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

Scuole di Dottorato in Matematica Pura e Computazionale

Seminario Dottorato 2024/25

Preface

This document offers an overview of the activity of Seminario Dottorato 2024/25.

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, 2025

Corrado Marastoni, Tiziano Vargiolu

Abstracts (from Seminario Dottorato's webpage)

Thursday 7 November 2024

Hawkes processes in cyber-risk analysis: modelization and optimal security investment

Beatrice Ongarato (Padova, Dip. Mat.)

With the rapid growth of the digital economy in recent years, cyber-risk has emerged as one of the most relevant and rapidly growing sources of risk. We provide an overview of the main concepts related to cyber-risk and examine the challenges involved in its quantification and modeling. We introduce Hawkes processes and explain their applicability in capturing the dynamics of cyber-attacks. Lastly, we present an ongoing project aimed at determining the optimal cyber-security investment strategy for an organization facing cyber-attacks. The problem is framed as a stochastic control problem with jumps and is addressed using Hamilton-Jacobi-Bellman (HJB) techniques. We introduce the main tools needed to solve this type of problem and show some preliminary numerical results.

Thursday 21 November 2024

Bridging Enumerative Geometry and Quantum Integrable Hierarchies

Ishan Jaztar Singh (Padova, Dip. Mat.)

Enumerative geometry explores the use of combinatorial and intersection theory techniques to solve counting problems in algebraic geometry. Integrable hierarchies, in contrast, consist of infinite sequences of partial differential equations with symmetries that have significance in mathematical physics. Both fields have seen substantial developments over the past half-century. This talk will focus on the infamous Witten-Kontsevich theorem, which establishes a deep connection between topological invariants of the moduli space of curves and the Korteweg-de Vries hierarchy. I will attempt to offer intuitive motivation and a formal statement of the theorem, and, time permitting, discuss its generalizations and the role of quantum hierarchies in this context.

Thursday 5 December 2024

Dynamics of Environment-Embedded Quantum Systems: An Introduction

Pietro De Checchi (Padova, Dip. Mat.)

Closed quantum systems are an idealization, their time evolution described by the Schrödinger Equation, i.e. by the action of unitary operators. The physics of a realistic quantum system, on the other hand, is bound to be disturbed by the environment in which it is naturally embedded and with which it inevitably interacts. The dimension of the space needed to fully describe the composite system increases, as one would have to include all, possibly infinite, environments variables, leading

to intractable problems. To reduce the system to a smaller subspace of interest and to describe its correct dynamics, many strategies have been developed. These systems have in general non-unitary dynamics and are known as Open Quantum Systems. During the talk, we will introduce some of the main approaches based on various techniques, from dynamical semigroup generators, stochastic unravellings and bottom-up modelling.

Thursday 19 December 2024

Representations of Quivers over Rings: Merging Commutative and Non-Commutative Results

Enrico Sabatini (Padova, Dip. Mat.)

In the vast universe of representation theory there are two very separate and different worlds: commutative rings and finite dimensional (non-commutative) algebras. The problem of characterising certain subcategories, like many other problems, has been solved in both fields. However, the main techniques used for one context are generally not transferable to the other. Recently, some authors have focused their interest on a special kind of algebras that partially merge the two fields. Here, the apparently different results have a surprising generalisation and a unifying proof. In this talk, I will give an overview of the two fields mentioned above, describe their main features and give an idea of what allows such characterisations; avoiding all the technicalities. Finally, I'll show how the generalisation works with the aid of some interesting examples.

Thursday 9 January 2025

Deep Unfolding: Bridging Optimization and Neural Network Interpretability

Erik Chinellato (Padova, Dip. Mat.)

Deep neural networks (DNNs) have revolutionized numerous fields due to their powerful ability to learn complex representations. However, their black-box nature and lack of interpretability in architecture and weight design remain significant challenges. After an introductory segment on DNNs and backpropagation learning, this seminar introduces the Deep Unfolding method as a promising alternative, bridging the gap between data-driven learning and model-based optimization. By unrolling iterative optimization algorithms into structured neural network architectures, Deep Unfolding provides a principled approach to network design, enabling interpretability and theoretical insights into their operation. We will explore how this method leverages domain knowledge, achieves faster convergence, and enhances performance in resource-constrained scenarios. The session will highlight many wide-ranging practical applications of Deep Unfolding, covering audio source separation and recognition, image denoising and state estimation.

Thursday 23 January 2025

Modeling Dark Matter: a Dynamics Study

Gaia Marangon (Padova, Dip. Mat.)

Dark matter is one of the most relevant and fascinating open problems in modern astrophysics. Since it cannot be directly observed, modeling it requires a balanced mix of physical intuition, mathematical deduction, and comparison with indirect experimental data. In this talk, I will briefly introduce the physical context motivating our research, specifically the problem of dark matter distributions around galaxies. Starting from the Schrödinger-Poisson system, the most commonly used model for dark matter dynamics, I will outline the main directions our work has taken. I will focus on two key aspects. First, I will discuss the issue of stationary states, ranging from numerical properties to comparison with experimental data. Then, I will propose a relativistic generalization of the model, the Klein-Gordon - Wave system. Its treatment by Hamiltonian perturbative techniques shows the potential of mathematical physics tools in building a comprehensive and reliable model.

Thursday 6 February 2025

Mixing times and cutoffs for Markov chains

Giacomo Passuello (Padova, Dip. Mat.)

How long does it take to shuffle a deck of 40 cards? This simple question, together with the seminal work of Aldous and Diaconis on the cutoff phenomenon, has generated, in the last 40 years, a rich research area in the field of discrete probability. A cutoff is a dynamical phase transition for a random process, which appears as the size of the system becomes large. It occurs when the distance to equilibrium of the process abruptly drops from its maximum value to zero at a critical time scale. Establishing the occurrence of the cutoff is a delicate matter, which may require a precise understanding of the spectral and diffusive properties of the underlying system. In this talk, I will review some basic concepts on Markov chains and their convergence to the stationary equilibrium. After that, I will introduce the concept of mixing time and discuss bounds on its limiting behaviour. Finally, I will focus on the cutoff phenomenon and present some results on the mixing time of the simple random walk on a directed random graph.

Thursday 20 February 2025

Mean Field Turnpike Theorems

Denis Shishmintsev (Padova, Dip. Mat.)

In the study of Mean Field Games (MFG), the Turnpike Property plays a crucial role in understanding the asymptotic behavior of large populations of agents. This property suggests that, for sufficiently long time horizons, the optimal trajectories of agents in a dynamic system converge to a steady-state or "turnpike" region, where their strategies remain approximately constant. The presence of the turnpike reflects the system?s tendency to stabilize and suggests that most of the

time, agents will follow similar paths despite starting from different initial conditions. In this introductory talk we investigate the turnpike property in the context of Lagrangian and Eulerian formulations of MFGs, which describe the agents either through their individual trajectories (Lagrangian) or through a distribution function over space (Eulerian). In both frameworks, we explore how the turnpike emerges and its implications for the long-term dynamics of the system, aiming at key applications in economics, control theory, and multi-agent systems.

Thursday 6 March 2025

An Integer Linear Programming Model for the Dynamic Airspace Configuration problem

Martina Galeazzo (Padova, Dip. Mat.)

Given central role of aviation as a transportation network and its remarkable economic impact, the air traffic demand is bound to increase. High traffic density in a given airspace region can cause safety issues and difficulties in monitoring tasks that can, in turn, result in flight delays. It is therefore crucial to efficiently organize the airspace structure to avoid under- and overloaded areas of the airspace. We begin by describing how the airspace is structured, introducing the concept of sector and configuration and their capacity, and how to quantify the air traffic excess associated to a configuration. We will then introduce Dynamic Airspace Configuration as a method for optimally meeting the air traffic demand by adopting different configurations over time, thus determining a sequence of configurations (configuration plan); we impose that such a sequence also satisfies some operational restrictions that smooth the configuration dynamics, as to avoid, e.g., too frequent switching between configurations. After recalling the basic definitions and tools of (Integer) Linear Programming, we will present an Integer Linear Programming model that provides a configuration plan that minimizes the traffic excess for a given time frame, and a polyhedral study that explains its good computational performance. We conclude by showing the numerical results obtained by testing the model on five days of historical data (summer 2019) over the Madrid Area Control Center, with a focus on the comparison of different time discretizations and different restrictions on the configurations' transitions.

Thursday 20 March 2025

Resonances and quasi-collisions in the Three–Body Problem

Xiang Liu (Padova, Dip. Mat.)

Mean motion resonance, a phenomenon occurring when two celestial bodies have orbital periods in a commensurable ratio, plays a pivotal role in both stabilizing and destabilizing motions within our Solar System. For highly eccentric orbits, quasi-collisions become a significant factor. When such eccentric orbits are trapped in resonance, perturbations can induce chaotic motions, leading to rapid changes in orbital elements and transitions of different dynamical states. This presentation will begin by introducing the concept of mean motion resonance within the framework of the restricted three-body problem. Subsequently, we will explore the application of Hamiltonian perturbation theory for low-eccentricity orbits. Finally, we will demonstrate the limitations of this

theory when applied to highly eccentric orbits.

Wednesday 2 April 2025

Lavrentiev Phenomenon and semicontinuous envelopment for integral functionals

Tommaso Bertin (Padova, Dip. Mat.)

The first part of the talk is devoted to introduce some basilar elements of the Direct Method of Calculus of variations. In particular we will see the Tonelli's Theorem about the strictly relationship between lower semicontinuity of an integral functional and convexity of the Lagrangian. In the second part we will explore the so called "Lavrentiev Phenomenon", i.e. the possibility for a functional to reach an infimum in a dense subset strictly greater than the infimum in the original set. We will see some classical examples and some recent results to avoid the Phenomenon. In particular we will focus on non convex and non continuous Lagrangian.

Tuesday 15 April 2025

Let's play symplectic billiards!

Alessandra Nardi (Padova, Dip. Mat.)

A mathematical billiard is a dynamical system describing the motion of a mass point (the billiard ball) inside a planar region (the billiard table). The ball moves with constant speed and without friction, following a rectilinear path. The straightforwardness and versatility of this model have made mathematical billiards an object of interest in many different contexts. Indeed, depending on the shape of the billiard table, they show a wide range of dynamical behaviors such as integrability, regularity, and chaoticity. Integrability remains an unanswered property, and the celebrated Birkhoff conjecture remains open. In 2018, P. Albers and S. Tabachnikov introduced a new interesting class of billiards, called symplectic billiards, as a natural variation of Birkhoff billiards with the inner area – instead of the length – as generating function. This talk will present the symplectic billiards dynamics and focus on recent rigidity results. Talk based on joint works with L. Baracco and O. Bernardi.

Thursday 8 May 2025

Inductive Methods in the Representation Theory of Finite Groups of Lie Type: An Introduction via GL(n,q)

Elena Collacciani (Padova, Dip. Mat.)

Representation theory of groups is the area of mathematics that studies how abstract groups can act on vector spaces – in other words, how to realize groups as collections of linear transformations. Among the many families of finite groups, those of Lie type form a particularly important class, appearing naturally in the classification of finite simple groups. In this talk, after a brief overview of classical results on the representation theory of finite groups, I will offer a glimpse into the

techniques used to construct and classify irreducible representations of finite groups of Lie type. To convey the core ideas while avoiding heavy technicalities, I will focus on the case of the General Linear group over a finite field. Special emphasis will be placed on the role of inductive methods - a fundamental paradigm in representation theory – which often reduce complex algebraic problems to more manageable combinatorial ones. To help develop intuition, I will also present concrete examples of the main objects involved.

Thursday 22 May 2025

Parameter estimation of integrated fractional Brownian motions with application to energy markets

Marco Mastrogiovanni (L'Aquila, Dip. Ing. Inf., Comp. Sci. e Mat.)

We investigate the statistical properties of time-averaged fractional Brownian motion (fBm), which naturally arises in the modeling of time series subjected to averaging transformations. The main motivation comes from electricity markets, where daily prices are typically obtained by averaging high-frequency data. While fBm-based models are popular in modeling electricity prices, treating averaged prices as direct realizations of fBm is theoretically inconsistent. Instead, time-integrated versions of fBm should be used to accurately reflect the effects of averaging. This seminar begins with an overview of electricity markets and an introduction to fractional Brownian motion (fBm) and its main characteristics. We then introduce the integral-mean process of fBm and analyze the impact of time-averaging on its properties. Using ergodic theory, we construct strongly consistent estimators for the Hurst parameter adapted to the averaged process and validate them through an extensive simulation study. Next, we extend our approach to linear combinations of two distinct timeaveraged fBm processes, again estimating the relevant parameters. Finally, we apply our methodology to empirical electricity spot price data and discuss potential future developments in this research area. (This is joint work with Yuliya Mishura, Stefania Ottaviano and Tiziano Vargiolu.)

Thursday 5 June 2025

Parallel parking 101

Marco Di Marco (Padova, Dip. Mat.)

Did you know that when you parallel-park your car, you're actually invoking the Chow-Rashevskii Theorem, a cornerstone of sub-Riemannian geometry? The non-commutativity of the car's "allowed directions" not only lets you move orthogonally to the road (and thus pull into a parking space) but also gives rise to striking phenomena in more "theoretical" settings: submanifolds that are perfectly regular from the intrinsic viewpoint of sub-Riemannian Heisenberg groups yet look fractal when seen through Euclidean lenses. Such subtleties make it difficult, in the sub-Riemannian setting, to prove classical results like Stokes' Theorem and Stepanov's Theorem, to establish fine properties of SBV functions, or even to tackle the seemingly simple task of computing the diameter of a ball. After an introduction to sub-Riemannian geometry and Heisenberg groups, I will

outline some of the key ideas behind the proofs of the aforementioned results. (Talk based on joint works with S. Don, A. Julia, S. Nicolussi Golo, A. Pinamonti, G. Somma, D. Vittone, and K. Zambanini.)

Thursday 19 June 2025

A controlled excursion into the Hamilton-Jacobi equation: from classics to mean field models

Giacomo Ceccherini Silberstein (Padova, Dip. Mat.)

Optimal control theory is a branch of Calculus of Variations aiming to guide efficiently a system to achieve a specific goal, minimizing a given "cost" along the way. In this seminar, we'll embark on a controlled excursion into the Hamilton-Jacobi equation, a powerful partial differential equation that encodes optimality conditions for this variational problem. We will start by presenting the classical Euclidean setting, where the system's state space is finite-dimensional and familiar. Then, we'll extend these fundamental ideas to the more complex mean field setting, where the dynamics play out on the space of probability measures, allowing us to understand collective behaviors in large systems.

Hawkes Processes in Cyber-Risk Analysis: Modelization and Optimal Security Investment

Beatrice Ongarato (*)

Abstract. With the rapid growth of the digital economy in recent years, cyber-risk has emerged as one of the most relevant and rapidly growing sources of risk. We provide an overview of the main concepts related to cyber-risk and examine the challenges involved in its quantification and modeling. We introduce Hawkes processes and explain their applicability in capturing the dynamics of cyber-attacks. Lastly, we present an ongoing project aimed at determining the optimal cyber-security investment strategy for an organization facing cyber-attacks. The problem is framed as a stochastic control problem with jumps and is addressed using Hamilton-Jacobi-Bellman (HJB) techniques. We introduce the main tools needed to solve this type of problem and show some preliminary numerical results.

1 Introduction

Cyber-risk has become a growing concern for businesses and institutions worldwide, with cyber-attacks and data breaches being ranked first in the top ten list of global risks by the 2023 AON Global Risk Management Survey reports[(1)]. Moreover, according to IBM, the global average cost of a data breach has raised to almost 5M USD in 2024, more than 10% higher with respect to the previous year[(2)]. Cyber-attacks are a threat to every industry: from healthcare to finance, from government to education, and potentially for every private company. It is urgent and essential that companies adequately protect themselves against cyber-attacks, which otherwise could cause enormous and irreparable damage. As noted in [17], [12], [7], addressing cyber-risk involves unique challenges that must be considered when quantifying and managing this specific type of risk. In particular, we highlight the following:

- **Limited historical data**: The emerging nature of this risk and reporting bias - companies are often hesitant to disclose incidents to protect their reputation - pose

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: beatrice.ongarato@phd.unipd.it . Seminar held on 7 November 2024.

(1)Source: https://www.aon.com/en/insights/reports/global-risk-management-survey/top-risks-facing-financial-institutions.

(2)Source: https://www.ibm.com/reports/data-breach.

challenges for building a reliable databases.

- **Dynamic risk type**: Cyber-risk rapidly evolves together with technology, making difficult to use past data for modelling future attacks.
- **Interdependence and accumulation risks**: The interconnected nature of digital infrastructures induce a dependence structure within and across company networks.
- **Complex impact determination**: It is difficult to quantify the economic consequences of a cyber incident.

This work explores two key aspects of cyber-risk analysis: the modeling of cyber-attacks and the optimization of an entity's investment in cyber-risk management. In Section 2, we provide an introduction to Hawkes processes and explain that they are an appropriate class of processes for describing cyber-attacks. In Section 3, we formulate an optimal security investment problem using stochastic control methods, in order to determine the optimal strategy to protect an entity from cyber-attacks. In Section 4, we conduct numerical experiments to illustrate the practical relevance of our model, while Section 5 presents our conclusions.

2 Modelization of cyber-attacks

For the majority of the following definitions, we refer to [11, Sections 2 and 3].

2.1 Counting Processes

Cyber-attacks' arrival can be modeled using a counting process, which records the occurrence of attacks over time.

Definition 1 (Counting process) A counting process is a stochastic process $(N_t)_{t\geq 0}$ taking values in $\mathbb{N}_0$ that satisfies $N_0 = 0$, it is almost surely finite and it is a right-continuous step function with increments of size 1. These processes are characterized by the sequence of random arrival times $\{\tau_1, \tau_2, \ldots\}$, where each τ_i marks the occurrence of an arrival.

A counting process $(N_t)_{t\geq 0}$ can be characterized also through its intensity $(\lambda_t)_{t\geq 0}$.

Definition 2 (Intensity) Consider $(N_t)_{t\geq 0}$ a counting process and let $(\mathcal{F}_t)_{t\geq 0}$ denote the history of the arrivals up to time t. Then N satisfies the following relationship

$$\mathbb{P}(N_{t+h} - N_t = m|\mathcal{F}_t) = \begin{cases} \lambda_t h + o(h) & m = 1 \\ o(h) & m > 1 \\ 1 - \lambda_t h + o(h) & m = 0 \end{cases},$$

where λ is called intensity of N.

At time t, the quantity N_t represents number of attacks arrived and λ_t the instantaneous probability of having a new attack at time t.

Poisson processes are a simple example of counting processes, characterized by a constant intensity $\lambda_t \equiv \lambda$. However, Poisson processes assume that events occur independently, which is often an unrealistic assumption in the context of cyber-attacks, where incidents tend to cluster and trigger subsequent events. To capture this characteristic, we introduce Hawkes processes in Section 2.2.

2.2 Hawkes Processes

First introduced by Alan G. Hawkes in [9], are counting processes characterized by their "self-exciting" behavior, meaning that the occurrence of an event increases the likelihood of occurrence of further events.

Definition 3 (Hawkes process) A counting process N is called a Hawkes process if its intensity is given by the stochastic process

$$\lambda_t = \alpha + (\lambda_0 - \alpha)e^{-\xi t} + \beta \int_0^t e^{-\xi(t-s)} \mathrm{d}N_s, \tag{1}$$

with parameters $\alpha \geq 0, \lambda_0, \xi, \beta > 0$.

The parameter $\alpha \geq 0$ can be interpreted as the constant reversion level, $\lambda_0 > 0$ is the initial intensity at time $t = 0$, $\xi > 0$ is the constant rate of exponential decay and $\beta > 0$ is the constant size of self-excited jumps. The process λ can also be expressed in stochastic differential equation (SDE) form, which is obtained applying Itô's Lemma, [10], to Eq. (1):

$$d\lambda_t = -\xi(\lambda_t - \alpha)dt + \beta dN_t, \quad \lambda_0 > 0. \tag{2}$$

Refer to Figure 1(a) for a representation of the Hawkes process N and its associated intensity λ. The self-exciting property of Hawkes processes is well-suited to model the shocks and persistent effects following cyber-attacks, especially considering the tendency of cyber incidents to cluster. This has been confirmed by Baldwin et al. [1], who analyzed threats to major Internet services using data from the SANS Institute. Additional validation comes from Bessy-Roland et al. [3], whose statistical analysis of the Privacy Rights Clearinghouse database highlights how Hawkes models effectively capture self-excitation and interactions in data breaches, proving that Poisson models are not suitable for such events.

2.3 Compound Hawkes processes

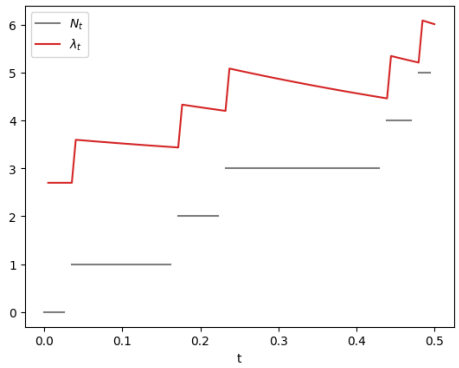
To account for the cumulative damage caused by attacks, it is necessary to introduce compound Hawkes processes, see [15].

Definition 4 (Compound Hawkes process) Let $(N_t)_{t\geq 0}$ be a Hawkes process as defined in Definition 3, we define the compound Hawkes process $(C_t)_{t\geq 0}$ at time t as:

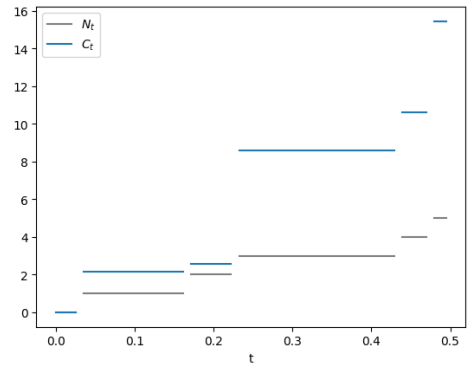
$$C_t = \sum_{i=1}^{N_t} \eta_i,$$

where η_i are independent and identically distributed (i.i.d.) random variables.

The random variable η_i represents the loss associated to the i-th attack and C_t the cumulated losses until time t. Refer to Figure 1(b) for a representation of the compound Hawkes C compared with the Hawkes process N.



(a) Trajectory of an Hawkes process N of the associated intensity λ.



(b) Trajectory of an Hawkes process N and and of the associated compound Hawkes C_t.

Figure 1

3 Optimal Security Investment: A Stochastic Control Approach

In this section, we introduce an optimal security investment problem. We consider an entity subject to cyber-attacks and assume it is sufficiently "large" (e.g., a corporation), leading to a clustered arrival of threats. To reduce its vulnerability, the entity invests in cyber-security. Our goal is to analyze the optimal investment strategy and quantify the benefits of such an investment. In Section 3.1, we briefly present the original Gordon-Loeb model and then introduce our dynamic extension in Section 3.2. In Section 3.3, we discuss the optimization problem, which determines the optimal cyber-security investment.

3.1 The Gordon-Loeb model

The Gordon-Loeb model, first introduced in 2002 by Gordon and Loeb [8], was one of the first studies to address the trade-off between investment and benefits in the context of cyber-security. Gordon and Loeb assume that an information set (e.g., IT system) is characterized by three parameters:

- p: the probability of a threat occurring.
- ℓ: the loss conditioned on an attack occurring.
- v: the vulnerability defined as the probability that a threat once realized (i.e., an attack) would be successful.

The expected loss from an attack if no investment in security is made is $vp\ell$. The entity can invest a certain amount z in security to reduce its vulnerability. This reduction is represented by a security breach probability function $S(z, v)$: after an investment z, a threat will penetrate the entity's IT system with probability $S(z, v)$. After investment, the expected loss is given by $S(z, v)p\ell$. Gordon and Loeb require S to satisfy the following assumptions:

Assumptions (A).

A1. $S(z, 0) = 0$ for all z i.e., an invulnerable information set remains invulnerable.

A2. For all v, $S(0, v) = v$ i.e., if there is no investment in security, then the vulnerability remains unaltered (equal to v).

A3. S is decreasing and convex w.r.t. z, meaning that $S_z(z, v) < 0$ and $S_{zz}(z, v) > 0$, for all $v \in (0, 1)$ and all z.

Remark 1 Gordon and Loeb consider two classes of security breach functions for which satisfy Assumptions (A):

$$S_I(z, v) = \frac{v}{(az + 1)^b} \qquad \text{and} \qquad S_{II}(z, v) = v^{az+1}.$$

for $a, b > 0$.

To find the optimal investment, Gordon and Loeb consider a cost-benefit approach, maximizing the following function:

$$\sup_{z \geq 0}(v - S(z, v))p\ell - z. \tag{3}$$

The first term represents the reduction in the expected loss as a result of the investment z in information security (benefit), while the second term subtracts the cost of investing. The optimal investment is given by z^* which satisfies the following first order condition:

$$-S_z(z^*, v)p\ell - 1 = 0.$$

Gordon and Loeb show that for the two classes of security breach functions in Remark 1, the optimal security investment is always less than $1/e$ times the expected loss,

$$z^* < \frac{1}{e}vp\ell. \tag{4}$$

3.2 Dynamic Extension of the Gordon-Loeb model

The Gordon-Loeb model is interesting due of of its simplicity and its ability to provide a benchmark for the maximum investment in cyber-security, see Eq. (4). However, given the complexity of managing cyber risk (see Introduction), we aim to develop a more sophisticated model that allows for dynamic responses to cyber-risk. Specifically, we extend the

Gordon-Loeb model to a continuous-time and stochastic framework, incorporating the key features outlined in Section 3.1.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $T > 0$ a terminal time. We assume that cyber-attacks arrive according to a Hawkes process $(N_t)_{t\in[0,T]}$, see Definition 3. We denote the jump times of N by $(\tau_i)_{i\in\mathbb{N}}$. For $t \in [0, T]$, the potential losses generated by all cyber-attacks occurring in the time interval $[0, t]$ are given by a compound Hawkes process, see Definition 4:

$$C_t = \sum_{i=1}^{N_t} \eta_i,$$

where $(\eta_i)_{i\geq 1}$ are a family of i.i.d. positive random variables which admit expectation, given by $\mathbb{E}[\eta_i] = \bar{\eta}$ for every i and independent with respect to N. We recall that each random variable η_i represents the loss associated with the i-th attack.

In the spirit of the Gordon-Loeb model, we assume that not all threats are successful: if the entity does not make any investment in security, the attacks penetrate the entity's IT system (or not) depending on its vulnerability $v \in (0, 1)$. The actual losses are given by:

$$L_t^0 = \sum_{i=1}^{N_t} \eta_i \cdot B_i^v,$$

where $(B_i^v)_{i\geq 1}$ is a family of i.i.d. Bernoulli random variables such that $B_i^v \sim \text{Be}(v)$, for every i.

Similarly to the Gordon-Loeb model, the entity can invest in cyber-security to reduce its vulnerability. We assume that, at each instant in time, the entity can invest a certain amount z_t. The process $z = (z_t)_{t\in[0,T]}$ is the investment rate and the cumulated investment at time t is given by $\int_0^T z_t \, \mathrm{d}t$. We assume that the control z belongs to the set $\mathcal{Z}$, described in Definition 5.

Definition 5 The filtration is $\mathbb{F} := (\mathcal{F}_t)_{t\in[0,T]}$, with $\mathcal{F}_t = \sigma(N_s, s \leq t) \vee \sigma(\eta_{\tau_i}, \tau_i \leq t)$. We define by $\mathcal{Z}$ the set of admissible strategies:

$$\begin{aligned} \mathcal{Z} := \{&(z_t)_{t\in[0,T]} \text{ such that } z_t \geq 0, z_t \ \mathbb{F}\text{-predictable and} \\ &\mathbb{E}\left[\int_0^T z_t \mathrm{d}t\right], \mathbb{E}\left[\int_0^T z_t^2 \mathrm{d}t\right] < \infty\}. \end{aligned} \tag{5}$$

Given $t \in [0, T]$, we denote by $\mathcal{Z}_t$ the class $\mathcal{Z}$ restricted to the time interval $[t, T]$.

