_{j=1}^{3}(\alpha_j\alpha_4 + \alpha_4\alpha_j)\partial_{x_j}.$$

Therefore, in order to get the Klein-Gordon equation (or, better, a system of decoupled equations) we have to look for $\{\alpha_j\}_{j=1}^4$ such that

$$\{\alpha_i, \alpha_j\} \coloneqq \alpha_i\alpha_j + \alpha_j\alpha_i = 2\delta_{ij}\mathbb{1}, \qquad \forall i, j = 1, \dots 4,$$

where δ_{ij} is the Kronecker delta. From this relation it is clear that α_j's cannot be scalars, but matrices. The smallest dimension in which these four matrices can be realized is $N = 4$. In a particular and widely used representation the α's matrices, called Dirac matrices, are given by

$$\alpha_j = \begin{pmatrix} 0_{2\times2} & \sigma_j \\ \sigma_j & 0_{2\times2} \end{pmatrix}, \quad j = 1, 2, 3, \qquad \alpha_4 = \begin{pmatrix} \mathbb{1}_{2\times2} & 0_{2\times2} \\ 0_{2\times2} & \mathbb{1}_{2\times2} \end{pmatrix},$$

Here $\sigma_j \in \mathcal{M}_{2\times2}(\mathbb{C})$ are the Pauli matrices:

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

At last, the Dirac equation reads as

$$i\partial_t\psi + \mathcal{D}\psi + m\alpha_4\psi = 0,$$

where $\psi \coloneqq \mathbb{R}_t \times \mathbb{R}_x^3 \to \mathbb{C}^4$, $\mathcal{D}$ is the Dirac operator defined as $\mathcal{D} \coloneqq -i\sum_j \alpha_j\partial_{x_j}$ and $m \geq 0$. The vector-valued wavefunction ψ on which the Dirac operator acts is called *spinor*. It is possibile to show that the obtained equation is Lorentz covariant, we refer to [3] (section 2.1) for the details.

This construction can be generalized for $x \in \mathbb{R}^n$, $n \in \mathbb{N}$. In this case, the α matrices are taken in $M_{N\times N}(\mathbb{C})$ with $N = 2^{\lceil \frac{n}{2} \rceil}$. For a reader with interests in algebra, we remark that the problem of finding these matrices is connected to the one of finding a basis for the Clifford algebra $Cl_{1,3}(\mathbb{R})$.

As we saw before, the dynamics of the Dirac equation is strictly connected to the one of the wave or Klein-Gordon equation, respectively if $m = 0$ or $m > 0$. Indeed, by the identity (11) we obtain that $u(t, x) \coloneqq e^{it(\mathcal{D}+m\alpha_4)}u_0(x)$ satisfies the Cauchy problem

$$\begin{cases} (\partial_{tt}^2 - \Delta + m^2)\mathbb{1}_N u = 0, \\ u(0, x) = u_0(x), \\ \partial_t u(0, x) = i(\mathcal{D} + m\alpha_4)u_0(x). \end{cases}$$

Hence, each component of u satisfies the same Strichartz estimates as for the $n-$dimensional wave or Klein-Gordon equation. Let us briefly recall them, for the sake of completeness (we refer to [7] (appendix A) for the ones of the Klein-Gordon equation).
Let u be the solution of the Cauchy problem associated with the massless Dirac equation

$$\begin{cases} i\partial_t u + \mathcal{D}u = 0, \\ u(0,x) = u_0(x). \end{cases}$$

Then, for any (p,q) wave admissible

$$\|u\|_{L_t^p L_x^q} \leq c\|u_0\|_{\dot{H}^s}$$

where $s = \frac{1}{2} + \frac{1}{p} - \frac{1}{q}$. Instead, if v is a solution of the Cauchy problem associated with the massive Dirac equation

$$\begin{cases} i\partial_t v + \mathcal{D}v + m\alpha_4 v = 0, \\ v(0,x) = v_0(x). \end{cases}$$

The following estimates hold for any (p,q) Schrödinger admissible:

$$\|v\|_{L_t^p L_x^q} \leq c\|v_0\|_{H^s}$$

with $s = \frac{1}{2} + \frac{1}{p} - \frac{1}{q}$. We observe that in the latter case the norm of the RHS is the non-homogeneous Sobolev norm. It can be defined for any $\gamma \in \mathbb{R}$ as

$$\|f\|_{H^\gamma} = \|\langle\xi\rangle^\gamma \hat{f}(\xi)\|_{L^2_\xi} = \|\langle\sqrt{-\Delta}\rangle^\gamma f\|_{L^2},$$

where $\langle\cdot\rangle$ denotes the Japanese bracket, $\langle x\rangle \coloneqq \sqrt{1+x^2}$.

Remark 1 As we have seen, there exists a strong link between the massless/massive Dirac equation and the wave/Klein-Gordon equation. We should however remark that the Dirac equation is also connected with the Schrödinger equation. Indeed, in the *non-relativistic limit* $c \to +\infty$ it is possible to find solutions of the Dirac equation that resemble suitably rescaled and modulated solutions of the Schrödinger equation and viceversa. We suggest the interested reader to look at [13] (chapter 2, Ex. 2.8) for references.