It is reasonable to assume that a more recent investment in security should be more effective than a past one (e.g. due to the obsolescence of technology). To describe this feature, we introduce a decaying rate $\rho > 0$ and define the process H as follows:

$$H_t = H_0 e^{-\rho t} + \int_0^t e^{-\rho(t-s)} z_s \mathrm{d}s.$$

The stochastic differential equation associated to H follows by applying Itô's Lemma, [10], and its given by

$$\mathrm{d}H_t = (-\rho H_t + z_t)\mathrm{d}t, \quad H_0 \geq 0. \tag{6}$$

Analogously to the Gordon-Loeb case, we introduce a probability breach function $S(H_t, v)$, that satisfies the properties listed in Assumptions (A). Hence, after investing in security, the actual losses of the entity are given by

$$L_t^z = \sum_{i=1}^{N_t} \eta_i \cdot B_i^{S(H_{\tau_i}, v)},$$

where $(B_i^{S(H_{\tau_i},v)})_{i\geq 1}$ is a family of Bernoulli random variables such that $\mathbb{P}(B_i^{S(H_{\tau_i},v)} = 1 | H_{\tau_i} = h) = S(h, v)$.

3.3 The optimization problem

Inspired by the benefit-cost approach presented in Eq. (3), we consider the following problem:

$$\sup_{z\in\mathcal{Z}} \mathbb{E}\left[L_T^0 - L_T^z - \left(\int_0^T \delta z_t + \frac{\gamma}{2} z_t^2 \mathrm{d}t\right) + U(H_T)\right], \tag{7}$$

where λ and H follows the dynamics in Eqs. (2), (6), respectively and the set of admissible controls is $\mathcal{Z}$ as defined in Eq. (5). The difference $L_T^0 - L_T^z$ represents the benefit obtained by the entity's when it invests in cyber-security. Differently from the problem in Eq. (3), where they assume a linear cost of investment, we consider a quadratic cost $\delta z_t + \frac{\gamma}{2} z_t^2$, $\delta > 0, \gamma > 0$. This choice is common in stochastic control literature. We also include a utility function $U(H_T)$, assuming U to be increasing and concave. The function U takes into account the efforts made by the entity before time T. In fact, the entity does not end up existing at time T, thus it needs some security investments for the future.

We can divide all terms in Eq. (7) by δ and compute the expectation of L^0, L^Z, using standard stochastic calculus techniques:

$$\mathbb{E}[L_T^0] = \mathbb{E}\left[\int_0^T v\bar{\eta}\lambda_t \mathrm{d}t\right], \quad \mathbb{E}[L_T^z] = \mathbb{E}\left[\int_0^T S(H_t, v)\bar{\eta}\lambda_t \mathrm{d}t\right].$$

We ultimately derive an equivalent problem, with a slight abuse of notation.

$$\sup_{z\in\mathcal{Z}} \mathbb{E}\left[\int_0^T \left[(v - S(H_t, v))\bar{\eta}\lambda_t - z_t - \frac{\gamma}{2} z_t^2\right] \mathrm{d}t + U(H_T)\right]. \tag{8}$$

From now on, we focus on the problem in Eq. (8). First of all, we introduce the following notation:

- $H_s^{t,h,z}$ is the process H evaluated at time $s > t$, starting at time t, with initial value h and associated to the control z. In particular,

$$H_s^{t,h,z} = h + \int_t^s (-\rho H_v^{t,h,z} + z_v)\mathrm{d}v.$$

- $\lambda_s^{t,\lambda}$ is the process H evaluated at time $s > t$, starting at time t, with initial value λ,

$$\lambda_s^{t,\lambda} = \lambda - \xi \int_t^s (\lambda_v^{t,\lambda} - \alpha)\mathrm{d}v + \beta \int_t^s \mathrm{d}N_v^{\lambda}.$$

- J is the revenue function, i.e. the function we aim at maximizing given the initial state (t, λ, h):

$$J(t,\lambda,h;z) = \mathbb{E}\left[\int_t^T \left[(v - S(H_s^{t,h,z}, v))\bar{\eta}\lambda_s^{t,\lambda} - z_s - \frac{\gamma}{2}z_s^2\right]\mathrm{d}s + U(H_T^{t,h,z})\right].$$

Consequently, we define the value function as

$$V(t,\lambda,h) = \sup_{z\in\mathcal{Z}_t} J(t,\lambda,h;z), \tag{9}$$

where $\mathcal{Z}_t$ has been defined in Definition 5. As a further assumption, we require that the function V as defined in Eq. (9) is at least $\mathcal{C}^1$, i.e., continuous and differentiable with continuous derivative in all its arguments.

Theorem 1 *V solves the following partial integral differential equation*

$$\begin{aligned} &\frac{\partial V}{\partial t} - \xi(\lambda-\alpha)\frac{\partial V}{\partial \lambda} - \rho h \frac{\partial V}{\partial h} + \lambda(V(t,\lambda+\beta,h) - V(t,\lambda,h)) \\ &\quad + (v - S(h,v))\bar{\eta}\lambda + \frac{\left(\frac{\partial V}{\partial h} - 1\right)^+}{\gamma}\left(\frac{\partial V}{\partial h} - 1 - \frac{\gamma}{2}\frac{\left(\frac{\partial V}{\partial h} - 1\right)^+}{\gamma}\right) = 0, \\ &\quad V(T,\lambda,h) = U(h). \end{aligned} \tag{10}$$

Moreover, the optimal control is given by

$$z^* = \frac{\left(\frac{\partial V}{\partial h} - 1\right)^+}{\gamma}. \tag{11}$$

Proof. Using similar techniques to [4, Theorem 2.2.2.], [2, Section 5.2, Eqs. (38a), (38b)], we can prove that the value function $V(t,\lambda,h)$ solves the Partial-Integro Differential-Equation in Eq. (11). □

Remark 2 The interpretation of the optimal control in Eq. (11) is the following: it is worth to invest if the benefit we obtain by doing so is larger than the marginal cost.

4 Numerical Results

Due to the strong non-linearity, the partial differential equation (PIDE) in (10) cannot be solved analytically and must be approached numerically. We solve it exploiting *method of lines*, i.e. we discretize the spatial derivatives (derivatives w.r.t. λ and h), transforming the

PIDE into a system of ordinary differential equations (ODEs), which can then be solved using standard solver for ODEs, e.g. scipy.integrate.solve_ivp, (`https://docs.scipy.org/doc/scipy/reference/generated/scipy.integrate.solve_ivp.html`). We do not provide a detailed description of the numerical method employed, refer to [16] for further details. The parameters chosen to perform the numerical analysis are in Tables 1, 2, 3.

S	v	a	b
S_I	0.65	$1 \cdot 10^{-5}$	1

Table 1: Security breach function.

λ	α	ξ	β	λ_0
	2.7	1.5	0.9	2.7

Table 2: Hawkes intensity.

Optimization	δ	γ	$\bar{\eta}$(k$)	$U(h)$	ρ	T
	0.01	0.2	2	$0.02\sqrt{h}$	0.03	0.5

Table 3: Optimization problem parameters.

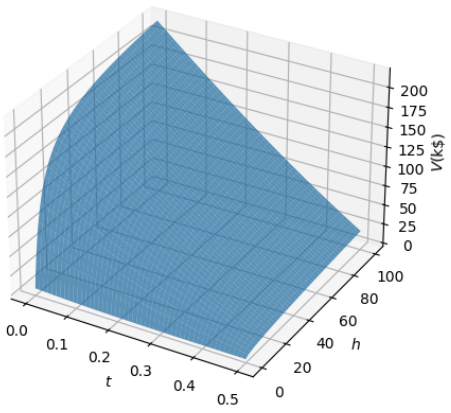
The parameters for the security breach function, Table 1, are analogous to those in [14], which are themselves slight variations of those in [8] and [13]. We choose the security breach function S as S_I in Remark 1, considering the h entry to be expressed in k$. Under this assumption, S_I becomes:

$$S_I(h, v) = \frac{v}{(az \cdot 10^3 + 1)^b}.$$

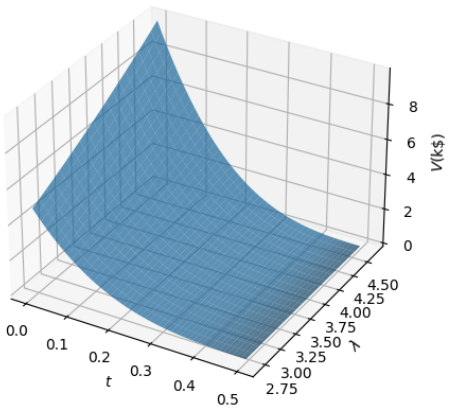
For the Hawkes intensity parameters, see Table 2, we choose those in [5]. In [5], the authors calibrate the Hawkes intensity parameters on real data, taken from the Hackmageddon database. We refer to [5, Section 4.1.1] for further details on the database. For the other parameters, we refer to Table 3.

Value function and optimal control We report in Figure 2 different representations for the value function and the optimal control. In Subfigures 2(a), 2(c) we plot the functions for h fixed, varying t and λ. We observe that the value function is increasing in h. Clearly, a higher cumulative initial investment h leads to a greater benefit, which is represented by a higher value function. On the other hand, the optimal control decreases in h as more money the entity has already invested, the less it should invest later. Both the value function and optimal control are decreasing functions of t. In fact, approaching maturity, the impact of the entity's actions becomes less significant in generating substantial benefits, and thus this causes an overall lower investment. In Subfigures 2(b), 2(d) we plot the functions for

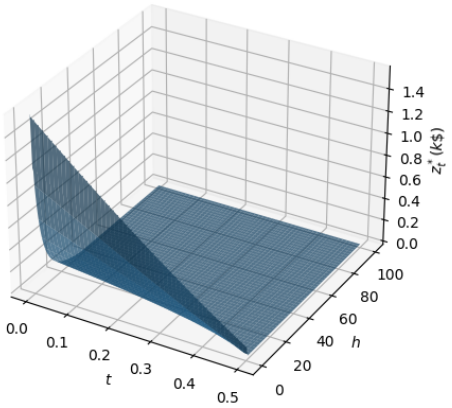
λ fixed, varying t and h. We highlight that the value function is increasing in λ. The same holds for the optimal control. A larger λ represents a higher risk that induces a higher benefit, if the entity invests wisely. In terms of optimal control, a larger risk should lead to a higher investment to mitigate it.



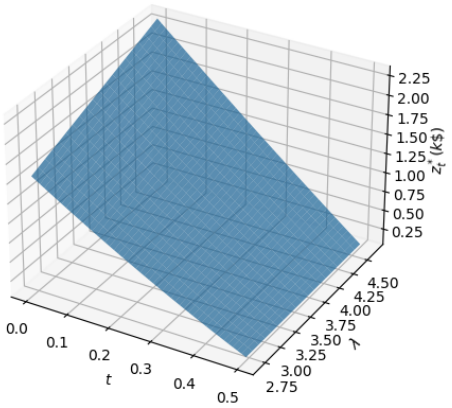
(a) Value function $V(t, \lambda, h)$ for $\lambda = 2.7$.



(b) Value function $V(t, \lambda, h)$ for $h = 0$.



(c) Optimal control $z_t^*(\lambda, h)$ for $\lambda = 2.7$.



(d) Optimal control $z_t^*(\lambda, h)$ for $h = 0$.

Figure 2: Value function and optimal control computed with *standard parameters set*, see Tables 1, 2, 3.

Comparison with Poisson model for arrival of attacks. In this paragraph, we reformulate the optimization problem by choosing as the counting process a Poisson process. We denote by P a Poisson process, i.e., a counting process having constant intensity λ^P. The optimization problem we aim at solving is the same of Eq. (8), but considering a constant intensity rather than a dynamic one. Under the Poisson's hypothesis, the problem becomes deterministic, as the only randomness lies in the Hawkes' dynamics. We denote by $V^P(t,h)$ the value function in this setting, which now solves the PDE:

$$\frac{\partial V^P}{\partial t} - \rho h \frac{\partial V^P}{\partial h} + \lambda^P (v - S(h,v))\bar{\eta} + \frac{\left(\frac{\partial V^P}{\partial h} - 1\right)^+}{\gamma} \left(\frac{\partial V^P}{\partial h} - 1 - \frac{\gamma}{2} \frac{\left(\frac{\partial V^P}{\partial h} - 1\right)^+}{\gamma} \right) = 0,$$

$$V^P(T,h) = U(h).$$

The corresponding optimal control is given by

$$z^{P*} = \frac{\left(\frac{\partial V^P}{\partial h} - 1\right)^+}{\gamma}.$$

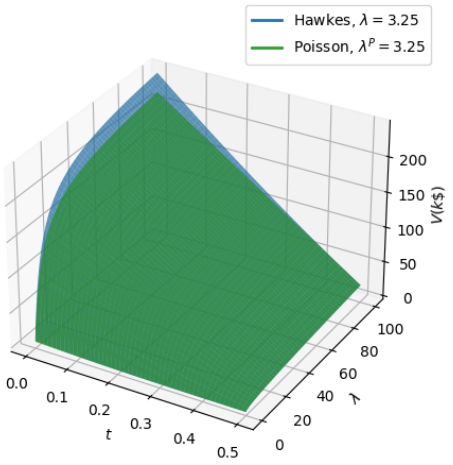
We consider a Poisson P and a Hawkes process N such that $\mathbb{E}[P_T] = \mathbb{E}[N_T]$. Recall that $\mathbb{E}[P_T] = T\lambda^P$ if the Poisson process has intensity λ^P, thus it follows that

$$\lambda^P = \frac{\lambda_0 \xi}{\xi - \beta} + \frac{1}{T(\xi - \beta)} (\lambda_0 - \frac{\lambda_0 \xi}{\xi - \beta})(1 - e^{-\xi T}) = 3.25,$$

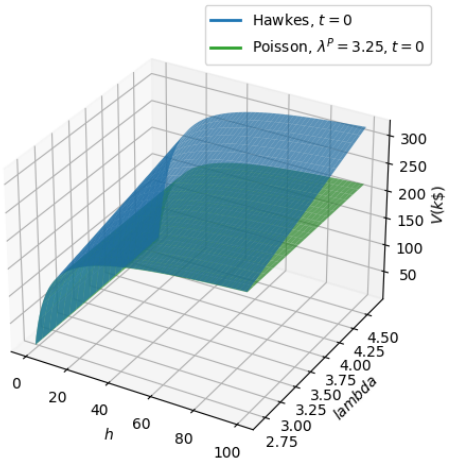
with the parameters choice in Tables 2, 3. Refer to [6, Theorem 3.6, Eq. (3.16)] for the last formula.

In Subfigure 3(a), we observe the Hawkes value function $V(t,\lambda,h)$ evaluated in λ^P, compared with the Poisson value function $V^P(t,h)$ and in Subfigure 3(c) we depict the corresponding optimal controls. In Subfigures 3(b) and 3(d), we compare the value function and optimal control, varying λ and h at time $t = 0$. Since $V^P(0,h)$ does not depend on λ, so we assume it to be constant for every λ. We observe that considering Hawkes processes instead of Poisson's causes a larger value function and optimal control.

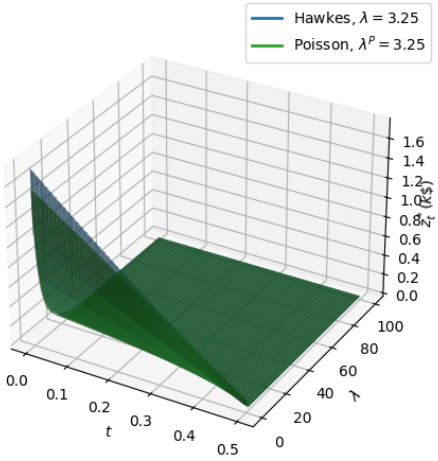
To further enhance this comparison, we also consider the behaviour of the optimal control along a specific trajectory of λ_t. In Figure 4, we observe that around the Hawkes' jump times (grey dotted lines), we tipically observe a jump also along the optimal control. We note that the optimal control for Hawkes is smaller than the optimal control for the Poisson process before the first jump. Then, when $\lambda_t(\omega) > \lambda^P$ it becomes larger.



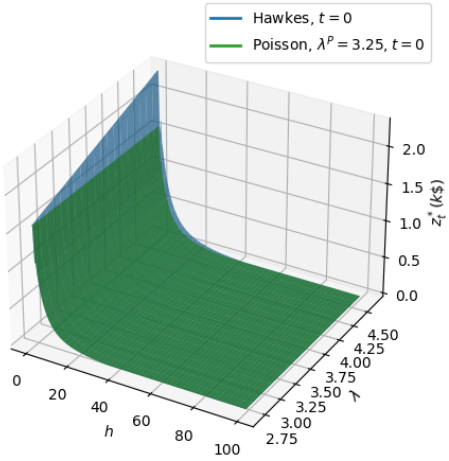
(a) Value function $V(t, \lambda, h)$ for $\lambda = 3.25$.



(b) Value function $V(t, \lambda, h)$ for $h = 0$.

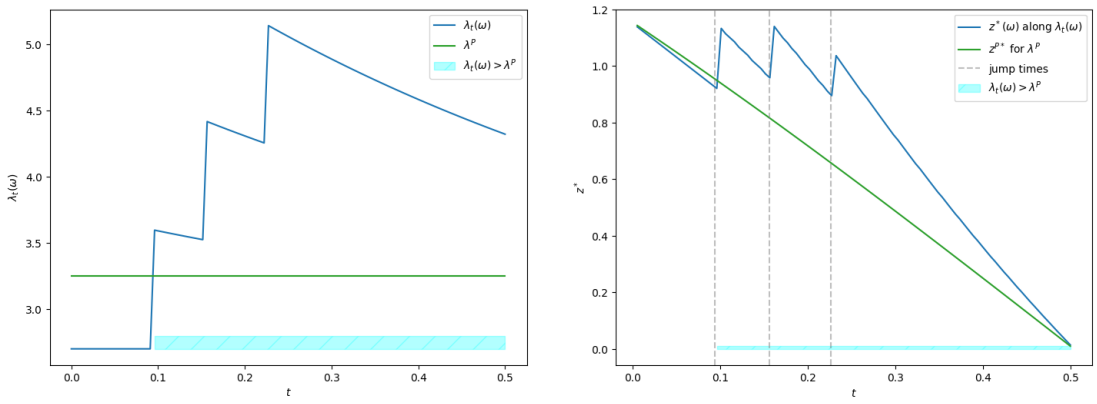


(c) Optimal control $z_t^*(\lambda, h)$ for $\lambda = 3.25$.



(d) Optimal control $z_t^*(\lambda, h)$ for $h = 0$.

Figure 3: Comparison between Hawkes and Poisson, $\lambda^P = 3.25$.



(a) Intensity trajectory.

(b) Optimal control along the trajectory.

Figure 4: Optimal control along a trajectory.

5 Conclusions

In this paper, we provide a dynamic version of the Gordon-Loeb model, exploiting instruments such as compound counting processes to describe the overall losses experienced by the entity and Hawkes processes to represent the cyber-attack arrivals. We then formulate an optimization problem which respect the cost-benefit tradeoff proposed in the original Gordon-Loeb setting and solve it with dynamic programming techniques. We characterize the solution via a partial-integro-differential equation that we solve numerically. We then perform some numerical tests, to study the main properties of the optimal investment strategy and to compare our Hawkes setting with a Poisson one. We realize that not considering a dynamic intensity instead of a constant one might leads to a sub-optimal investment rate. As a next step we aim at investigation of the value function properties, in particular the existence of a solution (in a suitable sense) to the PIDE and the verification theorem. We would also like to explore a scenario in which the entity may also enter into an insurance contract to cover losses from cyber-attacks. We may then study the optimal allocation of the entity's resources in security and insurance. Finally, a possible extension might regard taking into consideration a singular control problem.

References

[1] Adrian Baldwin, Iffat Gheyas, Christos Ioannidis, David Pym, and Julian Williams, *Contagion in cyber security attacks.* Journal of the Operational Research Society 68 (2017), no. 7, 780–791.

[2] Alain Bensoussan and Benoit Chevalier-Roignant, *Stochastic control for diffusions with self-exciting jumps: An overview.* Mathematical Control and Related Fields 14 (2024), no. 4, 1452–1476.

[3] Yannick Bessy-Roland, Alexandre Boumezoued, and Caroline Hillairet, *Multivariate Hawkes process for cyber insurance.* Annals of Actuarial Science 15 (2021), no. 1, 14–39.

[4] Bruno Bouchard, "Introduction to stochastic control of mixed diffusion processes, viscosity solutions, and applications in finance and insurance". Lecture Notes, 2007.

[5] Alexandre Boumezoued, Yousra Cherkaoui, and Caroline Hillairet, *Cyber risk modeling using a two-phase Hawkes process with external excitation.* ArXiv preprint arXiv:2311.15701 (2023).

[6] Angelos Dassios and Hongbiao Zhao, *A dynamic contagion process.* Advances in Applied Probability 43 (2011), no. 3, 814–846.

[7] Martin Eling and Jan Hendrik Wirfs, *Cyber risk: too big to insure? Risk transfer options for a mercurial risk class.* I.VW HSG Schriftenreihe 59 (2016).

[8] Lawrence A. Gordon and Martin P. Loeb, *The economics of information security investment.* ACM Transactions on Information and System Security (TISSEC) 5 (2002), no. 4, 438–457.

[9] Alan G. Hawkes, *Spectra of some self-exciting and mutually exciting point processes.* Biometrika 58 (1971), no. 1, 83–90.

[10] Kiyosi Itô, *On a formula concerning stochastic differentials.* Nagoya Mathematical Journal 3 (1951), 55–65.

[11] Patrick J. Laub, Young Lee, and Thomas Taimre, "The elements of Hawkes processes". Springer, 2021.

[12] Angelica Marotta, Fabio Martinelli, Stefano Nanni, Albina Orlando, and Artsiom Yautsiukhin, *Cyber-insurance survey.* Computer Science Review 24 (2017), 35–61.

[13] Alessandro Mazzoccoli and Maurizio Naldi, *Robustness of Optimal Investment Decisions in Mixed Insurance/Investment Cyber Risk Management.* Risk analysis 40 (2020), no. 3, 550–564.

[14] Henry R.K. Skeoch, *Expanding the Gordon-Loeb model to cyber-insurance.* Computers & Security 112 (2022), 102533.

[15] Gabriele Stabile and Giovanni Luca Torrisi, *Risk Processes with Non-stationary Hawkes Claims Arrivals.* Methodology and Computing in Applied Probability 12 (2010), 415—429.

[16] Si Yuan, *ODE-oriented semi-analytical methods.* Computational Mechanics in Structural Engineering (1999), 375–388.

[17] Gabriela Zeller and Matthias Scherer, *A comprehensive model for cyber risk based on marked point processes and its application to insurance.* European Actuarial Journal 12 (2022), no. 1, 33–85.

Bridging Enumerative Geometry and Quantum Integrable Hierarchies

Ishan Jaztar Singh (*)

Abstract. Enumerative geometry applies combinatorial and intersection theory techniques to solve counting problems in algebraic geometry. In contrast, integrable hierarchies are infinite sequences of partial differential equations with rich symmetries, playing a crucial role in mathematical physics. This note focuses on the renowned Witten-Kontsevich theorem, which establishes a deep connection between the topological invariants of the moduli space of curves and the Korteweg–de Vries hierarchy. We conclude with a discussion of its generalizations and the role of quantum hierarchies in this context.

1 Introduction

The moduli space M_g of algebraic curves of genus g classifies algebraic curves of genus g up to isomorphism. Its Deligne-Mumford compactification, $\overline{M}_g$, extends this classification to stable algebraic curves of genus g. To address more general enumerative geometry questions, one considers the moduli space $\overline{M}_{g,n}$ of stable algebraic curves of genus g with n marked points. This space plays a fundamental role in algebraic geometry, and its cohomology ring, $H^\bullet(\overline{M}_{g,n})$, encodes the necessary information for formulating intersection theory on $\overline{M}_{g,n}$.

Since the full structure of the cohomology ring is often too intricate to analyze, attention is typically restricted to the tautological ring $RH^\bullet(\overline{M}_{g,n})$. This is the smallest subring of $H^\bullet(\overline{M}_{g,n})$ generated by natural geometric constructions on $\overline{M}_{g,n}$. A key feature of the tautological ring is the presence of ψ-classes, which can be expressed as first Chern classes of certain natural line bundles on $\overline{M}_{g,n}$. These ψ-classes, along with their associated strata, generate the tautological ring. However, determining the complete set of relations among these generators remains an open problem.

In this note, we provide an informal overview of $\overline{M}_{g,n}$, focusing on essential definitions and key results. For a more comprehensive treatment, we refer to [Zvo14; Sch20]. Additionally, an accessible and detailed review of the genus-zero theory of the moduli space of curves can be found in [KV07].

(*) Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: jaztar@math.unipd.it . Seminar held on 21 November 2024.

2 Stable Curves and Graphs

The algebraic curves classified by $\overline{M}_{g,n}$ are connected, nodal, complex-projective curves of genus g, equipped with n marked points. From an analytic perspective, $\overline{M}_{g,n}$ can be viewed as the space classifying nodal genus g compact Riemann surfaces with n punctured points on their surface. A Riemann surface with n marked points is denoted as

$$(C, p_1, \ldots, p_n),$$

where $p_i \in C$ and all the marked points are distinct, i.e., $p_i \neq p_j$ for $i \neq j$. A curve $(C, p_1, \ldots, p_n)$ is said to be *stable* if the set of its automorphisms, $\mathrm{Aut}(C, p_1, \ldots, p_n)$, is finite. To determine whether a genus-g curve with n marked points is stable, we use the *stability condition*:

$$2g - 2 + n > 0. \tag{2.1}$$

Thus, curves corresponding to the pairs $(g, n) \in \{(0,0), (0,1), (0,2), (1,0)\}$ are not stable and are not classified. Each point of $\overline{M}_{g,n}$ represents an isomorphism class $[(C, p_1, \ldots, p_n)]$.

To define integrals, we require a compact space. The Deligne-Mumford compactification, $M_{g,n} \mapsto \overline{M}_{g,n}$, introduces blow-ups, resulting in nodal curves. These nodal curves, which consist of multiple components, appear on the boundary of $\overline{M}_{g,n}$, while $M_{g,n}$ forms a dense open subset in its interior. For an intuitive visualization of the points in $\overline{M}_{g,n}$, see the diagrams in Figure 1.

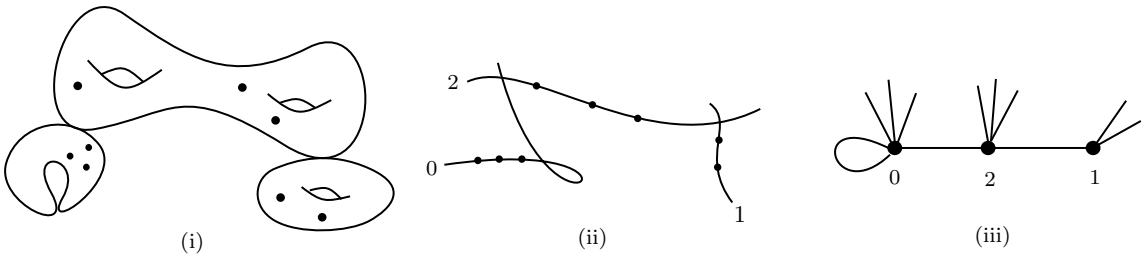


Figure 1: Equivalent descriptions of an isomorphism class in $\overline{M}_{4,8}$

In Figure 1(i), we present a topological perspective of *transversal* intersections in a nodal curve. This curve lies on the boundary of $\overline{M}_{4,8}$ as a result of the compactification. In Figure 1(ii), we provide a geometric perspective of a *reducible* nodal curve, where the transversal intersections are more clearly visible compared to (i). Additionally, we indicate the genus of each component on the side, indexed as $(0, 1, 2)$ in this case. In both Figures 1(i) and (ii), the curve is decomposed into three *irreducible* components, each belonging to a moduli space of lower genus with fewer marked points. Consequently, a curve $(C, p_1, \ldots, p_n)$ is considered stable if each of its irreducible components, $\tilde{C}$, satisfies the stability condition in Equation (2.1). More explicitly, each component $\tilde{C}$ must satisfy one of the following conditions:

- $\tilde{C}$ has genus zero and at least three marked points.

- $\tilde{C}$ has genus one and at least one marked point.
- $\tilde{C}$ has genus at least two.

For a detailed treatment of blow-ups in the genus-zero case, the interested reader may refer to [KV07], where the moduli space is expressed in terms of complex projective space:

$$\overline{M}_{0,n} = \overbrace{\mathbb{P} \times \ldots \times \mathbb{P}}^{n-3}.$$

In the algebraic geometric setting, $\overline{M}_{g,n}$ is defined as a smooth Deligne–Mumford stack, whereas in the differential geometric setting, $\overline{M}_{g,n}$ is considered a smooth complex orbifold. For a basic overview of orbifolds, see [Zvo14], and for more detailed insights, refer to [ALR07].