3.2 The Dirac equation on curved backgrounds

We now describe how the Dirac equation can be adapted to non-flat backgrounds. We stress the fact that, due to the rich algebraic structure of the Dirac operator, its generalization to curved spaces, even if classical, is significantly more delicate than the one of the Laplacian. The aim here is not to present this construction in the full generality. We refer to [12] and [5] for a gentle introduction. We restrict to the case where time and space are decoupled. The final goal will be to present some very recent results without proofs, contained in [6, 4], concerning respectively global-in-time and local-in-time Strichartz estimates for the Dirac equation on asymptotically flat manifolds and on compact manifolds without boundary.

As discussed in the previous subsection the construction of the Dirac equation on $\mathbb{R}^n$ relies on a family of matrices α^j such that[1]

$$\{\alpha^i, \alpha^j\} = 2\delta^{ij}\mathbb{1}. \tag{12}$$

Then, the Dirac operator was defined as $\mathcal{D}_{\mathbb{R}^n} = -i\alpha^j\partial_{x_j}$, using Einstein's notation for the sum over same indexes. We observe that δ^{ij} is (the inverse of) the metric that gives the Euclidean scalar product on $\mathbb{R}^n$. Let us now replace $(\mathbb{R}^n, \delta_{ij})$ with a complete manifold $\mathcal{M}$ with a given Riemannian metric $g_{\mu\nu}$ and inverse $g^{\mu\nu}$. The idea is then to look for some matrices γ^μ such that

$$\{\gamma^\mu, \gamma^\nu\} = 2g^{\mu\nu}(x)\mathbb{1} \tag{13}$$

and define the Dirac operator as $\mathcal{D}_{\mathcal{M}} = -i\gamma^\mu D_\mu$ with D_μ is the covariant derivative acting on spinor fields. In order to describe these gamma matrices, which are now depending locally on the manifold $\mathcal{M}$, the so called *n-bein* formalism is commonly used in the literature. Roughly speaking, one chooses a frame that locally sends the tangent space $T_{x_0}\mathcal{M}$ to the flat one. More precisely, we take matrices $e_a^\mu(x)$ such that

$$e_a^\mu(x)g_{\mu\nu}(x)e_b^\nu(x) = \delta_{ab} \quad \text{or equivalently} \quad e_a^\mu(x)\delta^{ab}e_b^\nu(x) = g^{\mu\nu}(x),$$

with $\mu, \nu, a, b \in \{1, \dots, n\}$. Then, it is not difficult to prove that the matrices defined as

$$\gamma^\mu := e_a^\mu(x)\alpha^a, \quad \mu = 1, \dots, n$$

satisfy the anticommuting relation (13). Moreover, the covariant derivative for a Dirac spinor is given by

$$D_\mu = \partial_\mu + B_\mu, \quad \mu = 1, \dots, n,$$

where

$$B_\mu = \frac{1}{8}\omega_\mu^{ab}[\bar{\gamma}_a, \bar{\gamma}_b], \quad \bar{\gamma}_0 = \alpha_0, \bar{\gamma}_j = \alpha_0\alpha_j, \quad j = 1, \dots, n$$

with ω_μ^{ab} a pure geometric factor, called *spin connection* that can be defined in terms of the n-bein $e_a^\mu(x)$ and the metric $g_{\mu\nu}$. Then, the Dirac equation on $\mathbb{R}_t \times \mathcal{M}_x$ reads as

$$i\partial_t\psi + \mathcal{D}_{\mathcal{M}}\psi + m\alpha^0\psi = 0.$$

Before describing two recent results concerning the validity of Strichartz estimates for the Dirac equation in two different curved settings, let us observe what we get when we "square" this equation; i.e., we compute

$$\big(i\partial_t + (\mathcal{D}_{\mathcal{M}} + m\alpha^0)\big)\big(-i\partial_t + (\mathcal{D}_{\mathcal{M}} + m\alpha^0)\big) = \partial_{tt}^2 - \Delta^S + m^2 + \frac{1}{4}\mathcal{R}_g.$$

Compared to the flat case, in the RHS we get an extra factor, namely the scalar curvature associated to the spatial metric g. Moreover, it is important to stress the fact that Δ^S is

[1] We raise and lower latin indexes multiplying by δ; $x^j = \delta^{ij}x_i$ and $x_j = \delta_{ij}x^j$.

not for the Laplace-Beltrami operator, but the *spinorial laplacian*. It can be expanded in terms of the Laplace-Beltrami operator, denoted as Δ_g:

$$\Delta^S = \Delta_g - \Omega_1 - \Omega_2,$$

where Ω_j $j = 1, 2$ are terms of order, respectively, one and zero[(2)]:

$$\Omega_1 = 2B^\mu \partial_\mu, \quad \Omega_2 = -\partial^\mu B_\mu + B^\mu B_\mu - \Gamma^{\mu\nu}_\nu B_\mu,$$

where $\Gamma^{\mu\nu}_\nu$ denote the standard Christoffel symbol. This means that it is not possible to apply effortlessly the available results for the wave/Klein-Gordon equation to the Dirac setting.

Let us now focus on two different possible choices for the metrics g.

i) *Asymptotically flat manifolds*: Let us take $n = 3$. We assume $g \in C^\infty(\mathbb{R}^3)$ to be "close" to the identity. That is, there exists a constant c_g and $\sigma \in (0, 1)$ such that for all $\alpha \in \mathbb{N}^3$, $|\alpha| = \alpha_1 + \alpha_2 + \alpha_3 \leq 3$ and all x

$$\left|\partial^\alpha (g_{ij}(x) - \delta_{ij})\right| \leq c_g \langle x \rangle^{-|\alpha|-1-\sigma}$$

where $\partial^\alpha = \partial^{\alpha_1}_{x_1} \partial^{\alpha_2}_{x_2} \partial^{\alpha^3}_{x_3}$. Then, it is proved in [6](Theorem 1.2) that the massless Dirac flow satisfies the Strichartz estimate:

$$\left\| e^{it\mathcal{D}} u_0 \right\|_{L^p_t(\mathbb{R}; L^q_x(\mathcal{M}))} \leq c \|u_0\|_{\dot{H}^s(\mathcal{M})}$$

for all wave admissible exponents, while in the massive case ($m > 0$) we have

$$\left\| e^{it(\mathcal{D} + m\alpha^0)} u_0 \right\|_{L^p_t(\mathbb{R}; L^q_x(\mathcal{M}))} \leq c \|u_0\|_{H^{s+\frac{1}{2}}(\mathcal{M})}$$

for all Schrödinger admissible exponents. In both cases, s is given by $s = \frac{1}{2} + \frac{1}{p} - \frac{1}{q}$, as in the flat case.

ii) *Smooth compact manifolds without boundary*: Let now $\mathcal{M}$ be a smooth compact Riemannian manifold without boundary of dimension $d \geq 2$ equipped with a spin structure. Differently from the cases above, we now estimate the L^p-norm in time restricted to a bounded interval $I \subset \mathbb{R}$. We have the following result (Theorem 2 in [4]):
for any wave admissible pairs (p, q)

$$\left\| e^{it(\mathcal{D} + m\alpha^0)} u_0 \right\|_{L^p_t(I; L^q_x(\mathcal{M}))} \leq c(I) \|u_0\|_{H^s(\mathcal{M})}$$

and for any Schrödinger admissible ones

$$\left\| e^{it(\mathcal{D} + m\alpha^0)} u_0 \right\|_{L^p_t(I; L^q_x(\mathcal{M}))} \leq c(I) \|u_0\|_{H^{s+\frac{1}{2p}}(\mathcal{M})}.$$

Moreover, we remark that in the latter case it is possibile to show that the obtained estimates are optimal for index $p = 2$ for the spheres of dimensions $d \geq 4$.

[(2)] Greek indexes are lowered and raised multiplying by $g_{\mu\nu}$ and $g^{\mu\nu}$.

References

[1] Bahouri H., *Littlewood-Paley Theory: A common Threat of Many Works in Nonlinear Analysis*. EMS Newsletter (2019).

[2] Bahouri H., Chemin J.-Y., Danchin R., "Fourier Analysis and Nonlinear Partial Differential Equations". Springer-Verlag, Vol. 115 (3-4), 2014.

[3] Bjorken J. D., Drell S. D., "Relativistic Quantum Mechanics". McGraw-Hill, International series in pure and applied physics, 1964.

[4] Cacciafesta F., Danesi E., Meng L., *Strichartz estimates for the half wave/Klein-Gordon and Dirac Equations on compact manifolds*. Math. Ann. 389 (2024), 3009–3042.

[5] Cacciafesta F., de Suzzoni A.-S., *Weak dispersion for the Dirac equation on asymptotically flat and warped product spaces*. Discrete Contin. Dyn. Syst. 39 (2019), 4359–4398.

[6] Cacciafesta F., de Suzzoni A.-S., Meng L., *Strichartz estimates for the Dirac equation on asymptotically flat manifolds*. Ann. Sc. Norm. Super. Pisa Cl. Sci. (2023).

[7] D'Ancona P., Fanelli L., *Strichartz and smoothing estimates of dispersive equations with magnetic potentials*. Comm. Part. Diff. Eq. 33 (4-6) (2008), pp. 1082–1112.

[8] Dirac P. A. M., *The Quantum Theory of the Electron*. Proceedings of the Royal Society of London, Vol. 117, No. 778 (1928), pp. 610-624.

[9] Hörmander L., "The Analysis of Linear Partial Differential Operators I: Distribution Theory and Fourier Analysis". Springer, Berlin, 2015.

[10] Keel M.,Tao T., *Endpoint Strichartz estimates*. Am. J .Math. 120(5) (1998), pp. 955–980.

[11] Linares F., Ponce G., "Introduction to Nonlinear Dispersive Equations". Universitext, Springer, New York, second ed., 2015.

[12] Parker L., Toms D., "Quantum Field Theory in Curved Spacetime: Quantized Fields and Gravity". Cambridge University Press, 2009.

[13] Tao, T., "Nonlinear dispersive equations: local and global analysis". CBMS Regional Conference Series in Mathematics (2006), Vol. 106, 373 pp..

[14] Thaller, B., "The Dirac Equation". Springer-Verlag, Texts and Monographs in Physics, 1992.

Topological Data Analysis (TDA) Basic concepts and applications

Cinzia Bandiziol [(*)]

In the last two decades, with the increasing need of analysing big amout of data, that usually are complex and of high dimension, it was revealed meaningfull and helpfull to discover new methodologies in order to provide new information from data. This has brought to the birth of Topological Data Analisys (TDA), whose aim is to extract intrinsic, topological features from data, related to the so called "shape of data". These kinds of features, collected in the so called Persistence Diagrams, has been winning in many different applications, mainly related to applied science, improving the performances of models and of classifiers, as in our context. Thanks to the strong theoretical basis behind, the TDA is very versatile and can be applied to data with a priori any kind of structure, as we will explain in the following. Therefore in literature there are a lot of applications of TDA to different fields like biology, chemistry, medicine, neuroscience, phisics only to name a few.