By treating $\overline{M}_{g,n}$ as an orbifold with a group action given by the automorphism group of the curves, its cohomology and homology can be defined analogously to those of manifolds. However, there are subtleties due to the presence of the group action. Additionally, the (complex) dimension of $\overline{M}_{g,n}$ is well-defined and corresponds to the dimension of its associated complex orbifold:

$$\dim_{\mathbb{C}} \overline{M}_{g,n} = 3g - 3 + n.$$

A nontrivial one-dimensional example is the moduli space of elliptic curves, $\overline{M}_{1,1}$, which is constructed as $\mathbb{C}_+/\mathrm{PSL}(2,\mathbb{Z})$, where $\mathrm{PSL}(2,\mathbb{Z})$ is the modular group (see [Eyn18, p. 84]).

For practical purposes, points in $\overline{M}_{g,n}$ can be interpreted in terms of graphs that are dual to their corresponding curves. An example of such a graph is shown in Figure 1(iii). The vertices of the graph encode the geometric genus of the curve, while the legs (or external edges) attached to each vertex represent the marked points. The edges between vertices correspond to the gluing or intersection of different components of the curve.

Definition 2.2 A *stable graph* Γ is a tuple,

$$\Gamma = \Big(\mathrm{V}(\Gamma), \mathrm{H}(\Gamma), \mathrm{L}(\Gamma), g : \mathrm{V} \to \mathbb{Z}_{\geq 0} \;\Big|\; v : \mathrm{H} \to \mathrm{V},\ \iota : \mathrm{H} \to \mathrm{H},\ l : \mathrm{L} \to \mathrm{N}\Big)$$

where $\mathrm{N} = \{1, \ldots, n\}$ and the tuple satisfies the following:

i. $\mathrm{V}(\Gamma)$ is a finite set of vertices v, while g is a map $v \mapsto g(v)$ that provides the geometric genus of the vertex.

ii. $\mathrm{H}(\Gamma)$ is a finite set of half-edges h, while v is a map $h \mapsto v(h)$ that sends half-edges to the vertex its attached to.

iii. $\mathrm{E}(\Gamma)$ is a finite set of edges $e = (h, h')$, that contains pairs of half-edges, such that the involution $\iota : h \mapsto h'$ is an involution.

iv. $\mathrm{L}(\Gamma) \subset H(\Gamma)$ is the set of half-edges that are fixed by ι, i.e. for any $h \in \mathrm{L}(\Gamma)$, $\iota(h) = h$. The map l is a bijective map that sends half-edges to the index of the marked points $\{1, \ldots, n\}$.

v. The graph Γ is connected and the data surrounding each vertex $v \in \mathrm{V}(\Gamma)$ satisfies stability condition:

$$2g(v) - 2 + n(v) > 0$$

Let $n(v)$ denote the number of half-edges around a vertex $v \in \mathrm{V}(\Gamma)$, so that the number of marked points $n(\Gamma)$ of the graph is given by,

$$n(\Gamma) = \sum_{v \in \mathrm{V}(\Gamma)} n(v) = |\mathrm{L}(\Gamma)|$$

Since self-intersection of a curve is included in $\overline{M}_{g,n}$, we did not restrict loops in out definition, loops being an edge attached to the same vertex. Hence, the genus $g(\Gamma)$ of the graph is given by,

$$g(\Gamma) = \sum_{v \in \mathrm{V}(\Gamma)} g(v) + 1 + |\mathrm{E}(\Gamma)| - |\mathrm{V}(\Gamma)|$$

Example 2.3 Consider the diagrams in Figure 1, its stable graph is given by the indexed figure 2. We then have the following data:

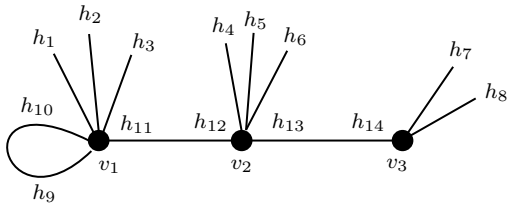


Figure 2: An example of a stable graph in $\overline{M}_{4,8}$.

- $\mathrm{V}(\Gamma) = \{v_1, v_2, v_3\}$, $\mathrm{H}(\Gamma) = \{h_1, \ldots, h_{14}\}$, $\mathrm{E}(\Gamma) = \{(h_{11}, h_{12}), (h_{13}, h_{14})\}$ and $\mathrm{L}(\Gamma) = \{h_1, \ldots, h_8\}$.
- The map g is defined by, $g(v_1) = 0, g(v_2) = 2$ and $g(v_3) = 1$. The map v is defined by, $v(\{h_1, h_2, h_3, h_9, h_{10}, h_{11}\}) = v_1$ and likewise for the other vertices. The map l is defined by $l(h_i) = i$ for $i = 1, \ldots, 8$.
- Hence, the number of marked points is $n(\Gamma) = 8$ and $g(\Gamma) = 4$.

Denote by $\mathrm{G}_{g,n}$ the set of isomorphism classes of stable graphs with $n(\Gamma) = n$ and $g(\Gamma) = g$. For any $\Gamma \in \mathrm{G}_{g,n}$, we define a moduli space M_Γ that classifies curves C whose associated stable graphs Γ_C are isomorphic to Γ. In fact, M_Γ is a closed subspace of $\overline{M}_{g,n}$, often referred to as a *stratum* of the moduli space of stable curves. Informally, it is plausible to observe that $\overline{M}_{g,n}$ is a disjoint union of all possible configurations of M_Γ:

$$\overline{M}_{g,n} = \coprod_{\Gamma \in \mathrm{G}_{g,n}} M_\Gamma.$$

Additionally, the dimension of each stratum is given by:

$$\dim_{\mathbb{C}} M_\Gamma = \sum_{v \in \mathrm{V}(\Gamma)} (3g(v) - 3 + n(v)) = \dim_{\mathbb{C}} \overline{M}_{g,n} - |\mathrm{E}(\Gamma)|.$$

This formula can be easily verified in the genus-zero case (see [KV07, p. 35]).

Subspaces M_Γ of $\overline{M}_{g,n}$ with $|\mathrm{E}(\Gamma)| > 0$ are often referred to as *boundary cycles* of $\overline{M}_{g,n}$, while those with $|\mathrm{E}(\Gamma)| = 1$ are called *boundary divisors*. Moreover, it can be shown that the boundary of $\overline{M}_{g,n}$ is the disjoint union of all its boundary divisors, i.e.,

$$\partial \overline{M}_{g,n} = \overline{M}_{g,n} \backslash M_{g,n} = \coprod_{\Gamma \in \mathrm{G}_{g,n}, |\mathrm{E}(\Gamma)|=1} M_\Gamma.$$

For an illustration of these statements, see [Sch20, p. 28].

3 Tautological Classes

A global understanding of the moduli space of stable curves $\overline{M}_{g,n}$ requires its topological data, particularly its homology and cohomology rings. We follow the standard approach by considering the singular cohomology ring $H^\bullet(\overline{M}_{g,n}, \mathbb{Q})$ and homology ring $H_\bullet(\overline{M}_{g,n}, \mathbb{Q})$ with rational coefficients. The singular cohomology ring maps to the Chow ring $A^\bullet(\overline{M}_{g,n})$, which is central in intersection theory within algebraic geometry. Whenever it is clear, we omit $\mathbb{Q}$ and simply write $H^\bullet(\overline{M}_{g,n})$. Only basic intersection theory is needed, as covered in [Nic11], with more detailed treatments in [HM91; Gat00].

Let X be a d-dimensional smooth connected complex projective variety. Denote by $\smile$ and $\frown$ the cup and cap products on $H^\bullet(X)$, given by

$$\begin{aligned} \smile &: H^k(X) \otimes H^l(X) \to H^{k+l}(X), \quad \alpha \otimes \beta \mapsto \alpha \smile \beta, \\ \frown &: H_k(X) \otimes H^l(X) \to H_{k-l}(X), \quad \alpha \otimes \beta \mapsto \alpha \frown \beta, \end{aligned}$$

satisfying the relation $\alpha \frown (\beta \smile \gamma) = (\alpha \frown \beta) \smile \gamma$. From the cap product and the identification $H_0(X) \simeq \mathbb{Q}$, we obtain the duality $H^k(X) \simeq H_k(X)^*$, where $H_k(X)^*$ is the dual space of $H_k(X)$.

The degree map is defined as

$$\deg : H^{2d}(X) \xrightarrow{\sim} \mathbb{Q}, \quad \deg(\alpha) = \int_X \alpha.$$

Since $\dim_{\mathbb{R}}(X) = 2d$, Poincaré duality provides the non-degenerate pairing

$$H^k(X) \otimes H^{2d-k}(X) \to \mathbb{Q}, \quad \alpha \otimes \beta \mapsto \deg(\alpha \smile \beta).$$

This implies the isomorphisms

$$H^k(X) \simeq H^{2d-k}(X)^* \simeq H_{2d-k}(X),$$

so that the map $H^k(X) \xrightarrow{\sim} H_{2d-k}(X)$ is given by $\alpha \mapsto [X] \frown \alpha$, where $[X]$ is the fundamental class of X (the Poincaré dual of X).

From a differential topology perspective, given a d-dimensional closed and oriented manifold X and a k-dimensional submanifold $S \subset X$, we associate S to its Poincaré dual $[S] \in H^{2(d-k)}(X)$ via

$$\int_S \iota^* \alpha = \int_X \alpha \smile [S],$$

where $\alpha \in H_c^{2k}(X)$ is a compactly supported k-form, and $\iota : S \hookrightarrow X$ is the inclusion map.

Given a morphism $f : X \to Y$, where $\dim X = d$ and $\dim Y = e$, the pushforward f_* is defined for the homology ring, while the pullback f^* is defined for the cohomology ring. Using Poincaré duality, where $H^l(X) \simeq H_{2d-l}(X)$ and $H_{2d-l}(Y) \simeq H^{2(e-d)+l}(Y)$, we obtain a well-defined pushforward in cohomology:

$$f_* : H^l(X) \to H^{2(e-d)+l}(Y).$$

The pushforward and pullback satisfy the following compatibility relations with respect to the cup and cap products:

$$f^*(\alpha \smile \beta) = f^*\alpha \smile f^*\beta,$$
$$f_*(f^*\beta \smile \alpha) = \beta \smile f_*\alpha.$$

We now extend these notions to the cohomology ring $H^\bullet(\overline{M}_{g,n})$ of the moduli space of stable curves. The *i-th cotangent line bundle* $\mathcal{L}_i$ is defined as

$$\mathcal{L}_i|_{(C,p_1,\dots,p_n)} = T^*_{p_i} C.$$

The *i-th ψ-class* is then given by the first Chern class of $\mathcal{L}_i$:

$$\psi_i = c_1(\mathcal{L}_i) \in H^2(\overline{M}_{g,n}).$$

A natural morphism called the *forgetful map*, denoted by π, is defined as

$$\pi : \overline{M}_{g,n+1} \to \overline{M}_{g,n},$$

which maps a tuple $(C, p_1, \dots, p_{n+1})$ with $(n+1)$ marked points to a tuple $(C', p'_1, \dots, p'_n)$ with n marked points. Two cases arise when removing a marked point: If the resulting curve remains stable, then $C' = C$ and $p'_i = p_i$. If the resulting curve is unstable, a contraction map $\phi : C \to C'$ collapses unstable irreducible components to a marked point $p'_i = \phi(p_i)$. These conditions ensure that π is well defined.

A nontrivial fact is that the ψ-classes are tautological for every $i = 1, \dots, n$, as they can be expressed as

$$\psi_i = -\pi_*([\delta_i] \smile [\delta_i]),$$

where $[\delta_i]$ is the Poincaré dual of the boundary divisor δ_i, illustrated in Figure 3.

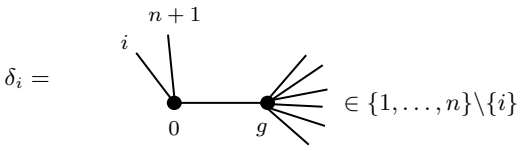


Figure 3: The graphical representation of the boundary divisor δ_i.

The *pullback relation* for the ψ-classes are given by

$$\pi^*\psi_i = \psi_i - [\delta_i].$$

This fundamental relation play a crucial role in computations.

Example 3.1 In genus zero, the i-th ψ-class admits an explicit expression:

$$\psi_i = [\delta_{i,jk}],$$

where $[\delta_{i,jk}]$ is the Poincaré dual of the boundary divisor $\delta_{i,jk}$. This divisor corresponds to a stable graph where the i-th marked point lies on one component, while the j-th and k-th marked points lie on the other, separated by an edge (see [Zvo14, p. 26]).

4 Witten-Kontsevich Theorem

Most of this section summarizes key results from Witten's groundbreaking article [Wit91], which connects intersection invariants of $\overline{M}_{g,n}$ to differential equations and the Korteweg–de Vries (KdV) hierarchy. We first define the basic objects and notation.

The *Witten-Kontsevich (WK) descendant correlators* are given by

$$\langle \tau_{i_1} \dots \tau_{i_n} \rangle_{g,n} = \int_{\overline{M}_{g,n}} \psi_1^{i_1} \dots \psi_n^{i_n},$$

where $\psi_k \in H^2(\overline{M}_{g,n})$. Let $(t_i : 0 \le i < \infty)$ be an ordered sequence of variables. The *genus-g Witten-Kontsevich potential* F_g is expressed as

$$F_g = \sum_{n, i_\bullet \ge 0} \langle \tau_{i_1} \dots \tau_{i_n} \rangle_{g,n} \frac{t_{i_1} \dots t_{i_n}}{n!}.$$

The *Witten-Kontsevich descendant potential* is given by

$$\mathcal{Z}^{\mathrm{WK}} = \sum_{g=0}^{\infty} \hbar^{2g} F_g.$$

Witten introduced two fundamental equations, later used to compute intersection invariants in genus zero and one. The *string equation* is

$$\frac{\partial F}{\partial t_0} = \frac{t_0^2}{2} + \sum_{k \ge 0} t_{k+1} \frac{\partial F}{\partial t_1},$$

while the *dilaton equation* is

$$\frac{\partial F}{\partial t_1} = \frac{1}{24} + \frac{1}{3}\sum_{k\geq 0}(2k+1)t_k\frac{\partial F}{\partial t_k}.$$

Setting $F = F_g$, these equations link differential equations to intersection theory.

Proposition 4.1 [Wit91] *The string equation is equivalent to*

$$\langle \tau_0\tau_{i_1}\dots\tau_{i_n}\rangle_{g,n+1} = \sum_{j=1}^{n}\langle \tau_{i_1}\dots\tau_{i_j-1}\dots\tau_{i_n}\rangle_{g,n},$$

which holds in $\overline{M}_{g,n}$ *for* $2g-2+n>0$.

Using the string equation, all genus zero ψ-class integrals can be computed by induction.

Corollary 4.2 *The genus-zero correlators satisfy*

$$\langle \tau_{i_1}\dots\tau_{i_n}\rangle_{0,n} = \frac{(n-3)!}{i_1!\dots i_n!}.$$

Proposition 4.3 [Wit91] *The dilaton equation is equivalent to*

$$\langle \tau_1\tau_{i_1}\dots\tau_{i_n}\rangle_{g,n+1} = (2g-2+n)\langle \tau_{i_1}\dots\tau_{i_n}\rangle_{g,n},$$

which holds in $\overline{M}_{g,n}$ *for* $2g-2+n>0$.

These relations are fundamental in the theory of the moduli space of stable curves and integrability. However, Witten's main conjecture is significantly more intricate. Although it has undergone substantial development, we will not explore it in detail. Kontsevich [Kon92] provided the first proof, with five different proofs now known.

Define $u(t_0,t_1,\dots)$ as a smooth function in the variables $(t_i : 0\leq i<\infty)$, and let $p_i(u,\dot{u},\ddot{u},\dots)$ be polynomials in the derivatives

$$\dot{u} = \frac{\partial u}{\partial t_0}, \quad \ddot{u} = \frac{\partial^2 u}{\partial t_0^2}, \quad \dots.$$

The *Korteweg–de Vries (KdV) hierarchy* is given by

$$\frac{\partial u}{\partial t_i} = \frac{\partial p_{i+1}}{\partial t_0},$$

where $p_1 = u$ and the higher-order terms satisfy the recursion

$$\dot{p}_{i+1} = \frac{1}{2i+1}\left(p_i\dot{u} + 2\dot{p}_i u + \frac{1}{4}\dddot{p}_i\right).$$

The first equation in this hierarchy, the *KdV equation*, is

$$\begin{aligned}\frac{\partial u}{\partial t_1} &= u\frac{\partial u}{\partial t_0} + \frac{1}{12}\frac{\partial^3 u}{\partial t_0^3} \\ &= \frac{\partial}{\partial t_0}\left(\frac{\partial u^2}{2} + \frac{1}{12}\frac{\partial^2 u}{\partial t_0^2}\right).\end{aligned} \tag{4.4}$$

The hierarchy consists of an infinite system of PDEs obtained recursively, such as

$$\frac{\partial u}{\partial t_2} = \frac{\partial}{\partial t_0}\left(\frac{u^3}{6} + \frac{1}{24}\left(2u\frac{\partial^2 u}{\partial t_0^2} + \left(\frac{\partial u}{\partial t_0}\right)^2\right) + \frac{1}{240}\frac{\partial^4 u}{\partial t_0^4}\right), \quad \ldots.$$

Define the generating function

$$\mathcal{Z}^{WK} = \sum_{g\geq 0} F_g,$$

which, unlike $\mathcal{F}^{WK}$, omits the $\hbar$ dependence for simplicity.

Theorem 4.5 [Wit91; Kon92] *The function*

$$u = \frac{\partial^2 \mathcal{Z}^{WK}}{\partial t_0^2}$$

satisfies the KdV equation (4.4).

To show that $\mathcal{F}^{WK}$ satisfies the full KdV hierarchy, one requires the initial conditions provided by the string and dilaton equations. This theorem enables recursive computation of all correlators (see [Zvo14, p. 60] for explicit recursion formulas).

5 Gromov-Witten Invariants and Hierarchies

A fundamental example in enumerative geometry is Kontsevich's formula for rational plane curves, which provides a recursive method for answering the question: *"How many rational plane curves of degree d pass through $3d-1$ given points in general position?"* This result arises from Kontsevich and Manin's development of Gromov-Witten theory [KM94]. In this section, we review the basic elements of Gromov-Witten classes and informally discuss their connection to integrable hierarchies. For an algebraic geometric perspective on Gromov-Witten classes, we follow [CK99]. Readers seeking further details are encouraged to consult [GP98; KM94].

Let X be a complex non-singular projective variety, and let $d \in H^2(X, \mathbb{Z})$. A tuple $(C, p_1, \ldots, p_n, f)$ is called a *stable map* if it satisfies the following:

- $(C, p_1, \ldots, p_n)$ is a compact and nodal Riemann surface of genus g with n marked points.
- $f : C \to X$ is a morphism such that every irreducible component $\tilde{C}$ of C satisfies one of the following:

i. $\tilde{C}$ has genus zero, at least three marked points, and is contracted (i.e., $f|_{\tilde{C}}$ is constant).

ii. $\tilde{C}$ has genus one, at least three marked points, and is contracted.

iii. $\tilde{C}$ has genus at least two or is not contracted.

A *stable map with class* d satisfies $f_*([C]) = d$, where $[C]$ is the Poincaré dual of C. If $d = [C']$ for a curve $C' \subset X$ and f is injective, then f parametrizes $C' = f(C)$. If X is zero-dimensional (i.e., a point), a stable map reduces to a stable curve. The *moduli space of stable maps* $\overline{M}^X_{g,n,d}$ classifies stable maps up to isomorphism. In particular, if $X = \mathrm{pt}$ and $d = 0$, we recover $\overline{M}_{g,n}$, and if $d = 0$ for general X, then every stable map is constant, yielding $\overline{M}^X_{g,n,0} \simeq \overline{M}_{g,n} \times X$.

As with the moduli space of curves, $\overline{M}^X_{g,n,d}$ can be viewed as a complex orbifold of dimension:

$$\dim_{\mathbb{C}}(\overline{M}^X_{g,n,d}) = (1-g)(\dim X - 3) + n - \int_d \omega_X.$$

Since this dimension may not be well-defined, the *expected dimension* e is defined via the virtual fundamental class $[\overline{M}^X_{g,n,d}]^{\mathrm{vir}} \in H_{2e}(\overline{M}^X_{g,n,d}) \simeq H^0(\overline{M}^X_{g,n,d})$. If the moduli space has pure dimension, then $\dim_{\mathbb{R}} \overline{M}^X_{g,n,d} = 2e$. Throughout this section, we assume rational cohomology.

Several important morphisms are associated with $\overline{M}^X_{g,n,d}$:

- The *evaluation map* at the i-th marked point:

$$\mathrm{ev}_i : \overline{M}^X_{g,n,d} \to X, \quad (C, p_1, \ldots, p_n, f) \mapsto f(p_i).$$

- The *forgetful morphism* that forgets the map f and stabilizes if necessary:

$$\mu : \overline{M}^X_{g,n,d} \to \overline{M}_{g,n}.$$

- The *forgetful morphism* that removes a marked point:

$$\pi : \overline{M}^X_{g,n+1,d} \to \overline{M}^X_{g,n,d}.$$

The *Gromov-Witten class* $\mathrm{I}_{g,n,d}$ is a map

$$\begin{aligned} \mathrm{I}_{g,n,d} : (H^\bullet(X))^{\otimes n} &\to H^\bullet(\overline{M}_{g,n}) \\ \gamma_1 \otimes \ldots \otimes \gamma_n &\mapsto \mathrm{I}_{g,n,d}(\gamma_1, \ldots, \gamma_n), \end{aligned}$$

which can be identified with $\mathrm{I}_{g,n,d} \in H^\bullet(\overline{M}_{g,n}) \otimes (H^\bullet(X)^*)^{\otimes n}$. It is explicitly given by

$$\mathrm{I}_{g,n,d}(\gamma_1, \ldots, \gamma_n) = \mu_*(\mathrm{ev}_1^*(\gamma_1) \smile \ldots \smile \mathrm{ev}_n^*(\gamma_n)).$$

Here, the pushforward μ_* is taken in the cohomology ring using Poincaré duality.

The corresponding *Gromov-Witten primary invariants* are defined as

$$\langle \mathrm{I}_{g,n,d}(\gamma_1,\dots,\gamma_n)\rangle = \int_{\overline{M}_{g,n}} \mathrm{I}_{g,n,d}(\gamma_1,\dots,\gamma_n).$$

The *i-th ψ-class* is defined as before. Moreover the *Gromov-Witten descendant invariants* are defined by

$$\langle \tau_{i_1}(\gamma_1)\dots\tau_{i_n}(\gamma_n)\rangle^X_{g,n,d} = \int_{\overline{M}_{g,n}} \mathrm{I}_{g,n,d}(\gamma_1,\dots,\gamma_n)\psi_1^{i_1}\dots\psi_n^{i_n},$$

where $\gamma_k \in H^\bullet(X)$ and $i_k \in \mathbb{Z}_{\geq 0}$. Computing Gromov-Witten invariants for a given variety X is a challenging task. Kontsevich's formula for rational plane curves provides a recursive method for the case $X = \mathbb{CP}^1$, as shown in [CK99, p.196]. One approach to gaining deeper insight into these invariants is through their generating function, inspired by the effectiveness of Witten's conjecture [Wit91]. The idea is to establish a connection between these generating functions and integrable hierarchies. In the current literature, the generating function associated with a Gromov-Witten theory is known as the *Gromov-Witten descendant potential*, analogous to the Witten-Kontsevich descendant potential.

The connection between Gromov-Witten theory and integrable hierarchies was first made explicit by Dubrovin and Zhang [DZ98; DZ01], who constructed an integrable hierarchy associated with the Gromov-Witten descendant potential. Givental's and Teleman's classification of semi-simple Gromov-Witten theories [Tel12; Giv01a; Giv01b; Giv04], combined with Dubrovin-Zhang's results, established that any semi-simple Gromov-Witten theory corresponds to an integrable hierarchy, now known as the *Dubrovin-Zhang (DZ) hierarchy*. This key result states that an explicit tau function of the DZ hierarchy encodes the potential of the underlying Gromov-Witten theory [BPS14].

More recently, Buryak [Bur15] constructed a new integrable hierarchy, the *Double Ramification (DR) hierarchy*, which is also associated with a given Gromov-Witten class. This hierarchy is built using the geometry of pairing $\lambda_g \mathrm{DR}_g(A)$ with a given Gromov-Witten class, where $\mathrm{DR}_g(A) \subset \overline{M}_{g,n}$ is the *double ramification cycle* and λ_g is the top Chern class of the Hodge bundle on $\overline{M}_{g,n}$. A fundamental conjecture, formulated by Buryak, states that in the semi-simple case, the DR hierarchy is related to the DZ hierarchy via a normal Miura transformation [Bur15; BDGR18], a statement now known as the *strong DR/DZ equivalence conjecture*. Furthermore, these authors [BGR19; BDGR20] proposed a set of conjectural relations in the tautological ring $RH^\bullet(\overline{M}_{g,n})$, which, if true, would imply the strong DR/DZ equivalence conjecture. Very recently, this set of conjectural relations was proven to be true [BLS24].

Building on the DR hierarchy, Buryak and Rossi introduced a deformation quantization of this integrable system, constructing the *Quantum Double Ramification (qDR) hierarchy* [BR16a]. In this framework, the hierarchy is developed using the geometry of pairing $\Lambda(\epsilon)\mathrm{DR}_g(A)$ with a given Gromov-Witten class, where $\Lambda(\epsilon)$ represents the sum of all Chern classes of the Hodge bundle on $\overline{M}_{g,n}$. Just as with the DR hierarchy, the entire qDR hierarchy can be reconstructed from its primary Hamiltonian density using the double ramification recursion relations [BR16b].

This leads to several natural questions:

- What additional enumerative geometric information does the quantized hierarchy encode beyond the classical DR hierarchy?
- What is the precise relationship between the qDR and DZ hierarchies?
- Can the Dubrovin-Zhang hierarchy itself be quantized, and if so, is there a potential connection between qDR and a possible quantum Dubrovin-Zhang (qDZ) hierarchy?

Addressing these questions is central to my doctoral research.

References

[ALR07] A. Adem, J. Leida, and Y. Ruan, "Orbifolds and string topology". Cambridge University Press, 2007.

[BLS24] X. Blot, D. Lewanski, and S. Shadrin, *On the strong DR/DZ conjecture*. In: (2024). arXiv: 2405.12334.

[Bur15] A. Buryak, *Double ramification cycles and integrable hierarchies*. In: Communications in Mathematical Physics (2015), 1085—1107. arXiv: 1403.1719.

[BDGR18] A. Buryak, B. Dubrovin, J. Guéré, and P. Rossi, *Tau-structure for the double ramification hierarchies*. In: Communications in Mathematical Physics (2018), 191–260. arXiv: 1602.05423.

[BDGR20] A. Buryak, B. Dubrovin, J. Guéré, and P. Rossi, *Integrable systems of double ramification type*. In: International Mathematics Research Notices 2020 24 (2020), pp. 10381–10446. arXiv: 1609.04059.

[BGR19] A. Buryak, J. Guéré, and P. Rossi, *DR/DZ equivalence conjecture and tautological relations*. In: Geometry & Topology 23.7 (2019), pp. 3537–3600. arXiv: 1705.03287.

[BPS14] A. Buryak, H. Posthuma, and S. Shadrin, *A polynomial bracket for the Dubrovin-Zhang hierarchies*. In: (2014). arXiv: 1009.5351.

[BR16a] A. Buryak and P. Rossi, *Double ramification cycles and quantum integrable systems*. In: Letters in Mathematical Physics 106.3 (2016), pp. 289–317. arXiv: 1503.03687.

[BR16b] A. Buryak and P. Rossi, *Recursion relations for Double Ramification Hierarchies*. In: Communications in Mathematical Physics 342.2 (2016), pp. 533–568. arXiv: 1411.6797.

[CK99] D. Cox and S. Katz, "Mirror symmetry and algebraic geometry". Monografias de Matematica. American Mathematical Society, 1999.

[DZ98] B. Dubrovin and Y. Zhang, *Bi-Hamiltonian hierarchies in 2D topological field theory at one-loop approximation*. In: Communciations in Mathematical Physics 198.2 (1998), 311–361. arXiv: hep-th/9712232.