The main tool of TDA is the so called Persistent Homology, that allow us to extract persistent topoloical features from data. This method derived directly from the Algebraic topology, that is a branch of math that uses tools from linear algebra to study topological spaces in order to find its algebraic invariants. An example of such invariants are the homology groups. Intuitively, given a topological space X, the n homology group, $H_n(X)$, consists of the n-dimensional holes that characterize the space itself. In application, users usually consider only 0,1,2 dimensional holes as we will explained later.

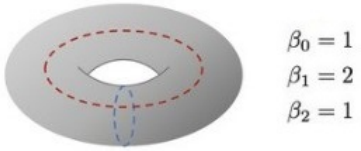


Figure 1: Thorus with its Betti Numbers

From a general point of view, if we have a topological space or simply a surface like a thorus, the aim is to count the number of connected components (0-dimensional holes),

[(*)]Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: bandizio@math.unipd.it . Seminar held on 20 June 2024.

cicles (1-dimensional holes) and cavities/voids (2-dimensional holes) with the idea that these numbers can indeed represent and characterize the space X from a qualitative and intrinsic point of view.

The numbers reported at the bottom of the figure represent such holes, the so called Betti number, for instance the rank of the homology groups previously mentioned. In the case of thorus, it is evident how there is only 1 connected component, 2 cicles as marked in the picture and obviously 1 cavity as tunnel inside the thorus. Now the idea is to understand how to extend this theory to be able to deal with discrete data as for example point clouds. For reach the purpose, it is needed to put some geometrical structure into data and this can be done thanks to the simplicial homology. This theory consists of applying homology to structures known as simplicial complexes, that are the generalization of triangulation of a topological space. Then, starting from a discrete dataset, as point cloud, the idea is to consider not only one simplicial complex build upon points, but a nested sequence of them, always more and more complex, and see which topological features appear and disappear through the evolution. This is the idea behind Persistent Homology.

The first concept to introduce is the simplicial complex,

Definition 1 A **simplicial complex** K consists of a set of simplices of certain dimensions and has to meet the following conditions:

- Every face of a simplex in K is also in K
- The non-empty intersection of any two simplices $\sigma_1, \sigma_2 \in K$ is a face of both σ_1 and σ_2

Figure 2: Simplices of low dimensions

We recall that the simplices of lower dimensions are as in the picture: a vertex (0 dim simplex), and edge (1 dim simplex), a triangle (2 dim simplex) and a thetraedron (3 dim simplex).

Another meaningfull ingredient to be able to extract topological information is the filtration. First, we introduce a **filtration** funtion,

$$f : K \to \mathbb{R} \text{ such that } f(\tau) \leq f(\sigma), \forall \tau \subset \sigma \text{ in } K.$$

Considering the simplicial complex of sublevel set $K_a = K(f^{-1}((-\infty, a]))$, where we say that a simplex $\sigma \in K_a \iff f(\sigma) \leq a$, we end up with the building of a filtration

$$\emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_{n-1} \subseteq K_n = K \textbf{ (Filtration)}$$

that goes from the empty simplicial complex to the whole K. Studing the evolution, one can recognize the appearence and the disappearance of features. For example, a connected component can appear in K_j and appear for the last time in K_{j+k} gluing then togheter with another one. Such a topological feature can be denoted using the corresponding indexes of the related simplicial complexes. In the aforementioned example, such feature is $p = (j, j + k)$. Taking the difference between them one obtains the lifetime of the feature, called **persistence**. All these points can then be collected in a multiset of points, $\{(b_i, d_i) \in \mathbb{R}^2 | i \in I\}$ the so called **Persistence Diagram** (PD).

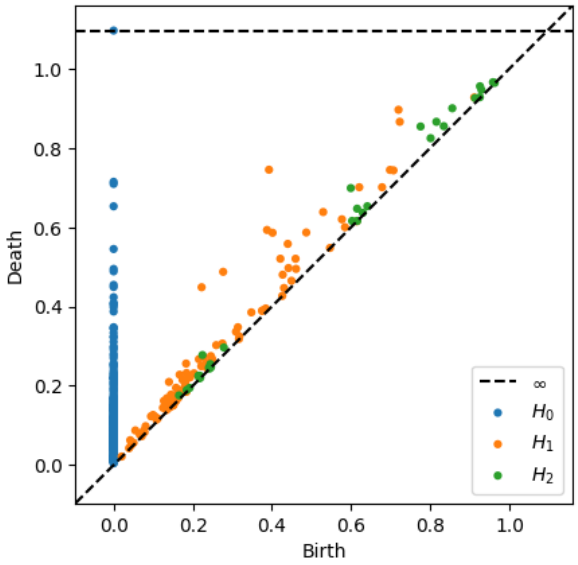


Figure 3: An example of Persistence Diagram

The Figure 3 is an example of PD collecting features of dimension 0 (in blue), of dimension 1 (in orange) and of dimension 2 (in green). Points close to diagonal represent features with short lifetime, and so usually are concern with noise, instead features far away are indeed relevant and meaningfull and, based on applications, one can decide to consider both or only the most relevant.

An interesting property of PD is its robustness or stability to noise. Before citing the most important result, we need to introduce some notions of distances in order to compare PDs each other.

Definition 2 Given two non empty sets $X, Y \subset \mathbb{R}^2$ with the same cardinality, the **Hausdorff distance** is

$$d_H(X,Y) = \max\{\sup_{x\in X}\inf_{y\in Y}\|x-y\|_\infty, \sup_{y\in Y}\inf_{x\in X}\|x-y\|_\infty\}$$

and the **Bottleneck distance** is defined as

$$d_B(X,Y) = \inf_{\gamma} \sup_{x \in X} \|x - \gamma(x)\|_\infty$$

where γ varies among all the bijections $\gamma : X \to Y$ and $\|\|_\infty$ is the usual supnorm.

Then

Theorem 1 *Let X and Y be finite subset of a metric space (M, d_M). Then*

$$d_B(D(X), D(Y)) \leq d_H(X,Y)$$

where $D(X), D(Y)$ are persistence diagrams related to X, Y.

The previous theorem, introduced in [1], means that if the original set is affected by noise and the distance between the original one is lower than ϵ than the related PDs differs for at most the same quantity.

The TDA can be applied to data with different discrete structures. First we see the example of point clouds.

In the case of point cloud, we can infer some geometrical structure through the Vietoris Rips complex.

Definition 3 Let $(\mathcal{X}, d)$ denote a metric space from which the samples are taken. Then the **Vietoris-Rips complex** for $\mathcal{X}$, attached to the parameter ϵ, denoted by $VR(\mathcal{X}, \epsilon)$, will be the simplicial complex whose vertex set $\mathcal{X}$ and where $\{x_0, \ldots, x_k\}$ spans a k-simplex if and only if $d(x_i, x_j) \leqslant 2\epsilon$ for all $0 \leqslant i, j \leqslant k$.

At ϵ varies, we obtain the Vietoris-Rips complexes that provide the elements of a **filtration** $\emptyset = K_1 \subset K_2 \subset \cdots \subset K_r$ with $K_i = VR(\mathcal{X}, \epsilon_i)$.

Graphically the PH pipeline is represented by Figure 4.

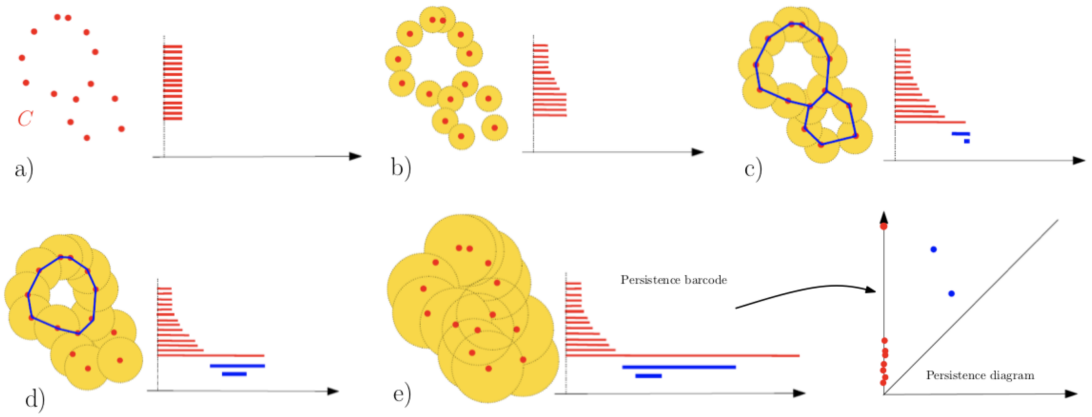


Figure 4: Persistent pipeline for point cloud data

The birth and the death of persistent features, in this example, is collected in the so called **Persistence Barcode** (PB), on the right of each subfigures. In subfigure a, we start with only n points that correspond to n different connected components and to n lines in the PB. It is evident that the point cloud is characterize by 1 connected component and two cicles, one bigger than the other one and we want to recover these information with PH. In step b, the radius of each ball increases and some of them intersect each other. Following the definition of VR complex, if it happens two points are connected with an edge and therefore the number of connected components decreases. In fact only 5 lines/connected components are still alive. In c, the radius further increases and we end up with only 1 connected component still alive and the apparence of two cicles, as expected. In d, only the bigger one cicle survive and finally the extreme case in subfigure e, has a fully connected structure and give no any further information. The final PB effectively shows the topological features predicted that can easily traslate into the PD, at the bottom right of the picture.

TDA is indeed versatile, and following more or less the same procedure, one can apply to other structures. In the case of grayscale image, one deals with a set of ordered pixels with the corresponding gray value as filtration function.

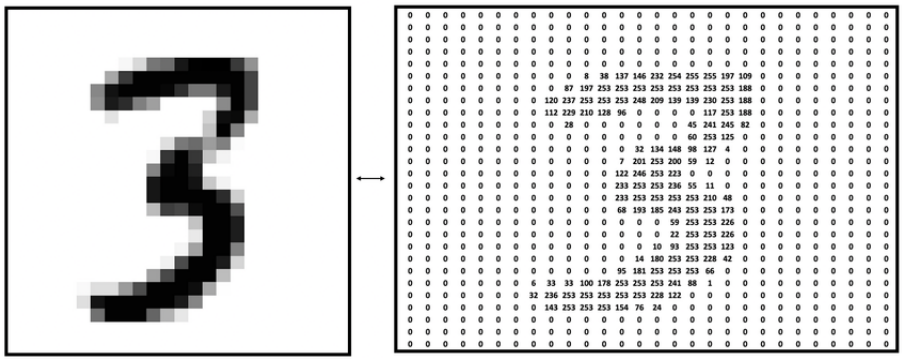


Figure 5: An image from MNIST

An example of handwritten digit taken from MNIST dataset, well known and really common in the classification community.