[DZ01] B. Dubrovin and Y. Zhang, *Normal forms of hierarchies of integrable PDEs, Frobenius manifolds and Gromov-Witten invariants*. In: Advances in Mathematics (2001), p. 189. arXiv: math/0108160.

[Eyn18] B. Eynard, "Lectures notes on compact Riemann surfaces". In: (2018). arXiv: 1805.06405.

[Gat00] L. Gatto, *Intersection theory on moduli spaces of curves*. Monografias de Matematica. Instituto Nacional de Matematica Pure e Aplicada, 2000.

[GP98] E. Getzler and R. Pandharipande, *Virasoro constraints and the Chern classes of the Hodge bundle*. In: Nuclear Physics B (1998), pp. 701–714. arXiv: math/9805114.

[Giv01a] A.B. Givental, *Gromov-Witten invariants and quantization of quadratic Hamiltonians*. In: Moscow Mathematics Journal 1.4 (2001), 551–568. arXiv: math/0108100v2.

[Giv01b] A.B. Givental, *Semisimple Frobenius structures at higher genus*. In: International Mathematics Research Notices 23 (2001), 1265–1286. arXiv: math/0008067.

[Giv04] A.B. Givental, *Symplectic geometry of Frobenius structures*. In: Frobenius manifolds. Quantum cohomology and singularities. Aspects of Mathematics E.36 (2004), 91–112. arXiv: math/0305409.

[HM91] J. Harris and I. Morisson, "Moduli of curves". Graduate Texts in Mathematics. Spring, 1991.

[KV07] J. Kock and I. Vainsencher, "An invitation to quantum cohomology. Kontsevich?s formula for rational plane curves". Birkhäuser Boston, MA, 2007.

[Kon92] M. Kontsevich, *Intersection Theory on the Moduli Space of Curves and the Matrix Airy Function*. In: Communications in Mathematical Physics 147 (1992), pp. 1–23.

[KM94] M. Kontsevich and Y. Manin, *Gromov-Witten classes, quantum cohomology, and enumerative geometry*. In: Communications in Mathematical Physics 164.3 (1994), 525–562. arXiv: hep-th/9402147.

[Nic11] L. Nicolaescu, *Intersection theory*. In: (2011). ePrint: Lecture Notes.

[Sch20] J. Schmitt, *The moduli space of curves*. In: (2020). ePrint: Lecture Notes.

[Tel12] C. Teleman, *The structure of 2D semi-simple field theories*. In: Inventiones Mathematicae 188.3 (2012), 525–588. arXiv: 0712.0160.

[Wit91] E. Witten, *Two-dimensional gravity and intersection theory on moduli space*. In: Surveys in differential geometry 1 (1991), 243–310.

[Zvo14] D. Zvonkine, *An introduction to moduli spaces of curves and their intersection theory*. In: (2014). ePrint: Lecture Notes.

Dynamics of Environment-Embedded Quantum Systems: An Introduction

Pietro De Checchi (*)

Abstract. Closed quantum systems are an idealization, their time evolution described by the Schrödinger Equation, i.e. by the action of unitary operators. The physics of a realistic quantum system, on the other hand, is bound to be disturbed by the environment in which it is naturally embedded and with which it inevitably interacts. The dimension of the space needed to fully describe the composite system increases, as one would have to include all, possibly infinite, environmental variables, leading to intractable problems. To reduce the system to a smaller subspace of interest and to describe its correct dynamics, many strategies have been developed. These systems have in general non-unitary dynamics and are known as Open Quantum Systems. We introduce some of the main approaches based on various techniques, from dynamical semigroup generators, stochastic unravelings and bottom-up modelling.

1 Introduction

In this work, I will introduce some of the most common approaches to describe the dynamics of quantum systems embedded in an environment.

The starting point is a brief recap of the basis of closed quantum systems and their dynamical evolution, moving from the most known wave-function frame, to the density matrix approach, which generalizes the former and allows for the description of a broader set of states. Using this basis, the composition of the system from different subsystems is introduced. The search for the description of the dynamics of only one subsystem, considering its interaction with the others, leads to contractive mappings and the theory of Open Quantum Systems.

In Figure 1, a commutative diagram as a conceptual map of the main routes and strategies to describe the maps $\mathcal{E}$. The exact one is the unfeasible one, treating the whole system up to being closed (the universe as a system), which I referred to as a *deterministic equation for probabilistic objects*. I will introduce three main feasible approaches: an algebraic derivation of the most used Markovian approximation of the dynamics, a bottom-up approach for a microscopic derivation, which can be thought of as *master equations for*

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: dechecch@math.unipd.it . Seminar held on 5 December 2024.

probabilistic objects, and finally one of the stochastic approaches, *probabilistic equations for probabilistic objects*.

Probabilistic intuition and physical interpretation are remarked on throughout the work.

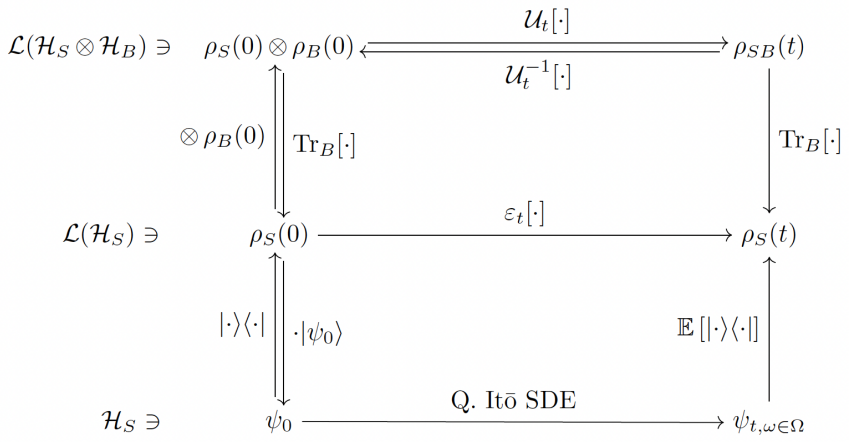


Figure 1: A conceptual map as a commutative diagram, illustrating the different strategies to describe the subsystem state $\rho_S(t)$ evolution in time, the spaces these objects live in.

1.1 Notation and Conventions

Bra-ket notation: ket $|v\rangle$ denotes a vector v in a complex vector space, the bra $\langle v| = (|v\rangle)^\dagger$ denotes the Hermitian conjugate of the ket v
Hilbert Spaces are denoted as $\mathcal{H}_A$ with the subscript A denoting which system the space refers to
Operators are notated in italic capital letters A, capital subscripts denoting the space they act on, and lowercase letters index the elements of the operator in matrix notation.
Unitaries transforms are denoted as the particular operators U
Maps and generators are denoted by calligraphic capital letters, unitary maps as the particular map $\mathcal{U}$
Indexes: Latin letters are used for the system of interest and Greek letters for the environment.
Orthonormality/orthonormal basis in a Hilbert spaces: $\{|i\rangle, \not\perp\} = \{|i\rangle \quad \text{s.t.} \quad \langle i|j\rangle = \delta_{ij}\}$, while the opposite is written with the simple negation: $\not\perp$, and orthogonality is stated as usual as $\perp$.

2 Basis of (Closed) Quantum Systems

Closed quantum systems are usually introduced during undergraduate courses and can be described in the wave function (wave-vector) formalism.

$$|\psi\rangle = \sum_{j=1}^{d} c_j |\phi_j\rangle\,, \quad |\psi\rangle \in \mathcal{H}_d = \mathbb{C}^d, \quad \{|\phi_j\rangle \perp^1 \mathcal{H}_d\}_{j=1}^{d} \tag{2.1}$$

where c_j are complex coefficients of each vector $|\phi_j\rangle$, which together compose an orthonormal basis of the Hilbert space $\mathcal{H}_d$.

The probabilistic nature of quantum mechanics depends on the meaning of the square norm of the coefficients of the state vectors. The coefficients $c_j \in \mathbb{C}$ are **probability amplitudes**:

$$\mathbb{P}(\text{meas. } |\phi_j\rangle) = |\langle\phi_j|\psi\rangle|^2 = |c_j|^2\,, \qquad \sum_{j=1}^{d} |c_j|^2 = 1 \tag{2.2}$$

We can summarize the physical meaning and interpretation of the objects we deal with in this formalism. The wavefunction ψ describes a probability amplitude, its norm squared $|\psi|^2$ a probability density, and the probability of measuring the system in d$\mathbf{r}$ being $|\psi|^2 \mathrm{d}\mathbf{r}$.

The interpretation is that identical measures on a set of identical replicas of the system give the probabilistic meaning, the action of the measure $\mathcal{M}_k$

$$|\psi\rangle \mapsto \frac{\mathcal{M}_k}{\sqrt{p_k}}|\psi\rangle \equiv |\psi_k\rangle \qquad p_k = \langle\psi|\mathcal{M}_k^\dagger \mathcal{M}_k|\psi\rangle \geq 0, \tag{2.3}$$

the state of the system instantaneously collapsing after the measurement, with probability p_k. Every physically measurable quantity is associated with an observable, i.e. a Hermitian operator with spectral decomposition:

$$A = \sum_a \lambda_a |a\rangle \langle a| \tag{2.4}$$

where $\{|a\rangle \langle a|\}$ is the set of projectors, projecting measures on the eigenvectors $\{|a\rangle\}$ subspaces. Expectation values are the mean value of the measure of an observable.

The time evolution of a closed quantum system, i.e. a system that evolves under the action of unitary operators, is given by the Schrödinger equation (SE):

$$\frac{\mathrm{d}}{\mathrm{d}t}|\Psi(t)\rangle = -\frac{i}{\hbar} H |\Psi(t)\rangle \tag{2.5}$$

where H is the Hermitian operator known as the Hamiltonian. In the following, for simplicity, we set $\hbar = 1$. The solution of this differential equation gives an equivalent postulate, and it highlights the presence of the unitary operator describing the evolution $U(t)$, called the *propagator*

$$|\Psi(t)\rangle = U(t)|\Psi(0)\rangle \tag{2.6}$$

.

When the quantum system is isolated, the Hamiltonian is time-independent, and the evolution of the statevector is simply given by:

$$|\Psi(t)\rangle = e^{-iHt}|\Psi(0)\rangle \tag{2.7}$$

hence $U(t) = e^{-iHt}$, with initial condition $U(0) = \mathbb{1}$. For simplicity, we will consider only time-independent Hamiltonians, for time-dependent Hamiltonians and semiclassical dissertations on light-matter interactions refer to any undergraduate textbook in physics.

3 From Closed to Open Quantum Systems

3.1 Density matrix formalism

Because of the probabilistic nature of quantum mechanics, as stated above, the observables of any quantum system can be obtained only by a statistical average of repeated measurements on copies of the systems, which is the same as measuring an ensemble of independent systems described by the same statevector. It is convenient then to introduce the density matrix formalism when dealing with ensembles of quantum systems. More importantly, the individual systems composing the ensemble can assume different states, and this is equivalent to saying that the state of a system is not always perfectly known.

Postulate 1 of Quantum Mechanics: State Space

i) To every quantum system is associated a Hilbert space $\mathcal{H} = \mathbb{C}^d$ equipped with inner product $\langle v|w\rangle = \sum_{j=1}^{d} v_j^* w_j = c_{vw} \in \mathbb{C}$

ii) **def. (density matrix)** A state of the system is a *positive semidefinite self-adjoint* linear operator acting on $\mathcal{H}$ with *unitary trace*

$$\rho = \sum_k p_k |\psi_k\rangle\langle\psi_k| \,, \quad |\psi_k\rangle \in \mathcal{H} \,, \quad \mathrm{Tr}\rho = 1 \,, \quad p_k \in [0,1] \,, \quad \rho \in \mathcal{L}(\mathcal{H})$$

iii) **def.** ρ is a *pure state* if $\rho = |\psi_k\rangle \langle\psi_k|$, i.e. $p_k = 1$ for some k.

We see that we can write a generic density matrix as the sum of pure states weighted by the probability of being measured.

Remark 1 The set

$$p = \left\{ p_k \,:\, p_k \in [0,1], \sum_{k=1}^{d} p_k = 1 \right\}$$

is a **probability distribution** of a pure states ensemble.

Remark 2 The trace operator ($\mathrm{Tr} : \mathcal{H} \to \mathbb{C}$) acts as an **average**.
Let $A = \sum_a \lambda_a |a\rangle \langle a| \in \mathcal{L}(\mathcal{H})$ Hermitian operator, the expectation value of A for the system ρ is

$$\langle A \rangle_\rho = \sum_a \lambda_a p_a = \mathrm{Tr}[A\rho]$$

Definition 1 (Purity) The purity of the state ρ is defined as the quantity

$$P \equiv \mathrm{Tr}(\rho^2) \tag{3.1}$$

A state is called *pure* if $P = 1$ and *mixed* if $P < 1$.

Remark 3 Note that the density matrix ρ is: (i) a state of the system by definition, (ii) a projection operator, (iii) a probability density.

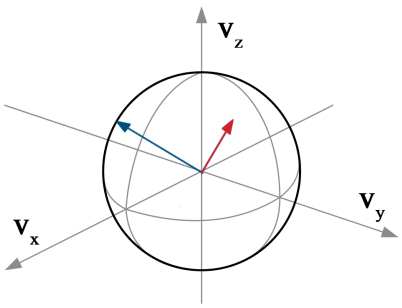


Figure 2: Bloch sphere representation for a qubit state. A generic pure state is represented with a blue arrow, a mixed one with a red arrow.

In this picture, the dynamics of the system are given by the Liouville-Von Neumann equation, which reads:

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho(t) = -i[H, \rho(t)] \tag{3.2}$$

where $[A, B]$ is the commutator $AB - BA$. Easily derived from the SE solution and the definition of the density matrix ρ, it has the formal solution:

$$\rho(t) = U(t)\rho(0)U^\dagger(t) = \mathcal{U}_t[\rho(0)] \tag{3.3}$$

where $\mathcal{U}_t[\cdot]$ is the dynamical map of the evolution. Again, the evolution of the system is unitary.

Let's visualise in a geometric representation a simple example, to understand the fundamental difference between the two formalisms. Take a two-level system $\rho \in \mathcal{L}(\mathcal{H} = \mathbb{C}^2)$, its density matrix is

$$\rho = \begin{pmatrix} a & b \\ b^* & 1-a \end{pmatrix} \tag{3.4}$$

by Hermiticity and unitary trace conditions. This, by the condition of positivity of the eigenvalues, can be parametrised and decomposed in a form:

$$\rho = \sum_{i=x,y,z} \sigma_i v_i = \frac{1}{2}\begin{pmatrix} 1+v_z & v_x - iv_y \\ v_x + iv_y & 1-v_z \end{pmatrix} \tag{3.5}$$

where σ_i are Pauli matrices $\sigma_x, \sigma_y, \sigma_z$ and the vector $v = (v_x, v_y, v_z)$ is called Bloch vector, with $\|v\| \leq 1$. Now, we can visualise in three dimensions the state, Fig. 2. This simple two-level system is what is usually referred to as a qubit in quantum computing.

The magnitude of the Bloch vector is related to the purity of the system, so vectors with unitary length describe pure states, otherwise mixed states. Phrased differently, pure states live on the sphere's surface, and general mixed states in the volume.

In closed dynamics, we can move only on the surface of a sphere or radius $\|v\|$, since unitary evolutions are nothing more than rotations of the states. What changes in open systems dynamics?

3.2 Complex systems: composition with an environment

What we have discussed so far are the closed dynamics of a single system. Let now consider a more complex setting, a total system composed of two parts. The second postulate of quantum mechanics follows.

Postulate 2 of Quantum Mechanics: Expansion Take two quantum systems and their Hilbert spaces,

$$\mathcal{H}_A\,,\ \{|a\rangle \perp^1 \mathcal{H}_A\}_1^N \quad ; \quad \mathcal{H}_B\,, \{|\mu\rangle \perp^1 \mathcal{H}_B\}_1^M$$

the composite Hilbert space of the super-system is:

$$\mathcal{H} = \mathcal{H}_A \otimes \mathcal{H}_B = \text{span}\{|a\rangle_A \otimes |\mu\rangle_B\} \tag{3.6}$$

We can define pure states ensembles as

$$\{|\psi_k\rangle \in \mathcal{H}\,,\ p_k\}, \quad |\psi_k\rangle = \sum_{i,\mu} c_{k;a,\mu} |a\rangle_A \otimes |\mu\rangle_B \tag{3.7}$$

and the associated generic density matrix

$$\rho = \sum_k p_k |\psi_k\rangle\langle\psi_k| = \sum_{ab\mu\nu} \lambda_{ab\mu\nu} |a\rangle_A \langle b| \otimes |\mu\rangle_B \langle\nu| \tag{3.8}$$

where

$$\lambda_{ab\mu\nu} = \sum_l p_l c_{l;i\mu} c^*_{l;j\nu} \tag{3.9}$$

This matrix describes the total system both when it is correlated and not correlated, with the latter being a special case: only when we can write $\lambda_{ij\mu\nu} = \lambda^A_{ij}\lambda^B_{\mu\nu}$ we can express the total density matrix as the tensor product of the density matrices of the two subsystems $\rho = \rho_A \otimes \rho_B$. In all the other cases the two systems are said to be *correlated*.

This is a realistic situation. We usually refer to only one part of the total system as the *system*, a small part ($\mathcal{H}_S$) of what can be a protein, a laboratory, the entire universe ($\mathcal{H}$), and the remaining sub-system as the *bath* or *environment*.

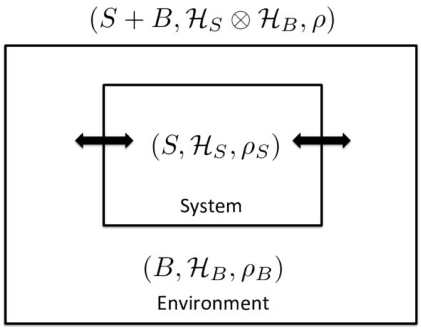


Figure 3: Schematic representation of an open quantum system.

Let's cast this to the typical case when studying molecular processes. Normally, the dimension of the Hilbert space of the system S is quite small, more generally it has a finite dimension d_S, while the bath contains all the solvent degrees of freedom, the molecular vibrations that are not of interest and so on, as such usually $d_B \to \infty$.

So what about when we are interested in a small part (S) of the total system (*universe*)? We introduce the partial trace, to recover only the information about our system S.

Definition 2 (Partial trace operator) Let $\mathcal{H} = \mathcal{H}_S \otimes \mathcal{H}_B$, and $A = N_S \otimes M_B$ generic operator acting on $\mathcal{H}$. The partial trace is the linear operator $\mathrm{Tr}_B : \mathcal{H} \mapsto \mathcal{H}_S$ acting

$$\mathrm{Tr}_B(N_S \otimes M_B) = N_S \mathrm{Tr}(M_B) = \tilde{N}_S \in \mathcal{H}_A \tag{3.10}$$

Therefore, take a generic universe described by the density matrix in (3.8), and look only at the subsystem S by applying the partial trace for the environment subsystem:

$$\rho_S = \mathrm{Tr}_B(\rho_{SB}) = \sum_{\omega \in \mathcal{H}_B} \langle \omega | \rho_{SB} | \omega \rangle = \sum_{\substack{a,b \in \mathcal{H}_S \\ \omega}} \lambda_{ab\omega} |a\rangle_S \langle b| \tag{3.11}$$

where $\{|\omega\rangle \perp^1 \mathcal{H}_B\}$ is basis of the environment space. The coefficients $\lambda_{ab\omega}$ and then the *reduced density matrix*, and the partial trace operation itself, still depend on the knowledge of the whole system. This shows the need for a simplified approach, an approximated model, to be able to describe such systems and their dynamics.

3.3 The need for a simplified approach

With equations (3.2) and (3.3) we can write the universe dynamics, and operating with the partial trace we can recover the exact dynamics of the subsystem of interest:

$$\rho_S(t) = \mathrm{Tr}_B\{U(t)\rho_{SB}(0)U^\dagger(t)\} \tag{3.12}$$

We can expand the partial trace over the bath, supposing to be able to know an orthonormal basis $\{|\mu\rangle \perp^1 \mathcal{H}_B\}$ for it:

$$\rho_S(t) = \sum_{\mu \in \mathcal{H}_B} \langle \mu | U(t)\rho(0)U^\dagger(t) | \mu \rangle \tag{3.13}$$

Now, an assumption needs to be made. We consider the initial state decoupled, i.e. factorized into the system and bath components, $\rho = \rho_S \otimes \rho_B$. Whilst it is not an approximation by itself, it is a very unrealistic and rare case, so we are narrowing the actual physical system we are describing or we can consider it as an approximation of a larger set. For example, it can be justified and accepted for an excitonic system if it can be considered instantaneously excited at $t = 0$ (sudden approximation), with no correlation between the excited system and the environment at the previous times.

Considering an orthonormal basis decomposition of the bath ($\rho_B = \sum_\nu \lambda_\nu |\nu\rangle \langle\nu|$) we can write:

$$\rho_S(t) = \sum_{\mu\nu \in \mathcal{H}_B} \sqrt{\lambda_\nu}\langle\mu|U(t)|\nu\rangle \rho_S(0) \sqrt{\lambda_\nu}\langle\nu|U^\dagger(t)|\mu\rangle \tag{3.14}$$

where we can identify the environment-averaged operators driving the dynamics of the system, the Kraus operators:

$$K_{\mu\nu}(t) = \sum_{\mu\nu} \sqrt{\lambda_\nu}\langle\mu|U(t)|\nu\rangle \tag{3.15}$$

and recast the time evolution of the system density matrix in what is called an Operator Sum Representation (Kraus OSR): [1]

$$\rho_S(t) = \sum_\alpha K_\alpha(t)\rho_S(0)K_\alpha^\dagger(t) \tag{3.16}$$

Hence, with just one assumption on the initial state of the universe system, we have written a simple form for the evolution of the system. Therefore, the OSR representation is a map of the evolution of the system, possibly exact.

$$\mathcal{E} : \mathcal{H}_S \to \mathcal{H}_S \tag{3.17}$$

$$\rho_S(0) \mapsto \rho_S(t) = \sum_\alpha K_\alpha(t)[\rho_S(0)]K_\alpha(t)^\dagger \tag{3.18}$$

This formulation holds the following properties:

(a) Trace preservation:

$$\mathrm{Tr}[\mathcal{E}(\rho)] = \sum_\alpha \mathrm{Tr}[K_\alpha(t)\rho K_\alpha(t)^\dagger] = \mathrm{Tr}[\sum_\alpha K_\alpha(t)^\dagger K_\alpha(t)\rho] = \mathrm{Tr}[\rho]$$

The fact that $\sum_\alpha K_\alpha^\dagger K_\alpha = I$ comes from the definition of Kraus operators, and can be intended later as a trace-preserving constraint.

(b) Linearity:

$$\mathcal{E}(c_1\rho_1 + c_2\rho_2) = c_1\mathcal{E}(\rho_1) + c_2\mathcal{E}(\rho_2)$$

By direct substitution, for any scalars c_1, c_2.

3a. Positivity:
$$\langle a|\mathcal{E}(A)|a\rangle \geq 0, \forall |a\rangle \in \mathcal{H}_S$$

3b. Complete positivity: any operator written in Kraus OSR satisfies the Let then denote $\mathcal{H}_R$ any ancillary Hilbert space, of dimension $k = \dim(\mathcal{H}_R)$, and the identity $\mathbb{1}_R^{(k)}$ for any k, then
$$\mathcal{E} \otimes \mathbb{1}_R^{(k)} \geq 0, \forall k > 0$$

Theorem 1 (Choi-Kraus theorem) *A linear map $\mathcal{E} : \mathcal{L}(\mathcal{H}) \to \mathcal{L}(\mathcal{H})$ is completely positive and trace-preserving (CPTP) if and only if it has a Kraus operator sum representation:*

$$\mathcal{E}[\cdot] = \sum_\alpha K_\alpha(t)[\cdot]K_\alpha(t)^\dagger$$

with the operator $K_\alpha(t) \in \mathcal{B}(\mathcal{H})$ such that:

$$\sum_\alpha K_\alpha(t)K_\alpha(t)^\dagger = \mathbb{1}$$

The proof of the theorem can be found in [1-3], in a more synthetic way in [4-7]. In particular, Choi's theorem gives sufficient conditions for a map to be completely positive, [3] and Kraus's theorem adds the condition for trace preservation and formalizes the operator sum representation. [1]

Then, we have a dynamical map for the evolution of a system from a time t_0, when the system satisfies the factorization condition, up to a generic time t, with the assumption of knowing the propagator for the universe and an orthonormal decomposition for the environment. The whole point is that these are quantities that we do not know, and even if we did, we can not start in an intermediate time $t' < t$ as we would not be in a factorized condition anymore. Further assumptions and approximations are then needed to have a structured and solvable description of the dynamics.

4 Gorini-Kossakowski-Sudarshan-Lindblad Form

Given a known system at time t_0 and its evolution to t_1, and the same for another time $t_2 > t_1$ one would expect by continuity of time, naïvely, the following composition law in (4.1). We have to rely on a second assumption, the Markovian nature of the map, so

$$\mathcal{E}_{t_2,t_0} = \mathcal{E}_{t_2,t_1}\mathcal{E}_{t_1,t_0} \tag{4.1}$$

i.e. the map $\mathcal{E}$ to be a contractive evolution family (contraction semigroup). That is a particular case, and to be true, and all the maps to be unital CPT maps, the map $\mathcal{E}$ is a contractive evolution family. This condition is called *divisibility condition*, and it means that the system loses memory of its evolution at any previous time. So, it is easy to see that the system has the Markov property, and therefore evolutions admitting CPT unital maps are called *Markovian evolutions*, and the system is referred as a *Markovian system.*

The key element of a semigroup of operators is its generator and, by Markovian assumption, it is linear, strongly continuous and admits an infinitesimal generator. We can then identify the generator of the semigroup by the solution of the first order ODE for the dynamics of the system:

$$\dot{\rho}_t = \mathcal{L}[\rho_t] \quad \rightarrow \quad \rho_t = e^{\mathcal{L}t}\rho_0 \tag{4.2}$$

Starting from the limit expression for the infinitesimal generator and using Kraus-Choi theorem, the Gorini-Kossakowski-Sudarshan and Lindblad Master Equations, [8, 9] is obtained, and the theorem follows:

Theorem 2 (Lindblad '76; Gorini,Kossakowski,Sudarshan '76) *The generator of any quantum operation satisfying the semigroup property must have the form*

$$\frac{d\rho_S(t)}{dt} = -i[H_S, \rho_S(t)] + \sum_k \gamma_k \left[L_k \rho_S(t) L_k^\dagger - \frac{1}{2} \left\{ L_k L_k^\dagger, \rho_S(t) \right\} \right] \tag{4.3}$$

where $\gamma_k > 0$ are relaxation rates, L_k are Lindblad operators and $\{\cdot, \cdot\}$ is the anticommutator.

Then, QME in the form of (4.3) are all and only admitting CTPT properties of the map, if not invoking non-Markovianity and various projection techniques bringing to complex and usually not solvable forms.

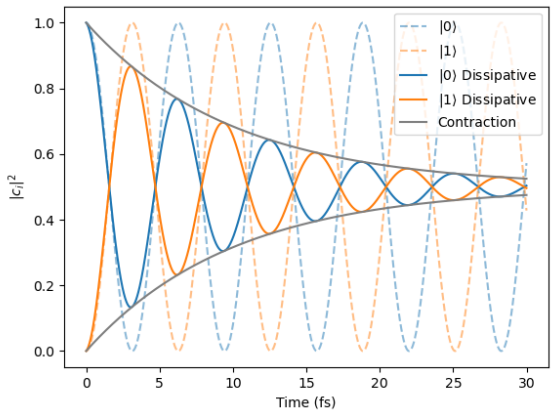


Figure 4: Example of the effect of dissipation of the dynamics and the contraction (grey enveloping line) of the space for a two-level quantum system with coupled states. For comparison, the dynamics of the closed system are in transparent dashed lines.

5 Bottom-up Approach: Redfield QME

There are many more methodologies to approach open quantum systems, even when still relying on most of the approximations and assumptions applied so far. Another tool

commonly used to deal with OQS is the Redfield Master Equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho_S(t) = -i[H_S, \rho_S(t)] - \sum_{\alpha\beta}[S_\alpha, Q_{\alpha\beta}\rho_S(t)] + [\rho_S(t)Q^\dagger_{\alpha\beta}, S_\alpha]$$

and we immediately see that it is not in Lindblad form: it is not a CP mapping! Then, why it is so common and important? In contrast to the Lindblad approach, this model is not simply algebraically derived and the application purely phenomenological. The derivation of the model is from a microscopically bottom-up approach, allowing for physical assumption for the system, the environment and their interaction.