In the case of graphs, or better **undirected graph**, $G = (V, E)$ with V the set of vertices and E the set of edges, we consider $f : V \to \mathbb{R}$ defined on its vertices, defining the **sublevel vertex function based** filtration by the nested sequence of subgraphs $G_\delta = (V_\delta, E_\delta)$ where $V_\delta = \{v \in V | f(v) \leq \delta\}$ and $E_\delta = \{(u, v) \in E | \max\{f(u), f(v)\} \leq \delta\}$. An example is given in Figure 6, where it is evident how the structure appears slowly along the filtration.

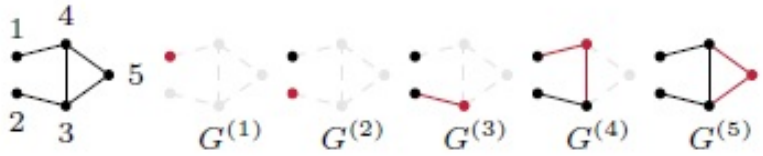


Figure 6: Example of graph filtration

The last example is related to 1-dimensional time series, that are object like $\{x_t \in \mathbb{R} | t = 1, \dots, T\}$. Thanks to the Taken's embedding, they can be translated into point cloud. If we fix the values for two parameters: $\tau > 0$ the delay parameter and $d > 0$ the dimension, in a suitable way, we end up with a subset of points in $\mathbb{R}^d$ composed by $v_i = \{x_i, x_{i+\tau}, \dots, x_{i+(d-1)\tau}\}$ for $i = 1, \dots, T - (d-1)\tau$.

For the purpose of the talk, our aim is to solve the classification problem, or better, to be able to classify PDs. The classification problem is indeed common in the Machine Learning or Deep Learning community.

Let $\mathcal{X} = \{x_i\}_{i=1,\dots,m}$ a dataset as subset of $\mathbb{R}^d$ with labels $\{y_i\}_{i=1,\dots,m}$ where $y_i \in \mathcal{Y} = \{-1, 1\}$ (binary problem). The classification task consists of finding out a function/classifier that, based on input data $(x_i, y_i)_{i=1,\dots,m}$, is able to predict the label of an unseen point $\bar{x}$. There exists several methods that solve such a problem as for example:

- SVM
- KNN
- Random Forest
- Neural Network
- ...

but for our purpose we are interested in **Support Vector Machine** (SVM). The geometric idea behind the method is well described in the picture.

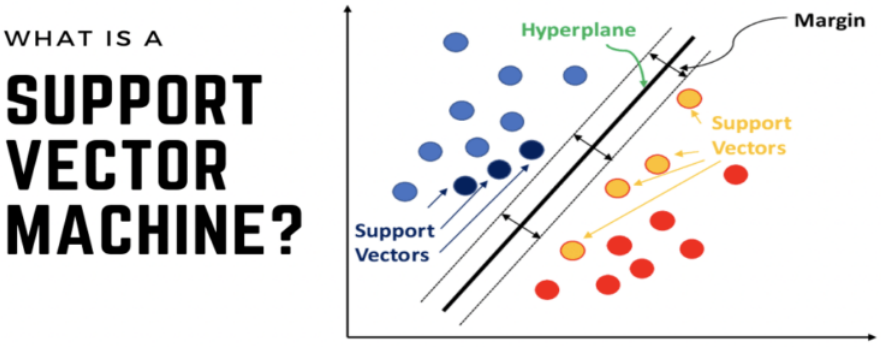


Figure 7: Geometric idea of SVM

The aim is to find out the hyperplane that is able to separate ,in the best possible way, points that belong to different classes and from here the name separating hyperplane. The best possible means that it separates the two classes with the higher margin, that is the distance between the hyperplane and the points of both classes. After some computations, such optimization problem turns out to have the following formulation.

The **SVM optimization problem** is given by [2]

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^m} \quad & W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\ \text{s. to} \quad & \sum_{i=1}^{m} \alpha_i y_i = 0 \\ & \alpha_i \geqslant 0 \ \forall i = 1, \dots, m \end{aligned}$$

$\alpha_i > 0$ are called **Support Vectors**, from here the name Support Vector Machine and $\langle \cdot, \cdot \rangle$ denotes inner product in $\mathbb{R}^d$. This formulation is able to face satisfactorily the classification task if data are linearly separable. In applications it happens frequently that data aren't linearly separable and so it is needed to introduce some nonlinearity and moving in higher dimensional space where, hopefully, that happens. This can be achieved with the use of kernels. Starting from the original dataset or feature space $\mathcal{X}$, the theory tells to introduce a feature map $\Phi : \mathcal{X} \to \mathcal{H}$ that moves data from $\mathcal{X}$ to an Hilbert space of function $\mathcal{H}$. The kernel is then defined as $\kappa(x, \bar{x}) := \langle \Phi(x), \Phi(\bar{x}) \rangle_{\mathcal{H}}$ (**kernel trick**). With kernels the optimization problem becomes

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^m} \quad & W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \kappa(x_i, x_j) \\ \text{s. to} \quad & \sum_{i=1}^{m} \alpha_i y_i = 0 \\ & \alpha_i \geqslant 0 \ \forall i = 1, \dots, m \end{aligned}$$

where kernel represents a generalization of the inner product in $\mathbb{R}^d$. We are interested in classifying PDs and obviously we need suitable definitions for kernels for PDs, the so called Persistence Kernels. In what follows we denote with $\mathfrak{D}$ the set of PDs.

The first kernel was described in [3]. The main idea is to compute feature map as the solution of a PDE. We consider $\Omega_{ad} = \{\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2 : x_2 \geqslant x_1\}$ and we denote with $\delta_{\mathbf{x}}$ the Dirac delta centered at $\mathbf{x}$. For a given $D \in \mathfrak{D}$, we consider the solution $u : \Omega_{ad} \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}$, $(\mathbf{x}, t) \mapsto u(\mathbf{x}, t)$ of the following PDE:

$$\begin{aligned} \Delta_{\mathbf{x}} u = \partial_t u \quad & \text{in } \Omega_{ad} \times \mathbb{R}_{\geqslant 0} \\ u = 0 \quad & \text{on } \partial\Omega_{ad} \times \mathbb{R}_{\geqslant 0} \\ u = \sum_{\mathbf{y} \in D} \delta_{\mathbf{y}} \quad & \text{on } \Omega_{ad} \times 0 \end{aligned}$$

The feature map $\Phi_\sigma : \mathfrak{D} \to \mathrm{L}^2(\Omega_{ad})$ at scale $\sigma > 0$ at D is defined as $\Phi_\sigma(D) = u\big|_{t=\sigma}$. This map yields the **Persistence Scale Space Kernel** (**PSSK**) K_σ on $\mathfrak{D}$ as:

$$K_\sigma(D, E) = \langle \Phi_\sigma(D), \Phi_\sigma(E) \rangle_{\mathrm{L}^2(\Omega_{ad})}.$$

But since it is known an explicit formula for the solution u, the kernel takes the form

$$K_\sigma(D, E) = \frac{1}{8\pi\sigma} \sum_{\mathbf{y} \in D, \mathbf{z} \in E} \exp(-\frac{\|\mathbf{y} - \mathbf{z}\|^2}{8\sigma}) - \exp(-\frac{\|\mathbf{y} - \bar{\mathbf{z}}\|^2}{8\sigma})$$

where $\mathbf{z} = (a, b)$, $\bar{\mathbf{z}} = (b, a)$, for any $D, E \in \mathfrak{D}$.

In [4], the authors introduce a new kernel where the idea is to replace each PD with a discrete measure. Starting with a strictly positive definite kernel, as for example the gaussian one $\kappa_G(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$, $\sigma > 0$ we denote the corresponding RKHS with $\mathcal{H}_{\kappa_G}$.

If $\Omega \subset \mathbb{R}^d$, we denote with $M_b(\Omega)$ the space of finite signed Radon measures and

$$E_{\kappa_G} : M_b(\Omega) \to \mathcal{H}_{\kappa_G}, \mu \mapsto \int_\Omega \kappa_G(\cdot, x) d\mu(x).$$

For any $D \in \mathfrak{D}$, if $\mu_D^w = \sum_{x \in D} w(x)\delta_x$, where the weight function satisfies $w(x) > 0$ for all $x \in D$ then

$$E_{\kappa_G}(\mu_D^w) = \sum_{x \in D} w(x)\kappa_G(\cdot, x).$$

The **Persistence Weight Gaussian Kernel** (**PWGK**) is defined as

$$K_G^w(D, E) = \exp\left(-\frac{1}{2\tau^2} \|E_{\kappa_G}(\mu_D^w) - E_{\kappa_G}(\mu_E^w)\|^2_{\mathcal{H}_{\kappa_G}} \right), \ \tau > 0$$

for any $D, E \in \mathfrak{D}$.

Another possible choice for κ was introduced in [5].

We consider μ and ν two nonnegative measures on $\mathbb{R}$ such that $\mu(\mathbb{R}) = r = |\mu|$ and $\nu(\mathbb{R}) = r = |\nu|$, we recall that the 1-Wasserstein distance for nonnegative measures is defined as

$$\mathcal{W}(\mu, \nu) = \inf_{P \in \Pi(\mu, \nu)} \int \int_{\mathbb{R} \times \mathbb{R}} |x - y| dP(x, y)$$

where $\Pi(\mu, \nu)$ is the set of measures on $\mathbb{R}^2$ with marginals μ and ν.