We start defining the Hilbert spaces of the reduced system, $\mathcal{H}_S$, of the environment $\mathcal{H}_B$, and of the composite super-system $\mathcal{H} = \mathcal{H}_B \otimes \mathcal{H}_B$. We define the total Hamiltonian for the system as

$$H = H_S \otimes \mathbb{1} + \mathbb{1} \otimes H_B + H_I \tag{5.1}$$

where $H_S, H_B \in \mathcal{H}_S, \mathcal{H}_B$ describe the two isolated subsystems and the interaction Hamiltonian $H_I \in \mathcal{H}$ is a bilinear coupling defined as

$$H_I = \sum_\alpha S_\alpha \otimes B_\alpha \tag{5.2}$$

where $S_\alpha \in \mathcal{B}(\mathcal{H}_S), B_\alpha \in \mathcal{B}(\mathcal{H}_B)$ are interaction operators acting in the respective Hilbert spaces.

The associated Liouville-Von Neumann equation of the universe system is then:

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho(t) = [H_S \otimes \mathbb{1} + \mathbb{1} \otimes H_B + H_I, \rho(t)] \tag{5.3}$$

and we transform to the so-called interaction picture, or Dirac picture, hence to a co-moving frame so that

$$\tilde{\rho}(t) = e^{i(H_S+H_B)t}\rho(t)e^{-i(H_S+H_B)t} \tag{5.4}$$

and the interaction Hamiltonian is transformed similarly, becoming a time-dependent Hamiltonian.

We can now write the Liouville equation $\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\rho}(t) = [H_I(t), \tilde{\rho}(t)]$, integrate it and inserting the integrated definition of the density matrix obtained into the initial Liouville equation. We obtain the following open integrodifferential equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\rho}(t) = -i[\tilde{H}_I(t), \tilde{\rho}(0)] - \int_0^t [\tilde{H}_I(t), [\tilde{H}_I(t'), \tilde{\rho}(t')]]\,\mathrm{d}t' \tag{5.5}$$

To obtain the system evolution, our interest, we have to trace over the bath degrees of freedom,

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\rho}_S(t) = -i\mathrm{Tr}_B\{[H_I(t), \tilde{\rho}(0)]\} - \int_0^t \mathrm{Tr}_B\{[H_I(t), [H_I(\tau), \tilde{\rho}(\tau)]]\}d\tau \tag{5.6}$$

which is recasting that is still the exact description of the dynamic of the whole system. Now, we need to add some approximations to find a solution, i.e. to obtain an equation in closed form.

(a) Factorization of the initial system

$$\rho(0) = \rho_S(0) \otimes \rho_B(0) = \rho_S(0) \otimes \rho_B^{eq}$$

(b) Born Approximation: We consider a perturbative interaction, we expand the total density matrix at any time t as the series

$$\rho(t) = \rho_S(t) \otimes \rho_B^{eq} + \mathcal{O}(\lambda)$$

and we neglect the higher-order terms.

We can now substitute the factorized expression for both the density matrix and the interaction Hamiltonian, recall the action and linearity of the partial trace operator, Definition 2, and the cyclic property of trace operators, so that

$$\mathrm{Tr}_B\left[H_I(t)\tilde{\rho}(t)\right] = S_\alpha(t)\tilde{\rho}_S(t)\mathrm{Tr}\left[B_\alpha(t)\tilde{\rho}_B^{eq}\right] \tag{5.7}$$

the trace on the r.h.s. the expectation value of the simple coupling operator B_α, and consider it as a mean-field effect, which would simply shift the Hamiltonian, or assume that it vanishes (meaning that we can consider it to be included in the system Hamiltonian and therefore be null). Hence the first term of (5.6) vanishes.

In the integral term, we get the trace of the different combinations of the product of two operators acting at different times on the bath subsystem, which is identified as the correlation function of the bath:

$$C_{\alpha\beta}(t,\tau) = \mathrm{Tr}[B_\alpha(t)B_\beta(\tau)\rho_B^{eq}] \tag{5.8}$$

Here, we can evaluate different system bath models, either from physical assumptions or from actual models of the surrounding environment. As simple examples, considering a δ-correlated environment, very fast correlation times w.r.t. the system evolution, we fall back to Markovian assumption which will eventually lead to the Lindblad equation. On the other hand, we can use models: in the Caldeira-Legget model the bath is modelled by a set of independent quantum oscillators, from which we extract a complex correlation, or from overdamped Brownian oscillators a Drude-Lorentz spectral density, i.e. Ohmic spectral density with Lorentzian cutoff, and many others.

We obtain then a non-Markovian QME for the system:

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\rho}_S(t) = -\sum_{\alpha\beta}\int_0^t C_{\alpha\beta}(t,t')[\tilde{S}_\alpha(t), \tilde{S}_\beta(t')\tilde{\rho}_S(t')] + C_{\alpha\beta}(t',t)[\tilde{\rho}_S(t')\tilde{S}_\beta(t'), \tilde{S}_\alpha(t)]\,\mathrm{d}t' \tag{5.9}$$

as it depends on the ρ_S at all previous times, and the bath too is correlated. Then, two more approximations are needed to obtain a form in the Markovian framework:

(c) the first Markov approximation is to assume that $\rho_S(t)$ varies slower than the decay of the correlation of the bath, allowing to change

$$\rho_S(\tau)d\tau \to \rho_S(t)d\tau$$

(d) the second Markov approximation to remove the time dependency of the coefficients in the nested commutator. To do so we follow the same reasoning as above, and consider that, as ρ_S decays slowly with respect to the correlation function of the bath, the kernel decays fast enough to extend the integration upper bound to infinity

$$\int_0^t \to \int_0^{+\infty}$$

We finally obtain an equation which is time-local and closed for the system dynamics, the Redfield QME in the rotated frame:

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\rho}_S(t) = -\int_0^\infty d\tau \sum_{\alpha\beta} \{C_{\alpha\beta}(\tau)[\tilde{S}_\alpha(t), \tilde{S}_\beta(t-\tau)\tilde{\rho}_S(t)] + \text{h.c.}\} \tag{5.10}$$

where h.c. indicates the hermitian conjugate of the term inside the parenthesis. Moving back to the Schrödinger picture, expanding the operators S in the H_S eigenbasis, the Redfield QME is obtained

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho(t) = -i[H_S, \rho(t)] - \sum_{\alpha,\beta}\sum_{\omega,\omega'} \Gamma_{\alpha\beta}(\omega') \left(\left[S^\dagger_{\mathrm{rot},\alpha}(\omega), S_{\mathrm{rot},\beta}(\omega')\rho(t) \right] + \text{h.c.} \right). \tag{5.11}$$

where the coefficients $\Gamma_{\alpha\beta}(\omega')$ are defined

$$\int_0^\infty C_{\alpha\beta}(\tau) e^{i\omega'\tau}\, \mathrm{d}\tau \tag{5.12}$$

Further assumptions can be considered and approximations can be applied, leading to different recastings of this form. One notable example is the recovery of a Lindblad form from this microscopic approach applying the secular approximation, commonly known also as rotating wave approximation, considering fast oscillating terms null. [5, 10]

6 Stochastic Schrödinger Equation

Approaching open quantum systems *via* stochastic Schrödinger equations is a method to use effective equations to describe the interaction of the quantum system with its environment. These methods are called *unravelings* of the respective QME, and gained of importance for many aspects, most notably their application to Quantum Computing implementations and control. In the unraveling scheme, the possibly-mixed density matrix of the system is obtained as the average of pure state dynamics with stochastic driving. The clear upside is the scaling of the numerical implementation, $\mathcal{O}(N)$ instead of $\mathcal{O}(N^2)$, the problem with the scaling with the number of trajectories required easily overcome by the

ability of efficiently parallelize on classical architectures, and since we deal with unitary evolutions we can implement is on QC architectures, harnessing the need for repeated runs *in-lieu* of the parallelization. On the downside, we are limited in the choice of operators depending on the stochastic process used and therefore limiting also the application for the QC applications.

The simple and general way to formulate an is starting with a linear homogeneous stochastic differential equation (SDE) of the following form,

$$\mathrm{d}\tilde{\psi}_t = A\tilde{\psi}_t\,\mathrm{d}t + B\tilde{\psi}_t\,\mathrm{d}X_t \tag{6.1}$$

where the operator A should contain the deterministic evolution of the system as if isolated, and where B is the operator of the stochastic fluctuations. This simple SSE approach does not ensure, in principle, the preservation of the norm of the state vector $\psi(t)$. Hence, normalized states must be introduced:

$$\psi_t = \frac{\tilde{\psi}_t}{\|\tilde{\psi}_t\|} \tag{6.2}$$

In this way, the mean density matrix describing a mixed state is unraveled into an ensemble of pure states. To do so, a change of distribution is required to ensure the correct mean, and this is given by a Girsanov transformation (Theorem A.45 in [11]) In terms of the SSE storm of trajectories, this transformation fulfilling this transformation is given by the *a priori* normalization of the wavefunction ensuring the martingale property $\|\psi(0)\|^2 = \|\psi(t)\|^2 = 1$. This gives the normalization condition

$$\mathbb{E}[||\psi_t||^2] = 1 \implies \mathrm{d}(\psi_t^\dagger \psi_t) = 0 \tag{6.3}$$

The form of the operators A and B are then obtained by computing the normalization above. Then, we can then write the normalized solutions, in terms of the average density matrix, as

$$\rho_t = \mathbb{E}\left[|\psi_t\rangle\langle\psi_t|\right] \tag{6.4}$$

notably the stochastic unraveling of the mean density matrix time evolution.

As a prototypical example, let's see how any QME in Lindblad form can be recovered from this approach. Set $\mathrm{d}X_t = \sigma\,\mathrm{d}W_t$, the non-normalized SSE is:

$$\mathrm{d}\tilde{\psi}_t = A\tilde{\psi}_t\,\mathrm{d}t + \sigma B\tilde{\psi}_t\,\mathrm{d}W_t \tag{6.5}$$

where W is a Wiener process, $\sigma > 0$ is its intensity. Note that in this setting with a single noise source, $\mathrm{d}W$ is a real scalar, so both B and A are operators with a matrix representation.

To ensure the martingale property, the differential of the squared wavefunction averaged over trajectories replicas must be null, $\mathbb{E}[\|\mathrm{d}\psi_t\|^2] = 0$. The noise components average to zero due to the properties of $\mathrm{d}W$, and to preserve the martingale property we have to set the operators in the time component to zero too:

$$A^\dagger + A + \sigma^2 B^\dagger B = 0 \tag{6.6}$$

We are not interested in the trivial solution when B is the null element: that would recover the Schrödinger equation for the system as if closed.

Remark 4 The operator A must contain the Hamiltonian component $(-iH/\hbar)$ to recover the Schrödinger equation in the limit case of nought noise.

Then, we set

$$A = -iH + C \tag{6.7}$$

where H is the Hamiltonian of the system of interest, and the renormalization term C is an unspecified renormalization operator term, leading to

$$A = -iH - \frac{1}{2}\sigma^2 B^\dagger B \tag{6.8}$$

without any constraints on the B operator. Then, the SSE obtained is:

$$\mathrm{d}\psi_t = (-iH - \frac{1}{2}\sigma^2 B^\dagger B)\psi_t \,\mathrm{d}t + \sigma B\psi_t \,\mathrm{d}W_t \tag{6.9}$$

which can be numerically implemented. We can compute now the differential of the outer product of the wavefunction with itself $\mathrm{d}(\psi\psi^\dagger)$, making use of Itō's product rule, obtaining the Quantum Stochastic Master Equation, not so useful as it is associated with the single unitary trajectory, and then take the expectation value to obtain the associated Quantum Master Equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho_t = -i\left[H, \rho_t\right] + \sigma^2\left(B^\dagger \rho_t B - \frac{1}{2}\left\{B^\dagger B, \rho_t\right\}\right) \tag{6.10}$$

where $\{\cdot,\cdot\}$ as usual is the anti-commutator. With time-independent σ and B operators, this is quite clearly a Lindblad form, with just one dissipation channel considered - that due to the single "noise bath". We therefore are able to identify the operator $B = L$ as the jump operator and $\gamma = \sigma^2$ as the relaxation rate. The method is easily generalized to multiple dissipation channels by the use of i.i.d. multidimensional baths each with its own intensity. Due to the absence of constraints on the jump operators L, which can also be non-Hermitian hence inducing asymmetric decaying, and we can recover all QMEs of Lindblad form.

Considering different stochastic differentials and stochastic processes that are not white, colored Gaussian and not-Gaussian, different memory effects can be introduced in the dynamics of the system, leading to QMEs not in Lindblad form QMEs, usually not closed, but correlated and with memory effects for which we have an unraveling that allows for numerical computation.

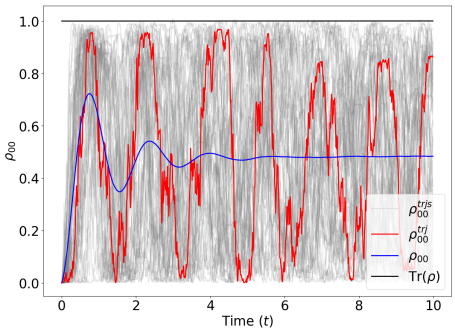


Figure 5: Stochastic unraveling of a simple Lindblad equation for a two-level system. For the quantity given by $(\psi(t))_0(\psi(t))_0^*$, a swarm of 50 different stochastic trajectories in grey lines in transparency and the example evolution of single trajectories, red line, are displayed. The mean time evolution of the element ρ_{00} (population) of the density matrix is in the blue line.

References

[1] K. Kraus, "Effects, and Operations: Fundamental Notions of Quantum Theory". Edited by K. Kraus, A. Böhm, J.D. Dollard, and W.H. Wootters (Springer Berlin, Heidelberg, Aug. 1983), p. 40, 10.1007/3- 540-12732-1.

[2] M.D. Choi, *Positive Linear Maps on C*-Algebras.* Can. J. Math XXIV, 520–529 (1972) 10.4153/CJM-1972-044-5.

[3] M.D. Choi, *Completely positive linear maps on complex matrices.* Linear Algebra and its Applications 10, 285–290 (1975) 10.1016/0024-3795(75)90075-0.

[4] D. Manzano, *A short introduction to the Lindblad master equation.* AIP Advances 10, 025106 (2020) 10.1063/1.5115323.

[5] H.-P. Breuer and F. Petruccione, "The Theory of Open Quantum Systems". 1st ed. (Oxford University Press, Oxford, Jan. 2002), 10.1093/acprof:oso/9780199213900.001.0001.

[6] D.A. Lidar, "Lecture Notes on the Theory of Open Quantum Systems". Los Angeles, Feb. 2019.

[7] Á. Rivas and S.F. Huelga, "Open Quantum Systems. An Introduction". 1st ed. (Springer Berlin, Heidelberg, Apr. 2011), 10.1007/978-3-642-23354-8.

[8] G. Lindblad, *Mathematical Physics On the Generators of Quantum Dynamical Semigroups.* Communications in Mathematical Physics 48, 119–130 (1976).

[9] V. Gorini, A. Kossakowski, and E.C. Sudarshan, *Completely positive dynamical semigroups of N-level systems.* Journal of Mathematical Physics 17, 821 (1976) 10.1063/1.522979.

[10] E.B. Davies, *Markovian master equations.* Communications in Mathematical Physics 39, 91–110 (1974) 10.1007/BF01608389.

[11] A. Barchielli and M. Gregoratti, "Quantum Trajectories and Measurements in Continuous Time: The Diffusive Case". Lect. Notes Phys. 782 (Springer, Berlin Heidelberg, 2009), 10.1007/978-3-642-01298-3.

Representations of Quivers over Rings: Merging Commutative and Non-Commutative Results

Enrico Sabatini (*)

Abstract. In the vast universe of representation theory there are two very separate and different worlds: commutative rings and finite dimensional (non-commutative) algebras. The problem of characterizing certain subcategories, like many other problems, has been solved in both fields. However, the main techniques used for one context are generally not transferable to the other. Recently, some authors have focused their interest on a special kind of algebras that partially merge the two fields. Here, the apparently different results have a surprising generalization and a unifying proof. We will give an overview of the two fields mentioned above, describe their main features and give an idea of what allows such characterizations. Finally, we will show how the generalization works with the aid of some examples.

A Bestiary of Algebraic Structures

Let us recall the definitions of some algebraic structures that will appear in the next:

- **Rings**: Triples $(R, +, \cdot)$ given by a set R and two associative operations such that $(R, +)$ is an abelian group, $(R, \cdot)$ is a monoid and the multiplication distributes over the sum.
  - When $(R, \cdot)$ is a commutative monoid, R is called commutative ring;
  - When $(R, \cdot)$ is an abelian group, R is a filed and will be denoted by $\mathbb{K}$.
- **R-Modules**: Triples $(M, +, \cdot)$ given by a set M, an associative operation such that $(M, +)$ is an abelian group and an operation $\cdot_R : R \times M \to M$ which distributes over both the sum in R and the sum in M and it is compatible with multiplication in R.
  - Example: if $R = \mathbb{K}$ is a field, M is a $\mathbb{K}$-vector space.

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: enrico.sabatini@phd.unipd.it . Seminar held on 19 December 2024.

- **R-Algebras**: Quadruples $(A, +, \cdot, \cdot_R)$ such that $(A, +, \cdot)$ is a ring and $(A, +, \cdot_R)$ is an R-module.
  - When $R = \mathbb{K}$ is a field and $\dim_{\mathbb{K}}(A) < \infty$, A is called finite-dimensional $\mathbb{K}$-algebra.
- **Categories**: Triples $\mathcal{C} = (\mathrm{Ob}(\mathcal{C}), \mathrm{Mor}(\mathcal{C}), \circ_{\mathcal{C}})$, given by a class of objects, a class of morphisms and an associative partial operation on morphisms such that for any $A \in \mathrm{Ob}(\mathcal{C})$ there is a morphism $\mathrm{id}_A \in \mathrm{Mor}(\mathcal{C})$ which is an identity for $\circ_{\mathcal{C}}$. Examples are:
  - $\mathcal{S}et$ - with sets as objects, functions as morphisms and composition as $\circ_{\mathcal{S}et}$;
  - $\mathcal{T}op$ - with topological spaces as objects and continuous functions morphisms;
  - $\mathcal{M}et$ - with metric spaces as objects and 1-Lipschitz functions as morphisms.

  In the following we will be mostly interested in the categories:
  - $\mathrm{Mod}(R)$ - with R-modules as objects and R-linear functions as morphisms;
  - $\mathcal{P}os$ - with partially ordered sets as objects and monotone functions as morphisms.
- **Subcategories**: A category $\mathcal{S}$ is a subcategory of a category $\mathcal{C}$ if $\mathrm{Ob}(\mathcal{S}) \subseteq \mathrm{Ob}(\mathcal{C})$, $\mathrm{Mor}(\mathcal{S}) \subseteq \mathrm{Mor}(\mathcal{C})$ and $g \circ_{\mathcal{S}} f = g \circ_{\mathcal{C}} f$ for any two composable morphisms $f, g \in \mathrm{Mor}(\mathcal{S})$.
  - Since every metric induce a topology such that 1-Lipschitz functions are continuous, there is a chain of subcategories $\mathcal{M}et \subseteq \mathcal{T}op \subseteq \mathcal{S}et$.

1 Introduction

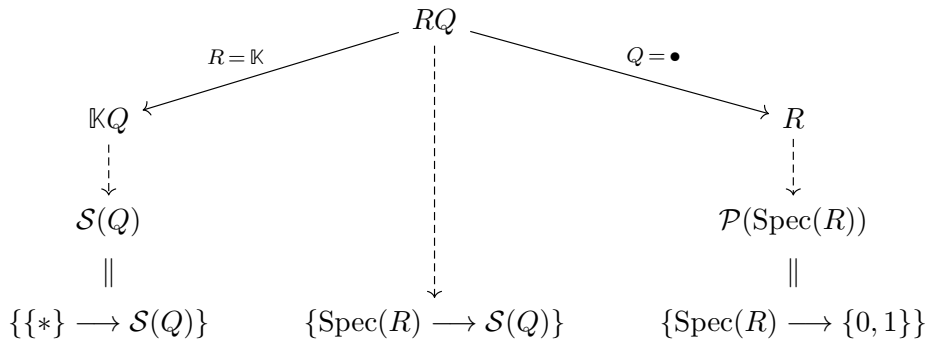
The main interest of representation theory is the study of the category of modules $\mathrm{Mod}(R)$ of a given ring R, and many problems in this area can be approached by studying suitable subcategories. Among all the rings there are two distinguished classes of particular interest, which are very different from each other: finite dimensional hereditary algebras and commutative noetherian rings. The aim of this paper is to give a rough idea of how to classify a certain class of subcategories in the two contexts.

In particular, any finite dimensional hereditary algebra (over an algebraically closed field) can be described as the path algebra $\mathbb{K}Q$ of a quiver Q, and in this case some classes of subcategories are in bijection with a poset $\mathcal{S}(Q)$ depending on the quiver (for example, the Cambrian lattice or the lattice of non-crossing partitions); on the other hand, for a commutative noetherian ring R, the classifications can be obtained via the prime spectrum $\mathrm{Spec}(R)$ (for example, by considering its power set $\mathcal{P}(\mathrm{Spec}(R))$), which is the poset formed by the prime ideals of the ring ordered by inclusion.

In the last section, we will consider a third class of rings which includes the two previous classes: the noetherian path algebras RQ; and we will show, through some examples and without giving the details, how some of the classifications, despite the very different tools

and techniques used in the two settings, find a unifying description. In particular, in this new context, the classifications are given by the poset of maps from $\mathcal{S}(Q)$ to $\mathrm{Spec}(R)$.

The following diagram schematically summarizes what we have explained above. In fact, a noetherian path algebra RQ is a finite dimensional hereditary algebra if $R = \mathbb{K}$ is a field and, in this case, $\mathrm{Spec}(R) = \{*\}$ consists of one point; or, it is a commutative noetherian ring if $Q = \bullet$ has one vertex and no arrows, and in this case the poset $\mathcal{S}(Q) = \{0, 1\}$ consists of two points (with the obvious order).



2 Finite Dimensional Hereditary Algebras

A quiver $Q = (Q_0, Q_1, s, t)$ is a directed graph, where Q_0 is the set of vertices, Q_1 is the set of arrows and $s, t : Q_1 \to Q_0$ are two functions assigning to each arrow its source and its target.

Example 2.1 The running examples for us will be:

$A_1 = \overset{1}{\bullet}$ - the quiver with one vertex and no arrows;

$A_2 = \overset{1}{\bullet} \xrightarrow{\alpha} \overset{2}{\bullet}$ - the quiver with two vertices and one arrow s. t. $s(\alpha) = 1$ and $t(\alpha) = 2$;

$L = \overset{1}{\bullet}\circlearrowright^{\alpha}$ - the quiver with one vertex and one arrow such that $s(\alpha) = t(\alpha) = 1$.

2.1 Path Algebras

A path in Q of length n is a sequence of n arrows $\omega = \alpha_n \dots \alpha_1$ such that $t(\alpha_i) = s(\alpha_{i+1})$ for any $i = 1, \dots, n-1$, by abuse of notation we will write $s(\omega) = s(\alpha_1)$ and $t(\omega) = t(\alpha_n)$. Moreover, for any vertex $i \in Q_0$ we assume that there exists a length-zero path ε_i with source and target i and we call it the lazy path at i (not to be confused with a loop based in i, which is a path of length 1). So we define the set of paths of Q to be

$$Paths(Q) := \{\alpha_n \dots \alpha_1 \mid n \in \mathbb{Z},\, \alpha_i \in Q_1,\, t(\alpha_i) = s(\alpha_{i+1})\} \cup \{\varepsilon_i \mid i \in Q_0\}$$

Given a field $\mathbb{K}$ and a quiver Q, we call path algebra of Q over $\mathbb{K}$ the (non-unitary) $\mathbb{K}$-algebra whose vector space and (non-unitary) ring structures are given respectively by

$$\mathbb{K}Q := \mathrm{span}_{\mathbb{K}}\langle \omega \mid \omega \in Paths(Q)\rangle \text{ with multiplication } \omega \cdot \omega' := \begin{cases} \omega\omega' & \text{if } t(\omega') = s(\omega) \\ 0 & \text{otherwise} \end{cases}$$

As stressed out by the adjectives in the brackets, this definition does not always give an algebra as defined in the "Bestiary", but the following holds.

Proposition 2.2 ([ASS06, Lemma II.1.4, Theorem VII.1.7])

(a) *The path algebra $\mathbb{K}Q$ has a unit element (i.e. an identity for the multiplication) if and only if the quiver Q has finitely many vertices and in this case $1_{\mathbb{K}Q} = \sum_{i \in Q_0} \varepsilon_i$;*

(b) *The path algebra $\mathbb{K}Q$ is finite dimensional if and only if the quiver Q has finitely many arrows and no oriented cycles (i.e. paths ω with $s(\omega) = t(\omega)$);*

(c) *If the condition above are satisfied, $\mathbb{K}Q$ is a finite-dimensional hereditary algebra (i.e. any submodule of a projective module is projective); moreover, if $\mathbb{K}$ is algebraically closed, any finite-dimensional hereditary $\mathbb{K}$-algebra A is isomorphic to a path algebra $\mathbb{K}Q_A$.*

Example 2.3

$\mathbb{K}A_1 := \mathrm{span}_{\mathbb{K}}\langle \varepsilon_1 \rangle \cong \mathbb{K}$ is a 1-dimensional $\mathbb{K}$-algebra;

$\mathbb{K}A_2 := \mathrm{span}_{\mathbb{K}}\langle \varepsilon_1, \varepsilon_2, \alpha \mid \alpha\varepsilon_1 = \varepsilon_2\alpha = \alpha\rangle \cong \begin{bmatrix} \mathbb{K} & 0 \\ \mathbb{K} & \mathbb{K} \end{bmatrix}$ is a 3-dimensional $\mathbb{K}$-algebra;

$\mathbb{K}L := \mathrm{span}_{\mathbb{K}}\langle \varepsilon_1, \alpha, \alpha^2, \ldots\rangle \cong \mathbb{K}[x]$ is an infinite-dimensional $\mathbb{K}$-algebra.

2.2 Representations of Quivers

Given a field $\mathbb{K}$ and a quiver Q, a representation of Q over $\mathbb{K}$ is a tuple $V = (V_i, V_\alpha)_{i \in Q_0, \alpha \in Q_1}$ where V_i is a $\mathbb{K}$-vector space for any vertex $i \in Q_0$ and $V_\alpha : V_i \to V_j$ is an $\mathbb{K}$-linear map for any arrow $\alpha : i \to j \in Q_1$. A morphism of representations $f : V \to W$ is a collection of $\mathbb{K}$-linear maps $(f_i : V_i \to W_i)_{i \in Q_0}$ such that for any $\alpha \in Q_1$ the following diagram commute

$$\begin{array}{ccc} V_{s(\alpha)} & \xrightarrow{V_\alpha} & V_{t(\alpha)} \\ \downarrow{\scriptstyle f_{s(\alpha)}} & & \downarrow{\scriptstyle f_{t(\alpha)}} \\ W_{s(\alpha)} & \xrightarrow{W_\alpha} & W_{t(\alpha)} \end{array}$$

Example 2.4

$A_1 = \overset{1}{\bullet}$ - representations are $\mathbb{K}$-vector spaces and morphisms $\mathbb{K}$-linear maps between them;

$A_2 = \overset{1}{\bullet} \xrightarrow{\alpha} \overset{2}{\bullet}$ - representations are given by $\mathbb{K}$-linear maps between two vector spaces and examples of morphisms are:

$$\begin{array}{ccc} \mathbb{K} & \xrightarrow{1} & \mathbb{K} \\ \downarrow{\scriptstyle 1} & & \downarrow{\scriptstyle 0} \\ \mathbb{K} & \xrightarrow{0} & 0 \end{array} \qquad \begin{array}{ccc} \mathbb{K} & \xrightarrow{1} & \mathbb{K} \\ \downarrow{\scriptstyle 1} & & \downarrow{\scriptstyle 1} \\ \mathbb{K} & \xrightarrow{1} & \mathbb{K} \end{array} \qquad \begin{array}{ccc} 0 & \xrightarrow{0} & \mathbb{K} \\ \downarrow{\scriptstyle 0} & & \downarrow{\scriptstyle 1} \\ \mathbb{K} & \xrightarrow{1} & \mathbb{K} \end{array}$$

$L = \overset{1}{\bullet} \circlearrowright^{\alpha}$ - representations are $\mathbb{K}$-vector spaces together with a $\mathbb{K}$-linear endomorphism.