Definition 4 Given $\theta \in \mathbb{R}^2$ with $\|\theta\|_2 = 1$, let $L(\theta)$ denote the line $\{\lambda\theta | \lambda \in \mathbb{R}\}$ and let $\pi_\theta : \mathbb{R}^2 \to L(\theta)$ be the orthogonal projection onto $L(\theta)$. Let $D, E \in \mathfrak{D}$ and let $\mu_D^\theta := \sum_{p \in D} \delta_{\pi_\theta(p)}$ and $\mu_{D\Delta}^\theta := \sum_{p \in D} \delta_{\pi_\theta \circ \pi_\Delta(p)}$ and similarly for μ_E^θ and $\mu_{E\Delta}^\theta$ where π_Δ is the orthogonal projection onto the diagonal. Then, the **Sliced Wasserstein distance** is

$$SW(D, E) = \frac{1}{2\pi} \int_{\mathbb{S}_1} \mathcal{W}(\mu_D^\theta + \mu_{E\Delta}^\theta, \mu_E^\theta + \mu_{D\Delta}^\theta) d\theta$$

Thus, the **Sliced Wasserstein Kernel** (**SWK**) is defined as

$$K_{SW}(D,E) := \exp\left(-\frac{SW(D,E)}{2\sigma^2}\right),\ \sigma > 0$$

for any $D, E \in \mathfrak{D}$. The last kernel available in literature is the Fisher Information Kernel introduced in [6], where the idea of the authors is to replace each PD with a probability distribution. So if $D \in \mathfrak{D}$

$$\rho_D(x) := \frac{1}{Z}\sum_{u\in D} N(x; u, \sigma I)$$

where N is a gaussian function, $Z = \int \sum_{u\in D} N(x; u, \sigma I)dx$ and I is the identity matrix. If we denote $\mathbb{P} = \{\rho | \int \rho(x)dx = 1, \rho(x) \geq 0\}$, we recall

Definition 5 Given two element in $\rho_i, \rho_j \in \mathbb{P}$, the **Fisher Information Metric** is

$$d_{\mathbb{P}}(\rho_i, \rho_j) = \arccos\left(\int \sqrt{\rho_i(x)\rho_j(x)}dx\right),$$

and we extend it to

Definition 6 Let $D, E \in \mathfrak{D}$. The **Fisher Information Metric** between D and E, is defined as

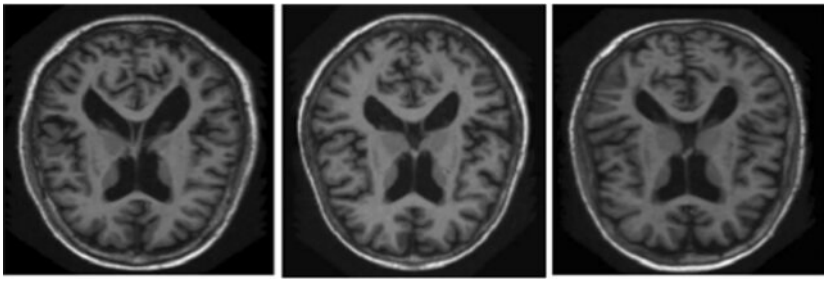
$$d_{FIM}(D,E) := d_{\mathbb{P}}(\rho_{D\cup E_\Delta}, \rho_{E\cup D_\Delta})$$

where $D_\Delta := \{\Pi_\Delta(u) | u \in D\}$ and Π_Δ is the projection on the diagonal $\Delta = \{(a,a) | a \geq 0\}$.

The **Persistence Fisher Kernel** (**PFK**) is then defined as

$$K(D,E) := \exp(-td_{FIM}(D,E)),\ t > 0, \text{ for any } D, E \in \mathfrak{D}.$$

Finally we report an application to real world data, taken from the world of neuroscience. As in [7], we consider the OASIS Brains Dataset that is a compilation of MRI and PET images and related clinical data and the aim is to predict if a person has the Alzheimer disease or not.



For this purpose, thanks to the estimation of corthical thickness on 32 points in both right and left hemisphere of the brain, we build the VR complexes and extract the PD collecting persistent features of dimension 1 and 2. Then we proceed in solving the classification task and evaluate the performances using the Accuracy. For balanced and binary classification problems, the accuracy is defined as

$$\text{accuracy} = \frac{\text{number of points classified correctly}}{\text{total number of points}}$$

and thus it is a measure of how good the classifier is. The aim is that such a quantity must be closer to 1. Results in table show that kernels are all quiet good in classification, achieving accuracy of more than 0.7.

Kernel	**Accuracy**
PSSK	0.78
PWGK	0.74
SWK	0.76
PFK	0.74

References

[1] D. Cohen-Steiner, H. Edelsbrunner, J. Harer, *Stability of persistence diagrams.* Discrete & computational geometry 37 (1) (2007), pp. 103–120.

[2] B. Scholkopf, A.J. Smola, "Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond". The MIT Press, 2002.

[3] J. Reininghaus, S. Huber, U. Bauer, R. Kwitt, *A Stable Multi-Scale Kernel for Topological Machine Learning.* Proceedings of the IEEE conference on computer vision and pattern recognition (2015), pp. 4741–4748.

[4] G. Kusano, K. Fukumizu, Y. Hiraoka, *Kernel method for persistence diagrams via kernel embedding and weight factor.* The Journal of Machine Learning Research vol. 18, no. 1 (2017), pp. 6947–6987.

[5] M. Carriere, M. Cuturi, S. Oudot, *Sliced Wasserstein kernel for persistent diagrams.* International Conference on Machine Learning, PMLR 2017, pp. 664–673.

[6] T. Le, M. Yamada, *Persistence fisher kernel: A riemannian manifold kernel for persistence diagrams.* arXiv preprint `arXiv:1802.03569` (2018).

[7] S. De Marchi, F. Lot, F. Marchetti, D. Poggiali, *Variably Scaled Persistence Kernels (VSPKs) for persistent homology applications.* Journal of Computational Mathematics and Data Science, Volume 4, August 2022.