Notice that representations of a quiver Q over a field $\mathbb{K}$ and morphisms between them form a category, we will denote it by $\mathrm{Rep}_{\mathbb{K}}(Q)$.

Theorem 2.5 ([ASS06, Corollary III.1.7]) *Given a quiver Q and a field $\mathbb{K}$, there is an equivalence of categories* $\mathrm{Mod}(\mathbb{K}Q) \cong \mathrm{Rep}_{\mathbb{K}}(Q)$.

It turns out that the category of modules $\mathrm{Mod}(\mathbb{K}Q)$ is not only very easy to visualize, since it is equivalent to $\mathrm{Rep}_{\mathbb{K}}(Q)$, but its subcategory of finite dimensional modules $\mathrm{mod}(\mathbb{K}Q)$ is also very easy to study and understand, since it can be decomposed into some building blocks. Indeed, there are some distinguished objects and morphisms which allow to recover the whole category.

Definition 2.6

(a) A representation V is said to be indecomposable if it is nonzero and has no nontrivial direct sum decomposition.

(b) A morphism $f : V \to W$ is said to be irreducible if it is neither a section nor a retraction and has no nontrivial factorization.

Proposition 2.7 ([ASS06, Theorem I.4.10, Lemma IV.1.6])

(a) *Every finite dimensional representation of a quiver decomposes uniquely as a direct sum of indecomposable representations;*

(b) *Every morphism between finite dimensional representations can be built from irreducible morphisms (and isomorphisms) by forming compositions, linear combinations and matrices.*

Remark 2.8 In very nice situations, such as when the quiver Q is a Dynkin diagram (see [ASS06, Section VII.2] for a complete list), indecomposable modules and irreducible morphisms of $\mathrm{Mod}(\mathbb{K}Q)$ are independent of the field and depend only on Q; for Dynkin diagrams this is known as Gabriel's theorem [ASS06, Theorem VII.5.10]). Thus, it is now conceivable why many classifications of subcategories of $\mathrm{Mod}(\mathbb{K}Q)$ also depend only on the quiver.

In particular, some classes of subcategories are in bijection with a lattice obtained from Q, we will mention later: the power set of Q_0, denoted by $\mathbf{P}(Q)$, the Cambrian lattice $\mathbf{C}(Q)$ and the lattice of non-crossing partitions $\mathbf{Nc}(Q)$.

3 Commutative Noetherian Rings

Example 3.1 The running examples for us will be:

$R = \mathbb{K}$ a field;

$R = \mathbb{Z}$ the ring of integers;

$R = \mathbb{C}[[x]]$ the ring of formal power series over the complex numbers.

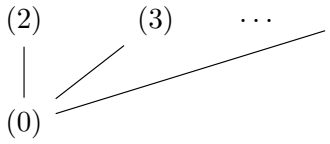
3.1 The Prime Spectrum

Given a commutative ring $(R, +, \cdot)$, we say that $\mathfrak{p} \subseteq R$ is an ideal if $(\mathfrak{p}, +)$ is a subgroup of $(R, +)$ and it is closed under multiplication by elements of R, i.e. $R \cdot \mathfrak{p} \subseteq \mathfrak{p}$. Moreover, an ideal $\mathfrak{p}$ is called prime if $R \setminus \mathfrak{p}$ is closed under multiplication, i.e. for any $s, t \in R \setminus \mathfrak{p}$ the element $s \cdot t$ is again in $R \setminus \mathfrak{p}$. We define the prime spectrum of R as

$$\mathrm{Spec}(R) = \{\mathfrak{p} \subseteq R \mid \mathfrak{p} \text{ is a prime ideal}\}$$

Example 3.2

$\mathrm{Spec}(\mathbb{K}) = \{*\}$ is equal to one point corresponding to the ideal (0);

$\mathrm{Spec}(\mathbb{Z}) =$ 

and $\mathrm{Spec}(\mathbb{C}[[x]]) =$ 

.

Remark 3.3 Recall that $(\mathrm{Spec}(R), \subseteq)$ is a poset. A subset $V \subseteq \mathrm{Spec}(R)$ is called specialization closed (or upper subset) if for any $\mathfrak{p} \in V$ and $\mathfrak{q} \in \mathrm{Spec}(R)$ such that $\mathfrak{p} \subseteq \mathfrak{q}$ it holds that $\mathfrak{q} \in V$. Denoting by $\mathcal{V}(\mathrm{Spec}(R))$ the set of specialization closed subsets of $\mathrm{Spec}(R)$, notice that

$$\mathcal{V}(\mathrm{Spec}(R)) = \{\text{Monotone maps } \mathrm{Spec}(R) \longrightarrow \{0, 1\}\}$$

3.2 Localization at a Prime

Given a prime ideal $\mathfrak{p} \in \operatorname{Spec}(R)$, we define the localization of R at $\mathfrak{p}$ as

$$R_{\mathfrak{p}} = \left\{ \frac{r}{s} \mid r \in R, s \in R \setminus \mathfrak{p} \right\} \Big/ \sim$$

where $\frac{r_1}{s_1} \sim \frac{r_2}{s_2}$ if there is $t \in R \setminus \mathfrak{p}$ such that $t(r_1 s_2 - r_2 s_1) = 0$, i.e. $\frac{r}{s} \sim \frac{tr}{ts}$ for any $t \in R \setminus \mathfrak{p}$.

Analogously, for an R-module M, we define the localization of M at $\mathfrak{p}$ as:

$$M_{\mathfrak{p}} = \left\{ \frac{m}{s} \mid m \in M, s \in R \setminus \mathfrak{p} \right\} \Big/ \sim \text{ where } \frac{m}{s} \sim \frac{tm}{ts} \text{ for any } t \in R \setminus \mathfrak{p}.$$

Example 3.4 Let $R = \mathbb{Z}$, $\mathfrak{p} = (0)$ and $M = \mathbb{Z}/n\mathbb{Z} = \{\overline{0}, \overline{1}, \dots, \overline{n-1}\}$ the group of integers modulo n. Then:

$\mathbb{Z}_{(0)} = \mathbb{Q}$ - it is the field of rational numbers;

$M_{(0)} = 0$ - indeed, each of its elements is of the form $\frac{\overline{m}}{s}$ with $\overline{m} \in M$ and $s \in \mathbb{Z} \setminus (0)$ and

$$\frac{\overline{m}}{s} \sim \frac{n\overline{m}}{ns} = \frac{\overline{nm}}{ns} = \frac{\overline{0}}{ns} = 0$$

This last example show us how localization can annihilate modules, thus it makes sense to introduce the following definition.

Definition 3.5 Given an R-module M, we call support of M over R the set

$$\operatorname{Supp}_R(M) = \{\mathfrak{p} \in \operatorname{Spec}(R) \mid M_{\mathfrak{p}} \neq 0\}$$

Analogously, for a subcategory $\mathcal{C} \subseteq \operatorname{Mod}(R)$, we define

$$\operatorname{Supp}_R(\mathcal{C}) = \bigcup_{M \in \operatorname{Ob}(\mathcal{C})} \operatorname{Supp}_R(M)$$

It is now imaginable how many classifications of subcategories of $\operatorname{Mod}(R)$ depend on the prime spectrum $\operatorname{Spec}(R)$. In particular, some classes of subcategories are in bijection with the power set $\mathcal{P}(\operatorname{Spec}(R))$ or with the set of specialization closed subsets $\mathcal{V}(\operatorname{Spec}(R))$.

4 Noetherian path algebras - Examples

The definitions of path algebras and representations of quivers, introduced in Section 2, remain unchanged if we consider them over a commutative noetherian ring R instead of a field $\mathbb{K}$. Analogously to Theorem 2.5, it is still valid that $\mathrm{Mod}(RQ) \cong \mathrm{Rep}_R(Q)$. Notice that when the ring $R = \mathbb{K}$ is a field, $RQ = \mathbb{K}Q$ is a finite dimensional hereditary algebra, and when the quiver $Q = A_1$ is just a vertex with no arrows, $RQ = R$ is a commutative noetherian ring.

The following table summarizes the examples of how some classifications, obtained separately in the two different contexts, have found a merged description in this new setting. We will not define either the subcategories or the lattices involved, as this is not relevant to the purpose of this paper. Instead, we refer the interested reader to the cited papers and the references therein.

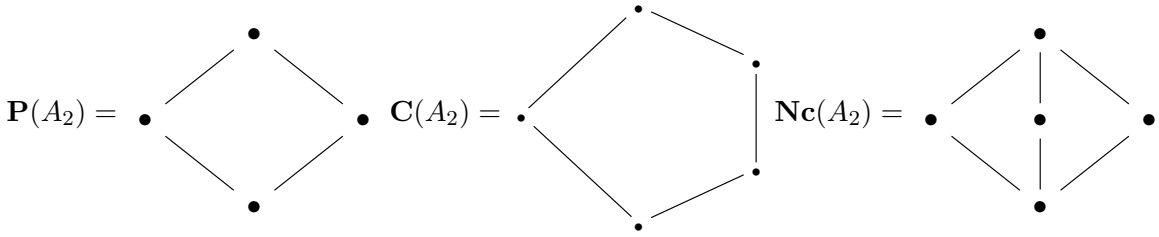
Subcategory	$\mathbb{K}Q$	R	RQ	Reference
Localizing	$\mathbf{Nc}(Q)$	$\mathcal{P}(\mathrm{Spec}(R))$	$\mathrm{Spec}(R) \xrightarrow{\mathcal{S}et} \mathbf{Nc}(Q)$	[AS16]
Smashing	$\mathbf{Nc}(Q)$	$\mathcal{V}(\mathrm{Spec}(R))$	$\mathrm{Spec}(R) \xrightarrow{\mathcal{P}os} \mathbf{Nc}(Q)$	
Serre	$\mathbf{P}(Q)$	$\mathcal{V}(\mathrm{Spec}(R))$	$\mathrm{Spec}(R) \xrightarrow{\mathcal{P}os} \mathbf{P}(Q)$	[IK24]
Torsion	$\mathbf{C}(Q)$	$\mathcal{V}(\mathrm{Spec}(R))$	$\mathrm{Spec}(R) \xrightarrow{\mathcal{P}os} \mathbf{C}(Q)$	
Wide	$\mathbf{Nc}(Q)$	$\mathcal{V}(\mathrm{Spec}(R))$	$\mathrm{Spec}(R) \xrightarrow{\mathcal{P}os} \mathbf{Nc}(Q)$	[Sab25]
t-structure	$\mathrm{Filt}(\mathbf{Nc}(Q))$	$\mathrm{Filt}(\mathcal{V}(\mathrm{Spec}(R)))$	$\mathrm{Spec}(R) \xrightarrow{\mathcal{P}os} \mathrm{Filt}(\mathbf{Nc}(Q))$	[Sab25]

Note that we have divided the table into three sections, each representing a different group of examples. In fact, the first group - which actually involves subcategories of the derived category - presents two classes of subcategories that are classified by the same object over $\mathbb{K}Q$ and by two different objects over R, while the opposite happens for the second group.

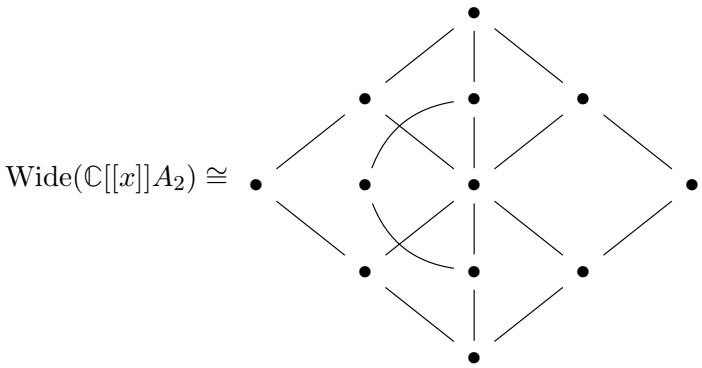
The first situation occurs because the prime spectrum of a field $\mathrm{Spec}(\mathbb{K})$ is a point and so any function from it to the poset $\mathbf{Nc}(Q)$ is automatically monotone; while the second situation occurs because for the quiver $Q = A_1$ we have that all three lattices coincide

$$\mathbf{P}(A_1) = \mathbf{C}(A_1) = \mathbf{Nc}(A_1) = \quad \begin{array}{c} \bullet_1 \\ | \\ \bullet_0 \end{array}$$

Apart from these limit cases, the examples suddenly become more interesting. Indeed, already for $Q = A_2$ we have that the three lattices are all different:



and, for example, for the noetherian path algebra $\mathbb{C}[[x]]A_2$ there is a poset isomorphism



References

[AS16] B. Antieau and G. Stevenson, *Derived categories of representations of small categories over commutative noetherian rings.* Pacific Journal of Mathematics, 283 (2016).

[ASS06] I. Assem, D. Simson, and A. Skowroński, "Elements of the Representation Theory of Associative Algebras: Techniques of representation theory". Cambridge University Press, 2006.

[IK24] O. Iyama and Y. Kimura, *Classifying subcategories of modules over noetherian algebras.* Advances in Mathematics, 446 (2024).

[Sab25] E. Sabatini, *Telescope conjecture for t-structures over noetherian path algebras.* ArXiv preprint, `arXiv:2505.20803` (2025).

Deep Unfolding: Bridging Optimization and Neural Network Interpretability

Erik Chinellato (*)

Abstract. Deep neural networks (DNNs) have revolutionized numerous fields due to their powerful ability to learn complex representations. However, their black-box nature and lack of interpretability in architecture and weight design remain significant challenges. After an introductory segment on DNNs and backpropagation learning, this seminar introduces the Deep Unfolding method as a promising alternative, bridging the gap between data-driven learning and model-based optimization. By unrolling iterative optimization algorithms into structured neural network architectures, Deep Unfolding provides a principled approach to network design, enabling interpretability and theoretical insights into their operation. We will explore how this method leverages domain knowledge, achieves faster convergence, and enhances performance in resource-constrained scenarios. The session will highlight many wide-ranging practical applications of Deep Unfolding, covering audio source separation and recognition, image denoising and state estimation.

1 Introduction

The focal point of this brief presentation is the *Deep Unfolding*, which we present in the context of bilevel optimization, where it has been widely and successfully developed for image restoration [2] and speech enhancement [6] as a way to avoid both hessian inversion and loss gradient approximation: while a simple gradient descent scheme is unfolded to solve the inner optimization, a backpropagation procedure is used to solve the outer optimization. The unfolding can be further generalized by the so called *untying*, first proposed in [6], a procedure in which some of the constraints on the parameters involved in the unfolded iterations are lifted, see Sec. 3 and 3.1. The resulting trained, unfolded network can be naturally interpreted as a parameter-optimized algorithm, effectively overcoming the lack of interpretability that characterizes most conventional neural networks. Moreover, it is often reported that in comparison with generic DNNs, unfolded networks have fewer parameters, therefore requiring less training data and computational resources; this makes them suitable for embedded computing. In Sec. 4-6 we present three paradigmatic applications of Deep Unfolding in the context of audio source separation, image denoising

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: chinella@math.unipd.it. Seminar held on 9 January 2025.

and state estimation.

2 Deep Neural Networks and interpretability

Deep Neural Networks (DNNs) are an effective tool to learn complex input-output relations. They are often used as an alternative to designing mathematical models of complex systems that can interact with the environment and respond to external excitations. Indeed, DNNs essentially consist in abstract parametric functions mapping an input vector $y = \hat{x}^{(0)} \in \mathbb{R}^n$ to an output $\hat{x}^{(K)} \in \mathbb{R}^m$ that closely approximates the response $S_y \in \mathbb{R}^m$ of the target system to the same input. The parameters of these functions are tuned (learned) to minimize the approximation error $e_y = \hat{x}^{(K)} - S_y$.

Given a collection of matrices $A^{(k)} \in \mathcal{M}_{m_k \times n_k}(\mathbb{R})$ for $k = 0, \dots, K-1$ satisfying $m_k = n_{k+1}$ $\forall k = 0, \dots, K-2$, $n_0 = n$, $m_{K-1} = m$ and a nonlinear activation function $\sigma : \mathbb{R} \longrightarrow \mathbb{R}$ to be applied elementwise to vectors, the simplest DNNs one can construct is the following K-layer, fully-connected network:

$$y = \hat{x}^{(0)} \xrightarrow{\sigma\left(A^{(0)}\,\cdot\right)} \hat{x}^{(1)} \xrightarrow{\sigma\left(A^{(1)}\,\cdot\right)} \cdots \xrightarrow{\sigma\left(A^{(K-2)}\,\cdot\right)} \hat{x}^{(K-1)} \xrightarrow{\sigma\left(A^{(K-1)}\,\cdot\right)} \hat{x}^{(K)} \dashleftarrow\dashrightarrow S_y$$

One such network is then trained by minimizing a chosen loss function $\mathcal{F}(x)$ applied to the output at the last layer $\hat{x}^{(K)}$, e.g. the 2-norm residue $\mathcal{F}(x) = \frac{1}{2}\|x - S_y\|_2^2$, and the parameters of the network are updated using simple gradient descent:

$$A_{\bar{i}\bar{j}}^{(k)} \Leftarrow A_{\bar{i}\bar{j}}^{(k)} - \mu \frac{\partial \mathcal{F}(\hat{x}^{(K)})}{\partial A_{\bar{i}\bar{j}}^{(k)}} \quad \forall \bar{i}, \bar{j}, \forall k = 0, \dots, K-1$$

In order to reduce computations, we use *backpropagation*, an algorithm that exploits the chain rule to drastically cut the number of operations required. In particular, assuming to have already computed $\frac{\partial \mathcal{F}}{\partial \hat{x}_{i_{k+2}}^{(k+2)}}$ $\forall i_{k+2}$, one obtains:

$$\begin{cases} \dfrac{\partial \mathcal{F}}{\partial \hat{x}_{i_{k+1}}^{(k+1)}} = \displaystyle\sum_{i_{k+2}} \frac{\partial \mathcal{F}}{\partial \hat{x}_{i_{k+2}}^{(k+2)}} \frac{\partial \hat{x}_{i_{k+2}}^{(k+2)}}{\partial \hat{x}_{i_{k+1}}^{(k+1)}} & \forall i_{k+1} \\ \dfrac{\partial \mathcal{F}(\hat{x}^{(K)})}{\partial A_{\bar{i}\bar{j}}^{(k)}} = \displaystyle\sum_{i_{k+1}} \frac{\partial \mathcal{F}}{\partial \hat{x}_{i_{k+1}}^{(k+1)}} \frac{\partial \hat{x}_{i_{k+1}}^{(k+1)}}{\partial A_{\bar{i}\bar{j}}^{(k)}} \end{cases}$$

Traditional DNNs struggle with *interpretability*: while on the one hand these networks can be trained to associate any input to its desired output (possibly by adjusting their depth K and width m_k), on the other hand their generic architecture and non-informativeness of the learned parameters make it impossible to extrapolate information or gain insight on the target complex system that generated those input-output pairs. For this reason, DNNs are often called *black-box* models. The following sections introduce the mathematical framework used by the Deep Unfolding to generate interpretable deep networks.

3 Bilevel optimization

Let us consider a parametric optimization problem depending on a family of parameters $\theta \in \Omega_\Theta$ and an *inner* objective function $\mathcal{F}_\theta^{in} : \Omega_X \times \Omega_Y \to \mathbb{R}$:

$$\begin{aligned} &\min \quad \mathcal{F}_\theta^{in}(x, y) \\ &\text{s.t.} \quad x \in \Omega_X \end{aligned} \tag{1}$$

where x is the quantity of interest (or state) to be optimized based on some given observation y. Denoting $\hat{x}(\theta; y)$ its minimizer:

$$\hat{x}(\theta; y) = \operatorname*{arg\,min}_{x \in \Omega_X} \mathcal{F}_\theta^{in}(x, y) \tag{2}$$

bilevel optimization frameworks, in their most simple form, strive to minimize an *outer* objective function $\mathcal{F}^{out} : \Omega_X \times \Omega_Y \to \mathbb{R}$, taking $\hat{x}(\theta; y)$ as an argument, with respect to the inner parameters θ:

$$\begin{aligned} &\min \quad \mathcal{F}^{out}(\hat{x}(\theta; y), y) \\ &\text{s.t.} \quad \theta \in \Omega_\Theta \end{aligned} \tag{3}$$

Such problems naturally arise when optimizing partially unknown models, in our notation associated to $\mathcal{F}_\theta^{in}$ in (1). The parameters θ can then be recovered using the following two approaches, leading to the choice of an appropriate outer objective $\mathcal{F}^{out}$ in (3):

- *Imposition of a priori optimality requirements.* By exploiting the unknown parameters θ, we can endow the minimizers $\hat{x}(\theta; y)$ with some desirable property chosen a priori, independent of the observations y. Such properties are usually enforced by a regularizer $\mathcal{R} : \Omega_X \to \mathbb{R}$, and depend on the specific task at hand. Common choices for $\mathcal{R}$ include the relaxed ℓ_1 sparsity constraint $\mathcal{R}(\,\cdot\,) = \|\cdot\|_1$ or the second order finite differences smoothness constraint $\mathcal{R}(\,\cdot\,) = \|D \cdot\|_2^2$, to name a few;
- *Learned from available data.* When training data $\mathcal{T} = \{y_i\}_{i=1,\dots,I}$ is available, as well as a map $\mathcal{P} : \Omega_X \times \Omega_Y \to \mathbb{R}$ encoding the cost of pairing the quantities of interest and observations together, we can search for the optimal parameters minimizing the pairing cost between $\hat{x}(\theta; y_i)$ and its associated observation y_i, for all available y_i in the training dataset. In this setting, a common choice for $\mathcal{P}$ is the simple $2-$norm squared $\mathcal{P}(\,\cdot\,,\,\cdot\,) = \|\,\cdot - \mathcal{S}(\,\cdot\,)\|_2^2$, where $\mathcal{S} : \mathcal{T} \subset \Omega_Y \to \Omega_X$ extrapolates some target quantity from the training observation data.

In complete generality, one can construct $\mathcal{F}^{out}$ by letting:

$$\mathcal{F}^{out}(x, y) = \mathcal{P}(x, y) + \mu \mathcal{R}(x) \tag{4}$$

for some penalty parameter $\mu > 0$, and minimize the global loss:

$$\mathcal{L} = \sum_{i=1}^{I} \mathcal{F}^{out}(\hat{x}(\theta; y_i), y_i) \tag{5}$$

3.1 Deep unfolding

We now introduce the deep unfolding concept within the previously presented bilevel optimization framework. We remark that this is not the only framework where one can apply this technique, for the same idea can be employed to a much broader class of optimization problems, whenever an iterative update scheme is given. Nevertheless, bilevel optimization offers a solid foundation upon which the deep unfolding has been widely and successfully developed, especially since it allows for designing highly interpretable neural architectures.

Assuming to have at our disposal an iterative update map $f_\theta^{in} : \Omega_X \times \Omega_Y \to \Omega_X$ for the inner problem, providing an estimate $\hat{x}^{(K)} = \hat{x}^{(K)}(\theta; y)$ of the minimizer $\hat{x}(\theta; y)$ after K updates starting from an initial guess $\hat{x}^{(0)}$:

$$\hat{x}^{(k)} = f_\theta^{in}(\hat{x}^{(k-1)}, y) \quad \forall k = 1, \dots, K \tag{6}$$

the deep unfolding technique regards the above iterations as a sequence of layers in a neural network architecture with parameters θ *tied across layers* and activation function f_θ^{in}, as shown in Figure 1:

$$\hat{x}^{(0)} \xrightarrow{f_\theta^{in}} \hat{x}^{(1)} \xrightarrow{f_\theta^{in}} \dots \xrightarrow{f_\theta^{in}} \hat{x}^{(K-1)} \xrightarrow{f_\theta^{in}} \hat{x}^{(K)} \dashrightarrow \mathcal{F}^{out}(\hat{x}^{(K)}, y)$$

Figure 1: Unfolded iterations of (6) into a $K-$layer, tied parameters architecture trainable end-to-end with the loss function $\mathcal{F}^{out}$.

The resulting scheme can then be trained end-to-end using $\mathcal{F}^{out}$ as the loss function.

The main feature differentiating deep unfolded networks from traditional deep networks is the use of problem-specific activation functions, rather than general-purpose ones, such as ReLU, sigmoid, etc. Indeed, by employing f_θ^{in} as the activation function, the resulting architecture will be characterized by a highly interpretable forward-run, closely mimicking the process required to obtain an optimal solution to the associated bilevel optimization inner problem (1). The unfolded network then expands the modelling capabilities of the outer problem (3) by relaxing the requirement of knowing the exact optimal solution $\hat{x}(\theta; y)$, replacing it instead with a fixed-iteration estimate $\hat{x}^{(K)}$, and allowing the use of backpropagation learning for the parameters θ. We remark, however, that in cases where $\hat{x}(\theta; y)$ admits a closed-form representation as a function of θ, one can directly optimize the parameters without unfolding the iterates of an update map (see Example 1 and [10]). As a consequence, the unfolding technique is particularly attractive when the explicit form of $\hat{x}(\theta; y)$ is not known.

The modelling flexibility of unfolded deep networks even allows to replace or complement the inner update map f_θ^{in} with other learnable extensions (e.g. RNN-like modules for inter-layer correlation enforcing [7, 8], effectively trading some interpretability of the resulting framework in favor of efficiency or robustness. Indeed, this can be exploited to alleviate the computational burden of high-cost, non-smooth or high-dimensional operators [11] typical of model-based approaches.

Moreover, the fixed depth of the network K, when seen as the index at which we truncate the iterative process (6), offers yet another powerful modelling opportunity to deep unfolded networks: indeed, the learnable parameters can be leveraged to achieve convergence after a fixed and predetermined number of updates K. This idea has been successfully employed to speed-up notoriously slow-converging iterative schemes, as we will briefly discuss in Section 5.2.

A further extension to deep unfolded architectures can be obtained by *untying* the parameters θ into a layer-specific collection $\{\theta^{(k)}\}_{k=0,\dots,K-1}$. The rationale behind the untying is to allow the resulting network, shown in Figure 2, to embody a more complex range of inference functions [6]:

$$\hat{x}^{(0)} \xrightarrow{f^{in}_{\theta^{(0)}}} \hat{x}^{(1)} \xrightarrow{f^{in}_{\theta^{(1)}}} \dots \xrightarrow{f^{in}_{\theta^{(K-2)}}} \hat{x}^{(K-1)} \xrightarrow{f^{in}_{\theta^{(K-1)}}} \hat{x}^{(K)} \dashrightarrow \mathcal{F}^{\,out}(\hat{x}^{(K)}, y)$$

Figure 2: Unfolded iterations of (6) into a $K-$layer, untied parameters architecture trainable end-to-end with the loss function $\mathcal{F}^{\,out}$.

4 Nonnegative matrix factorization

Given a nonnegative matrix $X \in \mathcal{M}_{N\times M}(\mathbb{R}^+)$, a factorization rank R and an error measure $D : \mathcal{M}_{N\times M}(\mathbb{R}) \times \mathcal{M}_{N\times M}(\mathbb{R}) \to \mathbb{R}$ between two matrices, the problem of computing a Nonnegative Matrix Factorization (NMF) of X consists in solving the following constrained, nonlinear optimization problem:

$$\begin{aligned} &\min && D(X, WH) \\ &\text{s.t.} && W \in \mathcal{M}_{N\times R}(\mathbb{R}^+) \\ &&& H \in \mathcal{M}_{R\times M}(\mathbb{R}^+) \end{aligned} \tag{7}$$

Essentially, in (7) the observed matrix X is approximately reconstructed, with respect to the error measure D, as the product of two nonnegative matrices W and H. The former plays the role of a dictionary whose columns are used to reconstruct X given the coefficients in the latter. There is extensive literature regarding this constrained factorization [4]: here we will recall only a few fundamental aspects of NMF that will be used in later sections.

Problem (7) is in general a nonlinear, non-convex, constrained optimization problem over two distinct sets of variables, W and H. For this reason, most iterative algorithms used to tackle this problem often guarantee convergence results for a very specific class of error measures D. One such class is the $\beta-$divergences, a $\beta-$parametric family of error measures acting componentwise on two input nonnegative matrices:

$$D_\beta(\,\cdot\,,\,\cdot\,) = \sum_{n=1}^{N}\sum_{m=1}^{M} d_\beta(\,(\,\cdot\,)_{nm}, (\,\cdot\,)_{nm}\,)$$

where:

$$d_\beta(z,y) = \begin{cases} \frac{z}{y} - \log\frac{z}{y} - 1 & \text{if } \beta = 0, \\ z\log\frac{z}{y} - z + y & \text{if } \beta = 1, \\ \frac{1}{\beta(\beta-1)}(z^\beta + (\beta-1)y^\beta - \beta z y^{\beta-1}) & \text{if } \beta \neq 0, 1. \end{cases} \tag{8}$$

As mentioned, (7) has two distinct sets of variables, which can be conveniently exploited to devise 2-Block Coordinate Descent (2-BCD) schemes. In these methods, each set of variables is cyclically updated while keeping the other fixed. For $\beta-$divergences with $\beta \in [1, 2]$ this approach is particularly effective since D_β is convex in its second argument, aiding the optimization of the non-fixed set of variables. Nevertheless, a common technique to produce iterative updates within the 2-BCD framework whenever the error measure is differentiable, is to consider the following euristic, multiplicative update rule of the nonnegative factors, starting from an initial estimate $W^{(0)}$, $H^{(0)}$:

$$
\begin{aligned}
W^{(k+1)} &= W^{(k)} \circ \frac{\left[\nabla_W D_\beta\left(X, W^{(k)}H^{(k)}\right)\right]_-}{\left[\nabla_W D_\beta\left(X, W^{(k)}H^{(k)}\right)\right]_+} \\
H^{(k+1)} &= H^{(k)} \circ \frac{\left[\nabla_H D_\beta\left(X, W^{(k+1)}H^{(k)}\right)\right]_-}{\left[\nabla_H D_\beta\left(X, W^{(k+1)}H^{(k)}\right)\right]_+}
\end{aligned}
\tag{9}
$$

where $\circ$, $\div$ and $[\,\cdot\,]_\pm$ are the elementwise product, division and positive/negative part. Despite their euristic nature, for some values of β (including $\beta = 1$, the Kullback-Leibler divergence, and $\beta = 2$, the squared Euclidean distance) such multiplicative updates actually have convergence guarantees [1, 4, 5]: limit points of the sequence $\{(W^{(k)}, H^{(k)})\}_{k\geq 0}$ are stationary points for problem (7) where $D = D_\beta$.

4.1 Audio source separation

Perhaps the most common application of NMF within the scientific machine learning realm is that of time-frequency audio source separation and detection. Given a (discrete-time) audio mixture $\{x_t\}$ decomposed as the sum of a clean component $\{s_t\}$ and noise $\{n_t\}$:

$$x_t = s_t + n_t$$

we seek to either recover $\{s_t\}$, or detect its presence within the observed mixture. This same task can be performed in the time-frequency domain by considering the (discrete) STFT of the above signals, leading to an analogous additive decomposition depending on discrete time t and frequency f:

$$X_{f,t} = S_{f,t} + N_{f,t} \tag{10}$$

Computing the spectrogram on both sides of (10) and assuming approximate spectrogram additivity, we obtain a nonnegative matrix relation in the form:

$$X \approx S_X + N_X$$

which can be tackled by NMF. Indeed, by interpreting the nonnegative decomposition offered by NMF as a sum of R rank-1 matrices obtained by coupling each column of W with the corresponding row of H, this family of algorithms actually allow to additively decompose any observed spectrogram X. The goal then becomes devising a NMF-based

scheme able to provide the desired additive decomposition into S_X and N_X. Clearly, applying directly updates (9) to X would produce completely unstructured nonnegative factors, unfit to be used for the source separation task. One possible approach, which will be considered in the following Section 4.2, is to assume given an optimal, task-specific, structured dictionary $\hat{W}$ containing all spectral patterns needed for the source separation (or detection). The quantity of interest then becomes the coefficient matrix H, whose rows will encode when such patterns are present in X, thus allowing to actually separate S_X and N_X given the knowledge of $\hat{W}$'s structure. Formally, the process just described considers the problem:

$$\begin{aligned} &\min && D_\beta(X, \hat{W}H) \\ &\text{s.t.} && H \in \mathcal{M}_{R\times M}(\mathbb{R}^+) \end{aligned} \tag{11}$$

which can be solved iteratively with the updates:

$$H^{(k+1)} = H^{(k)} \circ \frac{\left[\nabla_H D_\beta\left(X, \hat{W}H^{(k)}\right)\right]_-}{\left[\nabla_H D_\beta\left(X, \hat{W}H^{(k)}\right)\right]_+} \tag{12}$$

4.2 Deep-NMF

In this Section we show how the deep unfolding method can be applied to NMF with the goal of performing audio source separation. The general outline follows that of [6], where a first version of the resulting neural architecture was initially conceived.

When it comes to recovering the clean source spectrogram S_X from a given mixture X, updates (12) suffer from two main issues: the potential slow convergence of the iterates $H^{(k)}$, and the ability to reconstruct exclusively what is observed within the mixture spectrogram X. While the former point is self-explanatory, the latter requires further detailing. As mentioned in the previous Section, spectrogram additivity is, in general, only approximately satisfied. Indeed, in presence of overlapping frequency content, the phase component, which we discard when considering the spectrogram, plays a fundamental role in what can be observed in the audio mixture. In the worst case, if $\{s_t\}$ and $\{n_t\}$ share a portion of their frequency content and have opposite phases, some information about the clean signal will be lost: as a consequence, we would not be able to completely recover S_X from the mixture spectrogram X alone. Both of these concerns can be addresses by employing the deep unfolding within the bilevel optimization framework presented in Section 3. As far as the inner optimization problem is concerned, we consider (11) for $\beta = 1$ (the Kullback-Leibler divergence, vastly used for source separation tasks) and regard $\hat{W}$ as the parameters which will be later untied. In our previous notation, we get $\Omega_Y = \mathcal{M}_{N\times M}(\mathbb{R}^+)$, $\Omega_X = \mathcal{M}_{R\times M}(\mathbb{R}^+)$, $\Omega_\Theta = \mathcal{M}_{N\times R}(\mathbb{R}^+)$ and:

$$\mathcal{F}_W^{in}(H, X) = D_1(X, WH) \tag{13}$$

The iterative update scheme we employ is (12), which, for the particular choice $\beta = 1$, yields[(1)]:

$$f_W^{in}(H, X) = H \circ \frac{W^\top \left(\frac{X}{WH}\right)}{W^\top \mathbb{1}_{N\times M}} \tag{14}$$

Lastly, the outer objective to be used during training is given by:

$$\mathcal{F}^{\,out}(H, X) = \frac{1}{2}\|X \circ F - S_X\|_2^2 + \frac{\mu}{2}\|\hat{W}_S H_S - S_X\|_2^2 \tag{15}$$

where $F = \frac{\hat{W}_S H_S}{\hat{W}H} = \frac{\hat{W}_S H_S}{\hat{W}_S H_S + \hat{W}_N H_N}$ is a Wiener filter and $(\,\cdot\,)_S$, $(\,\cdot\,)_N$ denote, respectively, the known, fixed sub-blocks of the argument matrix used to reconstruct the spectrograms S_X, N_X. This choice of $\mathcal{F}^{\,out}$ forces the architecture to recover S_X using the optimal dictionary $\hat{W}$, which is used exclusively for reconstruction and not in the intermediate layers. Moreover, the two terms in (15) play opposite and complementary roles: while the Wiener filter term incentivizes the last coefficient matrix to describe the noise component N_X of the mixture, the penalty term enforces an accurate reconstruction of the clean spectrogram S_X. The resulting supervised learning paradigm allows the network to recover S_X even when partially disrupted by noise.

Although (14) and (15) completely define our unfolded Deep-NMF architecture, the standard backpropagation updates described in Section 2 would require a projector onto the nonnegative orthant $\mathcal{M}_{N\times R}(\mathbb{R}^+)$ in order to preserve the nonnegativity of the parameters $W^{(k)}$. An alternative approach, following NMF theory presented in Section 4, consists in updating the parameters in a multiplicative fashion akin to the first equation in (9), replacing D_β with our current global training objective $\mathcal{L}$:

$$W^{(k)} \Leftarrow W^{(k)} \circ \frac{[\nabla_{W^{(k)}}\mathcal{L}]_-}{[\nabla_{W^{(k)}}\mathcal{L}]_+} \qquad \forall k = 0, \dots, K-1 \tag{16}$$

This method requires an ad-hoc, split gradient backpropagation algorithm.

5 Iterative Shrinkage-Thresholding Algorithm

Given a noisy observation vector $y \in \mathbb{R}^p$ corrupted by additive Gaussian noise $N_y \sim \mathcal{N}(0, \sigma^2 \mathbb{1}_p)$ and a dictionary matrix $D \in \mathcal{M}_{p\times m}(\mathbb{R})$, the problem of computing a sparse representation of y in the given dictionary D consists in solving the following constrained, nonlinear optimization problem:

$$\begin{array}{ll} \min & \|\alpha\|_0 \\ \text{s.t.} & \|D\alpha - y\|_2^2 \leq p\sigma^2 \end{array} \tag{17}$$

[(1)] We remark that, in the context of NMF, the operators $[\,\cdot\,]_\pm$ should be interpreted algebraically, rather than analytically. For example, if $f(x) = |x| - x^2$, the usual positive/negative part split would be defined as $f_+(x) = \max\{0, f(x)\}$ and $f_-(x) = \max\{0, -f(x)\}$, while here we use $f_+(x) = |x|$ and $f_-(x) = x^2$.

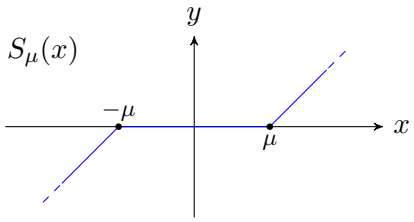
Due to the combinatorial nature of the 0−norm, which counts the nonzero entries of its argument, it is often relaxed to the 1−norm and the constraint is merged into a single loss function by introducing an appropriate Lagrange multiplier λ:

$$\min \quad \frac{1}{2}\|D\alpha - y\|_2^2 + \lambda\|\alpha\|_1 \tag{18}$$

Proximal gradient descent can then be employed to obtain a recursive optimization scheme for (18), called Iterative Shrinkage-Thresholding Algorithm (ISTA):

$$\alpha^{(k+1)} = \mathcal{S}_{\frac{\lambda}{c}}\left(\alpha^{(k)} - \frac{1}{c}D^\top\left(D\alpha^{(k)} - y\right)\right) \tag{19}$$

where c is the square spectral norm of D and $\mathcal{S}_\mu$ is the soft-thresholding operator:



5.1 Image denoising

Let us now consider the problem of denoising a grayscale image $S_Y \in \mathcal{M}_{N\times M}(\mathbb{R})$ corrupted by additive Gaussian noise $N_Y \sim \mathcal{N}(0, \sigma^2\mathbb{1})$. What we observe is therefore the image:

$$Y = S_Y + N_Y$$

A possible approach to recover S_Y is to divide Y into small square patches $y \in \mathcal{M}_{\sqrt{p}\times\sqrt{p}}(\mathbb{R}) = \mathbb{R}^p$, which allow a similar additive decomposition:

$$y = S_y + N_y$$

and reconstruct S_Y by assembling each denoised patch S_y. The denoising of each patch y can be performed with the ISTA scheme (19), where the recovered patch for the optimal sparse representation α^* of (18) is:

$$S_y \approx D\alpha^*$$

5.2 Deep-ISTA

The main drawback of the denoising procedure described in the previous Section is ISTA's slow convergence. Indeed, it is known that the algorithm may require thousands of iterations to converge to the optimal sparse representation α^*, thus applying ISTA to denoise each patch of Y would be prohibitive. Moreover, the number of iterations required may

even vary drastically depending on the noisy patch y. We will now detail how the Deep Unfolding can aid the denoising process: for a more comprehensive analysis of the complete algorithm, we refer the reader to [9].

The key aspect we can leverage to speed up the convergence is that an unfolded ISTA network must have a predetermined and fixed number of layers, which corresponds to the number of iterations we perform on any given patch y. As a consequence, by unfolding ISTA into a small $K-$layer network and choose the parameters to be learned, we can generate an ISTA-like algorithm with guaranteed approximate convergence in a small number of iterations. By choosing to learn the dictionary matrices, in our previous notation we can let $\Omega_Y = \mathbb{R}^p$, $\Omega_X = \mathbb{R}^m$, $\Omega_\Theta = \mathcal{M}_{p\times m}(\mathbb{R})$ and consider the bilevel framework having inner objective:

$$\mathcal{F}_D^{in}(\alpha, y) = \frac{1}{2}\|D\alpha - y\|_2^2 + \lambda\|\alpha\|_1 \tag{20}$$

update map:

$$f_D^{in}(\alpha, y) = \mathcal{S}_{\frac{\lambda}{c}}\left(\alpha - \frac{1}{c}D^\top (D\alpha - y)\right) \tag{21}$$

and outer objective:

$$\mathcal{F}^{out}(\alpha, y) = \frac{1}{2}\|D^{(K)}\alpha - S_y\|_2^2 \tag{22}$$

In particular, (22) enforces the unfolded network to accurately reconstruct the real image patch S_y with the last sparse representation vector $\alpha^{(K)}$ and learned dictionary $D^{(K)}$.

6 Kalman Filter

Let us now consider the problem of performing state estimation in a $p-$parametric Discrete Linear Time-Invariant (DLTI) system in state-space form:

$$\begin{aligned} x(k+1) &= A(k,p)x(k) + B(k,p)u(k) + v(k) \\ y(k) &= Cx(k) + w(k) \end{aligned} \tag{23}$$

where $x(k)$ is the state vector, $y(k)$ are measurements and $v(k) \sim \mathcal{N}(0, Q(k))$, $w(k) \sim \mathcal{N}(0, R(k))$ are independent, Gaussian model and measurement errors. Within this framework, the Kalman Filter (KF) offers a way of estimating the state at time $k+1$, denoted $\hat{x}(k)$, given the measurements $y(1), \dots, y(k)$. Namely, the KF is a recursive predictor-corrector algorithm in which the predictor updates the state estimate based on the deterministic components of (23), while the corrector modifies the latter by considering the stochastic components. The recursive scheme is the following:

$$\hat{x}(k) = \hat{x}(k \,|\, k-1) + \mathcal{K}_G^{(k)}\delta\hat{x}(k) \tag{24}$$

where $\hat{x}(k \mid k-1) = A(k-1,p)\hat{x}(k-1) + B(k-1,p)u(k-1)$ is the predictor, $\delta\hat{x}(k) = y(k) - C\hat{x}(k \mid k-1)$ is the innovation part of the corrector and $\mathcal{K}_G^{(k)} = P(k)C^\top R(k)^{-1}$ is the Kalman gain matrix, which depends on estimated state covariance:

$$P(k) = \left[\left(Q(k-1) + A(k-1,p)P(k-1)A(k-1,p)^\top\right)^{-1} + C^\top R(k)^{-1}C\right]^{-1}$$

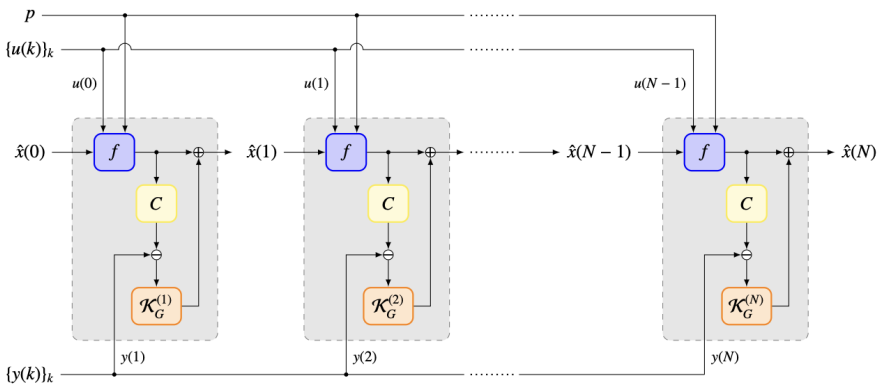
The KF working assumptions are quite restrictive: in fact, it requires a linear governing model (23), presence of Gaussian additive noise, as well as knowledge of the model and measurement error covariance matrices $Q(k)$, R. The following section focuses on how to exploit Deep Unfolding to generalize the KF to less restrictive working assumptions.

6.1 Deep-KF

The Deep Kalman Filter, briefly presented in [3], considers a generalization of (23) to a possibly nonlinear deterministic state-update map and general model and measurement noises:

$$\begin{aligned} x(k+1) &= f\left(x(k), p, u(k)\right) + v(k) \\ y(k) &= Cx(k) + w(k) \end{aligned} \tag{25}$$

The network is obtained by unfolding the Kalman-like recursive estimated-state relation (24) associated with (25) and letting the gain matrices $\mathcal{K}_G^{(k)}$ and the parameter p be the learnable weights. The resulting network is the following:



Given some measurements $\{y(k)\}_k$ the purpose of the network is to learn to encode the stochastic component of the underlying reference model (25) into the gain matrices $\mathcal{K}_G^{(k)}$ and optimize the p parameter. Compared to the KF, the Deep Kalman Filter addresses all main drawbacks of the original algorithm and poses as a machine learning alternative characterized by high flexibility and interpretability thanks to the Deep Unfolding technique.

References

[1] D.P. Bertsekas, "Nonlinear Programming". Athena Scientific, 1999.

[2] Silvia Bonettini, Giorgia Franchini, Danilo Pezzi, and Marco Prato, *Learning the image prior by unrolling an optimization method.* In Proceedings of the 30th European Signal Processing Conference, pages 952–956, 2022.

[3] Erik Chinellato and Fabio Marcuzzi, *State estimation of partially unknown dynamical systems with a deep kalman filter.* Springer LNCS 14836, Computational Science ? ICCS 2024, Part V, Chapter 22, pages 307–321, 2024.

[4] Nicolas Gillis, "Nonnegative Matrix Factorization". SIAM, 12 2020.

[5] L. Grippo, *On the convergence of the block nonlinear gauss-seidel method under convex constraints.* Operations Research Letters, 26: 127–136, 2000.

[6] John Hershey, Jonathan Le Roux, and Felix Weninger, *Deep unfolding: Model-based inspiration of novel deep architectures.* ArXiV, 09 2014.

[7] Seyed Amir Hossein Hosseini, Burhaneddin Yaman, Steen Moeller, Mingyi Hong, and Mehmet Akçakaya, *Dense recurrent neural networks for accelerated mri: History- cognizant unrolling of optimization algorithms.* IEEE Journal of Selected Topics in Signal Processing, pages 1–1, 2020.

[8] Guy Revach, Nir Shlezinger, Xiaoyong Ni, Adria Escoriza, Ruud van Sloun, and Yonina Eldar., *Kalmannet: Neural network aided kalman filtering for partially known dynamics.* IEEE Transactions on Signal Processing, 70: 1–1, 2022.

[9] Meyer Scetbon and Michael Elad, *Deep k-svd denoising.* IEEE Transactions on Image Processing, 2021.

[10] Pablo Sprechmann, Roee Litman, Tal Ben Yakar, Alexander M Bronstein, and Guillermo Sapiro, *Supervised sparse analysis and synthesis operators.* In Advances in Neural Information Processing Systems, volume 26. Curran Associates, Inc., 2013.

[11] Liang Zhang, Gang Wang, and G.B. Giannakis, *Real-time power system state estima- tion and forecasting via deep unrolled neural networks.* IEEE Transactions on Signal Processing, 2019.

Dynamical Models for Dark Matter

Gaia Marangon (*)

Abstract. Dark matter is one of the most relevant and fascinating open problems in modern astrophysics. Since it cannot be directly observed, modeling it requires a balanced mix of physical intuition, mathematical deduction, and comparison with indirect experimental data.

In these notes, I will briefly introduce the physical context motivating our research, specifically the problem of dark matter distributions around galaxies. Starting from the Schrödinger-Poisson system, the most commonly used model for dark matter dynamics, I will outline the main directions our work has taken.

I will focus on two key aspects. First, I will discuss the issue of stationary states, ranging from numerical properties to comparison with experimental data. Then, I will propose a relativistic generalization of the model, the Klein-Gordon - Wave system. Its treatment by Hamiltonian perturbative techniques shows the potential of mathematical physics tools in building a comprehensive and reliable model.

1 Dark Matter: a Physical Introduction

In these notes I will outline the problem of defining and analyzing a dynamical model for dark matter distributions around galaxies. This investigation, which primarily develops on the mathematical aspects of the problem, is, however, strictly related to the physical context, which serves both as the onset of the research and as the benchmark to check the predictive potential of our model. I will therefore start from a brief presentation of the physical context, introducing the involved experimental data and the physical assumptions underlying our modeling choices.

1.1 Physical Context: Experimental Rotation Curves

The fundamental scenario we consider is that of spiral galaxies. Spiral galaxies, such as the Milky Way, roughly consist of a flat disk of stars, possibly endowed with a small spherical bulge in the center and embedded with a cloud of interstellar gases. This whole cluster of visible matter rotates around its center. By measuring its rotational velocity at different radial positions, one obtains the so called rotation curves, which are one of the

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: marangon@math.unipd.it . Seminar held on 23 January 2025.

key experimental measurements in galactic studies. An example of these velocity profiles is illustrated in Figure 1.

Classical physics provides a straightforward prediction for the observed rotation curves:

$$V(R) = \sqrt{\frac{G\,4\pi \int_0^R \rho(s)\, s^2\, \mathrm{d}s}{R}}\,. \tag{1}$$

According to this formula, the velocity $V(R)$ at any specific radial position R depends solely on that position and on the total mass $M(R) = 4\pi \int_0^R \rho(s)\, s^2\, \mathrm{d}s$ enclosed within that radius, which is expressed as the integral of the matter density $\rho(R)$ on the sphere of radius R. This relationship implies that knowledge of the galaxy's mass distribution $\rho(R)$ enables the prediction of its rotation curve. However, when considering only the "visible" mass, observable through telescopes, the predicted curves deviate significantly from experimental data, see again Figure 1. This substantial mismatch provided the first historical evidence for the existence of dark matter, suggesting the presence of a vast, invisible ("dark") mass distribution surrounding the visible galactic structure.

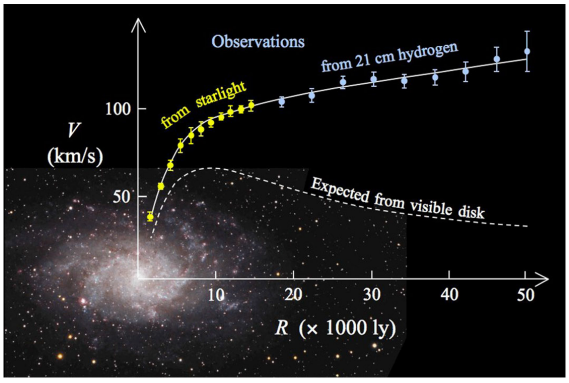


Figure 1: Observed rotation curve (yellow and blue data) compared to physical prediction (dashed line). Figure from Salucci (2019).

The aim of our research, built on extensive literature works in this field, is therefore to provide a dynamical description of this large cluster of dark matter, explaining the experimental rotation curves and at the same time being consistent with established physical principles.

To illustrate how this is performed, I will start by highlighting the basic physical assumptions that serve as a foundation for the dynamical model we study. In particular, I will focus on the cluster of dark matter, neglecting in the first place the visible stellar cluster around which it develops. The assumptions will therefore concern solely the nature of dark matter particles - which is, at present, still unknown.

1.2 Physical Assumptions

The fundamental physical hypotheses for our dark matter model consider a large cluster of particles interacting exclusively through gravitational forces. Only reciprocal interactions between the particles are considered, omitting any external potential.

The description of this system requires two primary fields: a particle density field $\psi(x,t)$ and a gravitational field $\phi(x,t)$. The particle density field is a scalar field depending on space and time, whose squared modulus represents the matter distribution throughout the galaxy. The gravitational field, also a scalar field depending on space and time, describes the mutual gravitational attraction between dark matter particles. Our aim is to define a dynamics for these two fields, which should exhibit a coupled behavior: the particle distribution generates the gravitational potential, which in turn influences the movement of the particles.

Once we set the dynamics, our primary interest lies in the stationary states. Current dark matter distributions, in fact, are presumed to have evolved from the Big Bang to reach a stable configuration, making these states most relevant for comparison with experimental data.

1.3 Schrödinger-Poisson Model and Outline of the Research

The most natural and widely studied approach to modeling this system is the Schrödinger-Poisson model, which combines Schrödinger dynamics for the matter field $\psi(x,t)$ with Poisson dynamics for the gravitational field $\phi(x,t)$. In non-dimensional units, and assuming $x \in \mathbb{R}^3$ for physical reasons, the Schrödinger-Poisson model reads:

$$\begin{cases} i\partial_t \psi = (-\triangle + 2\phi)\psi \\ \triangle \phi = |\psi|^2 \end{cases} \tag{2}$$

Observe that the matter field serves as a source term in the Poisson equation, while the gravitational potential appears as a potential term in the Schrödinger equation, thus realizing the desired coupling.

Our research explores multiple aspects of this model through several complementary approaches. We conducted extensive numerical simulations to study the structure of stationary states, revealing quantitative laws that prove valuable for comparison with experimental data.

We also investigated the model's physical foundations, particularly its connection to quantum many-body theory through mean-field derivations, providing more rigorous justification for previously ad hoc assumptions.

Then, we observed that the Schrödinger-Poisson model does not admit the possibility of any relativistic behavior for the dark matter particles. While appropriate for current dark matter distributions due to their low velocities, this prevents from applying the Schrödinger-Poisson system to early-universe predictions, since at that time dark matter is assumed to be characterized by high velocities and therefore to display a clear relativistic behavior. This deficiency prompted us to develop a relativistic extension of the

Schrödinger-Poisson model, namely the Klein-Gordon–Wave model. Through careful analysis using Hamiltonian perturbative techniques, we demonstrated that the Schrödinger-Poisson model could be recovered as a particular limit of this more general framework.

Finally, we addressed the question of stability, which represents a crucial challenge in these models. In the early 2000s, numerical simulations suggested that excited stationary states in the Schrödinger-Poisson model are unstable, but comprehensive analytical considerations on the topic are - to the present - still lacking. We therefore addressed the stability issue through analytical investigations, focusing both on the Schrödinger-Posson and on the Klein-Gordon–Wave models, primarily restricting to simplified conditions such as linear approximations or finite-dimensional restrictions. A more thorough characterization of the stability problem for the complete models remains a challenging open problem, to be addressed in future investigations.

In these notes, I will focus particularly on two aspects: the structure of stationary states, including their comparison with experimental data, and the relativistic extension through the Klein-Gordon–Wave model, illustrating its relationship to the Schrödinger-Poisson framework. These topics exemplify the rich mathematical structure and physical relevance of our research.

2 Analysis of Stationary States

In analyzing the stationary states of our system, we seek solutions where the gravitational potential is time-independent while the matter field may have a time-oscillating phase, according to the standar approach for Schrödinger-like problems:

$$\begin{cases} \text{Matter field: } e^{i\omega^2 t} f(x) \\ \text{Potential: } \phi(x) \end{cases} \tag{3}$$

When we substitute these solutions into the model equations, we obtain a static problem where the Poisson equation can be solved explicitly, yielding the potential in the form of a convolution integral. Substituting this result back into the Schrödinger equation leads to a non-linear eigenvalue problem on the matter field $f(x)$, known as Choquard equation:

$$\begin{cases} \triangle f(x) + \frac{1}{2\pi}\left(f^2 * |x|^{-1}\right) f(x) = \omega^2 f(x) & \text{with } (f^2 * |x|^{-1}) \equiv \int_{\mathbb{R}^3} \frac{f^2(y)}{|x-y|}\, \mathrm{d}^3 y \\ \phi(x) = -\frac{1}{4\pi}(f^2 * |x|^{-1}) \end{cases} \tag{4}$$

In this eigenvalue problem, the function $f(x)$ represents the eigenfunction (still related to the matter distribution, $f^2(x)$), while ω^2 serves as the eigenvalue, with physical meaning related to the system's energy. The non-linearity of this problem makes it particularly challenging to treat.

For the spherically symmetric case, the analytical studies of Lieb (1977) and Lions (1980) revealed the existence of an infinite discrete family of solutions, $\{\omega_n^2, f_n(r), \phi_n(r)\}_{n=0}^{\infty}$. Each solution, termed eigenstate, comprises an eigenvalue ω_n^2, an eigenfunction $f_n(r)$, and an associated eigenpotential $\phi_n(r)$, starting with the ground state (n = 0) and extending to infinitely excited stationary states.

2.1 Numerical Simulations: the Eigenstate Structure

Due to the absence of analytical expressions for these stationary states, numerical simulations become essential for understanding their structure. Consider, for example, the eighth excited stationary state, reported in Figure 2: the eigenfunction $f_8(r)$, illustrated in Figure 2(a), shows an oscillatory behavior with decreasing amplitudes, followed by monotonic decay after the final oscillation. This structure has a clear physical interpretation in terms of dark matter distribution $f_n^2(r)$: it suggests that dark matter organizes itself in concentric spherical shells separated by voids, with denser shells near the center and progressively less dense shells moving outward. The density eventually decays to negligible levels after the $(n+1)$-th shell, defining an approximately finite support for the physical density distribution. This three-dimensional structure is illustrated in Figure 2(b), which reports the modulus $|f_8|(r)$ rather than the matter density $f_8^2(r)$ to facilitate visualization.

To connect with experimental observations, we need to predict rotation curves. Remember that the velocity at any radial position depends solely on the mass enclosed within that radius, as in Eq. (1). Given our mass distribution $f_n^2(r)$, we can compute analytically the associated prediction for the rotation curve - which we term eigenvelocity, $v_n(r)$ - by rewriting Eq. (1) in non-dimensional units:

$$v_n(r) = \sqrt{\frac{\int_0^r f_n^2(s)\, s^2 \, \mathrm{d}s}{r}}\,. \tag{5}$$

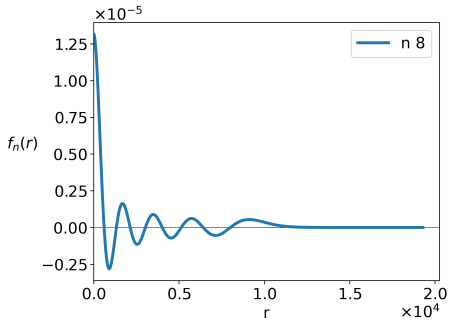
Figure 2(d) exemplifies the eigenvelocity $v_8(r)$ for the eighth excited stationary state. It exhibits an initial increase at small radii, followed by an oscillatory region, and finally a decay that coincides with the location of the last significant mass shell. Observe that the mid-range oscillating region resembles the overall behavior of the experimental data in Figure 1, suggesting a promising predictive power for the model.

2.2 Scalings with n: the Heuristic Laws

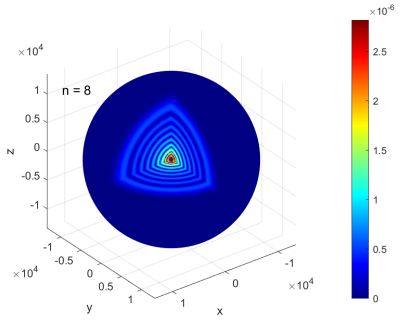
In our work, we simulate Schrödinger-Poisson stationary states up to high excitation ($n \leq 80$), exploring how the eigenfunction and eigenvelocity profiles evolve with increasing excitation number n. The results are collected in two companion papers, Marangon et al. (a) and Marangon et al. (b).

Numerical simulations reveal remarkable regularities in various aspects of the solutions, which are exemplified in Figure 3 for several excitation indices. For instance, the central peak amplitude in the eigenfunction decreases with excitation index, while the radial extent of the eigenfunction oscillatory region increases (see Figure 3(a)). Similarly, the mid-range oscillating region of the eigenvelocity has an average linear trend, with slopes that decrease with the excitation index (see Figure 3(b)).

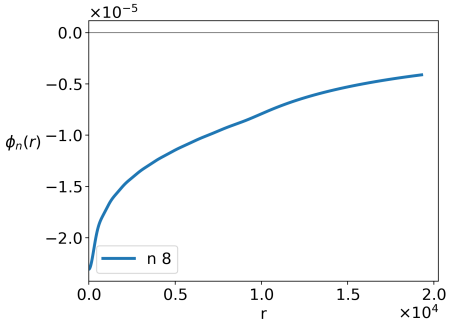
These behaviors, which can be investigated for a whole set of properties of the eigenstates, follow precise quantitative rules, which emerge clearly from our numerical simulations and which are comprehensively described in papers Marangon et al. (a) and Marangon et al. (b) These heuristic laws prove extremely valuable for predicting properties of eigenfunctions and velocity curves without resorting to extensive simulations, greatly facilitating comparison with theoretical results from the literature and with experimental data.



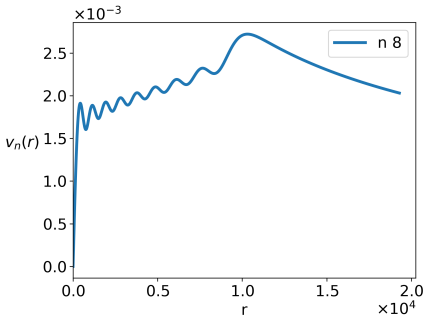
(a) Eigenfunction $f_8(r)$.



(b) Three dimensional structure.

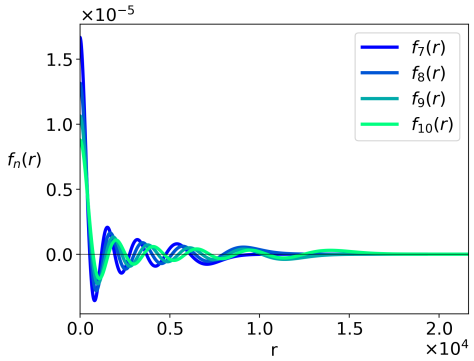


(c) Eigenpotential $\phi_8(r)$.

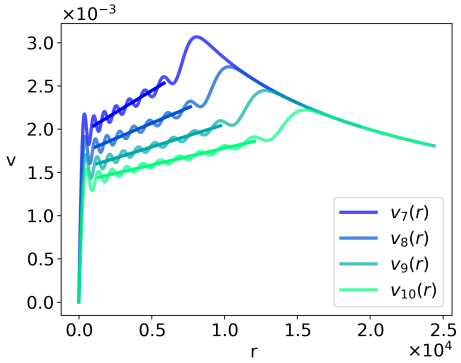


(d) Eigenvelocity $v_8(r)$.

Figure 2: Eighth eigenstate: (a) eigenfunction $f_8(r)$; (b) three-dimensional section of the dark matter distribution, represented as $|f_8|(r)$ rather than $|f_8|^2(r)$ for visualization purposes; (c) eigenpotential $\phi_8(r)$; (d) eigenvelocity $v_8(r)$, defined by Eq. (5).



(a) Eigenfunctions $f_n(r)$.



(b) Eigenvelocities $v_n(r)$.

Figure 3: Examples of eigenstates (eigenfunction in (a), eigenvelocity in (b)) for several excitation indices $n = 7, 8, 9, 10$. The plots give an intuition on the regularity of the stuctures across different values of n.

2.3 Comparison with Experimental Data

The practical value of these heuristic laws becomes evident when comparing model predictions with actual galactic observations. Specific features of the eigenvelocities, such as the position and velocity of the first and last local maxima, are detected and expressed as functions of the excitation index n. These features are then constrained to match the corresponding physical values, extracted from experimental rotation curves. The resulting relations, based on the heuristic laws, allow us to find optimal values for the free parameters of the model, in a much easier way then simply testing a uniform grid of parameters, that would be the default strategy without heuristic laws.

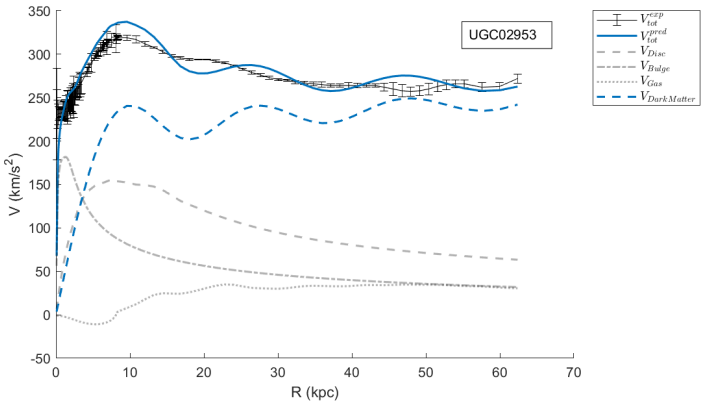


Figure 4: Fit of experimental rotation curve for the UGC02953 galaxy (data from Lelli et al. (2016)). Visible mass contributions (disk, bulge, gas) are shown in gray. Dark matter contribution, obtained by expressing the eigenvelocity in physical units, is shown in dashed blue. The total predicted curve, in solid blue, matches the experimental data, in black, remarkably well. The optimal fit of the free parameters is based on the use of the heuristic laws characterizing the Schrödinger-Poisson eigenstates.

Figure 4 reports a concrete example of such comparison, plotting the experimental rotation curve for the UGC02953 galaxy (in black, from Lelli et al. (2016)) against the optimal model prediction (in solid blue). To provide a reliable prediction, we must account for the presence of visible matter (stellar disk, stellar bulge and gas), whose distribution contributes as well to generating velocity components:

$$\begin{cases} V_{Disk}(R) &= \sqrt{\dfrac{G\,4\pi\int_0^R \rho_{Disk}(s)\,s^2\,\mathrm{d}s}{R}} \\ V_{Bulge}(R) &= \sqrt{\dfrac{G\,4\pi\int_0^R \rho_{Bulge}(s)\,s^2\,\mathrm{d}s}{R}} \\ V_{Gas}(R) &= \sqrt{\dfrac{G\,4\pi\int_0^R \rho_{Gas}(s)\,s^2\,\mathrm{d}s}{R}} \end{cases} . \tag{6}$$

Such contributions, shown in gray in Figure 4, are obtained from luminosity observations reported in the literature (see Lelli et al. (2016)). In order to obtain the total predicted rotation curve $V_{tot}^{pred}(R)$, they must be combined with the eigenvelocity $V_{DarkMatter}(R)$,

shown in dashed blue and obtained by expressing $v_n(r)$ in physical units. The overall prediction is naturally obtained by summing the mass distributions - or, in terms of velocities:

$$V_{Tot}(R) = \sqrt{V^2_{Disk}(R) + V^2_{Bulge}(R) + V^2_{Gas}(R) + V^2_{DarkMatter}(R)}\,. \tag{7}$$

Through this formula, the dark matter contribution emerging from our model combines with the visible components to produce a total predicted rotation curve that matches experimental data remarkably well. This fitting process required careful consideration of multiple factors, including the extraction of stellar velocity profiles from literature data and the incorporation of external source effects in our model. Despite its complexity, the process demonstrates the model's ability to produce predictions that align well with observational data.

3 Relativistic Extension: The Klein-Gordon–Wave Model

The Schrödinger-Poisson model, while effective, lacks relativistic nature. This limitation is evident in its non-covariant structure, where time and space play fundamentally different roles - time appears as a first derivative while space appears as a second derivative. To incorporate relativistic effects, we define the Klein-Gordon–Wave model, replacing the Schrödinger equation with a Klein-Gordon equation for the matter field $u(x,t)$ and the Poisson equation with a Wave equation for the potential $\phi(x,t)$. This modification is obtained in a natural and physically intuitive way, by assuming a massive relativistic dispersion relation for the matter field and a massless relativistic dispersion relation for the gravitational field, and then by first-quantizing them. The resulting system is naturally endowed with a covariant structure, with both space and time appearing as second derivatives.

$$\begin{cases} \Box u = (\lambda + 2\phi)u & (\Box := \triangle - \partial_{t^2}) \\ \Box\phi = u^2 \end{cases} \tag{8}$$

Here, the squared matter field $u^2(x,t)$ still represents the density distribution of dark matter, while $\phi(x,t)$ is still the gravitational potential, denoted by the same notation. The parameter λ, instead, is a specific feature of the Klein-Gordon–Wave model. In the non-dimensional system (8), this parameter collects the physical information on the amount of dark matter in the distribution, being related to both particle number and individual particle mass. Its role will be crucial in determining the model's behavior, as shown in the following.

Before delving in thorough considerations on the model's behavior, let us remark some notable consequences of adopting the Klein-Gordon–Wave model, which emerge directly from its analytical form. First, in the Schrödinger-Poisson model, the Poisson equation's lack of time derivatives implies an instantaneous adjustment of the potential on the value of the source:

$$\phi(x,t) = -\frac{1}{4\pi}\int \frac{|\psi|^2(y,\,t)}{|x-y|}\,\mathrm{d}^3y\,.$$

The wave equation in our new model, instead, ensures that the gravitational potential adjusts with finite speed, introducing a more realistic delay in gravitational interactions:

$$\phi(x,t) = -\frac{1}{4\pi}\int \frac{u^2(y,\, t-|x-y|)}{|x-y|}\,\mathrm{d}^3y\,.$$

Additionally, while the Schrödinger-Poisson model preserves particle number, $\partial_t\left(\int_{\mathbb{R}^3}|\psi|^2\mathrm{d}^3x\right) = 0$, the Klein-Gordon–Wave model breaks this conservation law, $\partial_t\left(\int_{\mathbb{R}^3} u^2\,\mathrm{d}^3x\right) \neq 0$, allowing for particle creation and annihilation, another physically desirable feature.

3.1 Klein-Gordon–Wave Regimes

Once established these straightforward observations, it is natural to investigate more thoroughly the model's behavior, and particularly the stationary states, checking the possible similarities with the Schrödinger-Poisson system.

When we examine stationary states - solutions independent of time, $u(x)$, $\phi(x)$ - we find they satisfy the same non-linear eigenvalue problem as in the Schrödinger-Poisson case:

$$\begin{cases} -\triangle u - \frac{1}{2\pi}(u^2 * |x|^{-1})u = -\lambda u \\ \phi(x) = -\frac{1}{4\pi}(u^2 * |x|^{-1}) \end{cases} \tag{0.1}$$

with $*$ still denoting the convolution. The eigenvalue, however, is now represented by the parameter λ. This creates an interesting constraint: stationary states exist only for specific values of λ, $\lambda \lesssim 1$ - which means, recalling the physical meaning of λ, they occur only for specific amounts of matter in the cluster.

This constraint suggests the Klein-Gordon–Wave model could describe the phenomenon of evaporation: the system may radiate matter, gradually reducing the amount of clustered particles, and this radiation process may stop when the amount of mass becomes compatible with the λ-value of an eigenstate. Curiously, this type of phenomenon may effectively describe the behavior of primordial black holes - astrophysical objects that were suggested by Carr and Hawking (1974) as dark matter candidates. These objects are believed to have undergone an evaporation process which has currently stopped, and the current mass estimates for these objects yield compatible $\lambda \sim 1$ values, suggesting they might effectively be understood as stationary states of our model.

This analysis suggests that the requirement $\lambda \lesssim 1$ is necessary to realize stationary states, and that this first regime can actually explain existing physical situations. One may wonder if the opposite regime $\lambda \gg 1$, not compatible with stationary states, is also physically relevant. Indeed, the galactic scale distributions described in the first part of these notes yield a value $\lambda \sim 10^6$, which definitely belongs to this second regime and therefore seems to preclude stationary states. This seems to be a problem, since physical intuition suggests galactic distributions should be stationary. This apparent limitation can be overcome through an elegant manipulation: through perturbative techniques, valid in this regime, we derive an approximate dynamics, whose stationary states can be compared with experimental data. We therefore solve the issue by modeling physical distributions through "approximate" stationary states, the approximation meaning that these states

are stationary for the approximate dynamics rather than for the complete Klein-Gordon–Wave dynamics. Remarkably, this approximate dynamics turns out to be the Schrödinger-Poisson model itself, and its stationary states are precisely the ones described in Section 2, successfully fitting the experimental data.

The overall scenario thus suggests that the Klein-Gordon–Wave model is indeed a good relativistic model for dark matter dynamics, encompassing two distinct physical regimes and recovering the affirmed Schrödinger-Poisson model as a limiting case.

As a last step of these notes, I will briefly describe how to derive the approximated dynamics - the Schrödinger-Poisson model - starting from the more general Klein-Gordon–Wave model. As anticipated, this process exploits perturbative techniques typical of Hamiltonian systems - elegantly showing how mathematical physics can both validate and provide rigorous justification for effective physical models. A more detailed description can be found in Marangon et al. (c).

3.2 From Klein-Gordon–Wave to Schrödinger-Poisson

We now restrict to the $\lambda \gg 1$ regime and describe how perturbative techniques can be applied.

Let us start by observing that the Klein-Gordon–Wave model has an Hamiltonian structure. Using $(u, p_u := \partial_t u)$, $(\phi, p_\phi := \partial_t \phi)$ as conjugated variables and passing to first-order-in-time equations, the Klein-Gordon–Wave system can be expressed as:

$$\begin{cases} \partial_t u &= \frac{\delta H}{\delta p_u} &= p_u \\ \partial_t p_u &= -\frac{\delta H}{\delta u} &= -(\lambda + 2\phi)u + \triangle u \\ \partial_t \phi &= \frac{\delta H}{\delta p_\phi} &= p_\phi \\ \partial_t p_\phi &= -\frac{\delta H}{\delta \phi} &= \triangle \phi - u^2 \end{cases}$$

with Hamiltonian:

$$H(u, \phi, p_u, p_\phi) := \int \left(\frac{p_u^2 + \lambda u^2 + |\nabla u|^2}{2} + \frac{p_\phi^2 + |\nabla \phi|^2}{2} + u^2 \phi \right) \mathrm{d}^3 x \,.$$

The presence of an Hamiltonian structure enables the use of a wide variety of techniques, characteristic of this class of systems. To be able to exploit them, let us first perform a convenient change of variables. First, we adjust the time scale, using the same notation for simplicity:

$$u \to u \,; \qquad \phi \to \phi \,; \qquad x \to x \,; \qquad t \to \frac{1}{\sqrt{\lambda}} t \,.$$

Then, we pass to the so called Birkhoff or Dirac variables. This change of variables is non canonical, meaning that it does not preserve the form of the Hamiltonian equations. We therefore need to recompute the Poisson parenthesis as well:

$$\psi := \frac{u + i p_u}{\sqrt{2}} \,; \qquad \overline{\psi} := \frac{u - i p_u}{\sqrt{2}} \,; \qquad \{\psi(x), \psi(y)\} = -i\delta(x - y) \,.$$

The Hamiltonian $\widetilde{H}(\psi, \overline{\psi}, \phi, p_\phi)$ in the new set of variables has a convenient structure. We can isolate the parameter $\varepsilon := \lambda^{-1}$ and observe that it is small, $\varepsilon \ll 1$ in the $\lambda \gg 1$ regime we're analyzing. It can therefore act as a perturbative parameter, splitting the Hamiltonian in an unperturbed term H_0 and a correction εH_1, smaller in ε:

$$\widetilde{H}(\psi, \overline{\psi}, \phi, p_\phi) := \underbrace{\int \left(|\psi|^2 + \frac{p_\phi}{2} \right) \mathrm{d}^3 x}_{H_0} \underbrace{+\frac{1}{\lambda}}_{+\varepsilon} \underbrace{\int \left(\frac{\left| \nabla\psi + \nabla\overline{\psi} \right|^2}{4} + \phi \frac{(\psi + \overline{\psi})^2}{2} + \frac{|\nabla\phi|^2}{2} \right) \mathrm{d}^3 x}_{H_1}$$

This perturbative structure is convenient for the application of a relevant Hamiltonian technique, leading to the following Proposition.

Proposition 1 *There exists a canonical transformation of the fields, mapping the Hamiltonians:* $H_0 + \varepsilon H_1 \quad \to \quad H_0 + \varepsilon \langle H_1 \rangle_0 + \mathcal{R}$, *where* $\langle H_1 \rangle_0$ *denotes the time average of* H_1 *along the unperturbed flow* Φ_0^s *at* $p_\phi = 0$, *and the remainder* $\mathcal{R}$ *is of order* $\mathcal{O}(\varepsilon^2)$.

The resulting Hamiltonian $H_0 + \varepsilon \langle H_1 \rangle_0 + \mathcal{R}$ is in the so called Normal Form, where we have further explicated the perturbative structure, isolating a remainder $\mathcal{R}$ which is of smaller order, $\mathcal{O}(\varepsilon^2)$. Neglecting it, we obtain a first order approximated Hamiltonian:

$$H_S := H_0 + \varepsilon \langle H_1 \rangle_0 = \int \left(|\psi|^2 + \frac{p_\phi}{2} \right) \mathrm{d}^3 x + \varepsilon \int \left(\frac{|\nabla\psi|^2}{2} + \phi\, |\psi|^2 + \frac{|\nabla\phi|^2}{2} \right) \mathrm{d}^3 x$$

By writing the associated Hamiltonian equations, we obtain the desired approximate dynamics. With a phase shift $\psi \to e^{-it}\psi$ and a time rescaling $t \to \frac{t}{\varepsilon}$, the Hamiltonian equations associated to H_S read:

$$\begin{cases} i\partial_t \psi = -\frac{1}{2} \triangle \psi + \phi\psi \\ \partial_t \phi = \frac{1}{\varepsilon} p_\phi \\ \partial_t p_\phi = \triangle \phi - |\psi|^2 \end{cases} \quad \Rightarrow \begin{cases} i\partial_t \psi = -\frac{1}{2} \triangle \psi + \phi\psi \\ \varepsilon \frac{\partial^2 \phi}{\partial t^2} = \triangle \phi - |\psi|^2 \end{cases}$$

We recognize here a Schr̈odinger-Wave model. Observe, finally, that the time derivative in the Wave equation is weighted by the perturbative parameter $\varepsilon \ll 1$. As a consequence, it vanishes in the $\varepsilon \to 0$ limit, reducing the Wave equation to the Poisson equation. This concludes the procedure, proving that the popular Schrödinger-Poisson model can be obtained, in a suited limit, as the first order Normal Form truncation of the more general Klein-Gordon–Wave model.

References

B.J. Carr and S.W. Hawking, *Black holes in the early Universe.* Monthly Notices of the Royal Astronomical Society, 168:399–416, August 1974. doi: 10.1093/mnras/168.2.399. URL https://ui.adsabs.harvard.edu/abs/1974MNRAS.168..399C.

Federico Lelli, Stacy S. McGaugh, and James M. Schombert, *SPARC: mass models for 175 disk galaxies with Spitzer photometry and accurate rotation curves.* The Astronomical Journal, 152 (6):157, nov 2016. doi: 10.3847/0004-6256/152/6/157. URL https://dx.doi.org/10.3847/0004-6256/152/6/157.

Elliott H. Lieb, *Existence and uniqueness of the minimizing solution of Choquard's nonlinear equation.* Studies in Applied Mathematics, 57(2):93–105, 1977. doi: https://doi.org/10.1002/sapm197757293. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sapm197757293.

P.L. Lions, *The Choquard equation and related questions.* Nonlinear Analysis: Theory, Methods & Applications, 4(6):1063–1072, 1980. ISSN 0362-546X. doi: https://doi.org/10.1016/0362-546X(80)90016-4. URL https://www.sciencedirect.com/science/article/pii/0362546X80900164.

Gaia Marangon, Antonio Ponno, and Lorenzo Zanelli (a), *On the scaling properties of excited stationary states of the Schrödinger-Poisson model.* Submitted to Journal of Mathematical Physics, Special Topic (Dec. 2024).

Gaia Marangon, Antonio Ponno, and Lorenzo Zanelli (b), *Scaling of highly excited Schrödinger-Poisson eigenstates and universality of their rotation curves.* Physical Letters A 555 (2025), 130761. doi: https://doi.org/10.1016/j.physleta.2025.130761. URL https://www.sciencedirect.com/science/article/pii/S0375960125005419.

Gaia Marangon, Antonio Ponno, and Lorenzo Zanelli (c), *From Klein-Gordon-Wave to Schrödinger-Wave: a Normal Form Approach.* Submitted to Journal of Physics A. URL https://doi.org/10.48550/arXiv.2504.20576.

Paolo Salucci, *The distribution of dark matter in galaxies.* The Astronomy and Astrophysics Review, 27(1):2, February 2019. doi: 10.1007/s00159-018-0113-1. URL https://link.springer.com/article/10.1007/s00159-018-0113-1.

Mixing times and cutoffs for Markov chains

Giacomo Passuello (*)

Abstract. How long does it take to shuffle a deck of 40 cards? This simple question, together with the seminal work of Aldous and Diaconis on the cutoff phenomenon, has generated, in the last 40 years, a rich research area in the field of discrete probability. A cutoff is a dynamical phase transition for a random process, which appears as the size of the system becomes large. It occurs when the distance to equilibrium of the process abruptly drops from its maximum value to zero at a critical time scale. Establishing the occurrence of the cutoff is a delicate matter, which may require a precise understanding of the spectral and diffusive properties of the underlying system. In this talk, I will review some basic concepts on Markov chains and their convergence to the stationary equilibrium. After that, I will introduce the concept of mixing time and discuss bounds on its limiting behaviour. Finally, I will focus on the cutoff phenomenon and present some results on the mixing time of the simple random walk on a directed random graph.

1 Introduction

Markov chains, and Markov processes in general, constitute one of the most studied family of random dynamics. They capture a very simple though powerful feature shared by many random systems: absence of memory. This property is present in a variety of processes and phenomena in natural, social, and economic science, such as opinion dynamics, the spread of epidemics, the motion of interacting particles, and fluctuations of stock prices. The theory of Markov processes is well developed in both discrete and continuous setting. In what follows, we will focus on the former, which has the same main qualitative features of the continuous one, benefiting of a more intuitive formalism. Many spin systems lying at the interface between probability and statistical mechanics, such as the Moter Model, the Contact Process, and the Ising Model, can be modelled by a Markov chain, and their scaling limit is able to produce a complex behaviour and the arousal of non-trivial phenomena known as phase transitions.

Because of their intrinsic algorithmic structure, Markov chains provide a great tool for computer scientists and statisticians, as they can be easily used to simulate physical, biological, and social systems. In particular, they are often used to design Monte-Carlo

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: giacomo.passuello@phd.unipd.it . Seminar held on 6 February 2025.

methods, which can produce approximate random samples of a target probability distribution. In this regard, the study of the mixing behaviour of random systems, namely their rate of convergence to the equilibrium, constitutes a deep theoretical tool, providing effective stopping criteria.

In the last decades, there has been a further step, given by the approach to markovian dynamics on random graphs. These random environments capture the typical connectivity features of huge networks such as the World Wide Web, social networks, citation networks, without the bias given by the partial observation real data. We refer to [14] for an introduction to random graphs.

This note aims to provide a self-contained presentation of some basic notions on Markov chains, mixing times, random graphs and some results on the convergence to equilibrium random walks on directed random graph models. We refer to [16, 15] for related readings.

2 Markov Chains

Consider a probability space $(\Omega, \mathcal{F}, \mathbf{P})$.

Definition Given a discrete set V, a Markov chain with state space V is a family of random variables $X = (X_t)_{t\in\mathbb{N}}$ with values in V such that the *Markov property* holds: for all $t \in \mathbb{N}$, and all $x_0, \dots, x_{t-1}, x, y \in V$, it holds

$$\mathbf{P}\left(X_{t+1} = y \mid X_t = x, X_{t-1} = x_{t-1}, \dots, X_0 = x_0\right) = \mathbf{P}(X_{t+1} = y \mid X_t = x).$$

In words, for a Markov chain, future depends on the past only via the most recent information.

This definition can be easily generalized to non-discrete times, and to non-discrete state spaces V, leading to the general concept of Markov process. In that case the state space has to be endowed with a Borel σ-algebra and the process $(X_t)_{t\in\mathbb{R}^+}$ needs to be adapted to a filtration $(\mathcal{F}_t)_{t\in\mathbb{R}^+}$, that is an increasing sequence of sub-σ-algebras of $\mathcal{F}$. Then, the Markov property above turns to ask that for every $t \in \mathbb{R}^+$ and $A \in \mathcal{F}$,

$$\mathbf{P}\left(X_{t+1} \in A \mid \mathcal{F}_t\right) = \mathbf{P}(X_{t+1} \in A \mid X_t).$$

From now on we will always consider a *time-homogeneous* Markov chain X that is, for every $x, y \in V$ the transition probabilities $\mathbf{P}(X_{t+1} = y \mid X_t = x)$ do not depend of t. In that case, it is possible to define the transition matrix P of the chain by

$$P(x, y) \coloneqq \mathbf{P}(X_1 = y \mid X_0 = x), \qquad x, y \in V,$$

and given an initial distribution μ, supported on V, at any given time $t \in \mathbb{N}$, the distribution of the process at time t is given by powers of P:

$$\mathbf{P}_\mu(X_t = y) = \mu P^t(y) \coloneqq \sum_{x\in V} \mu(x) P^t(x, y).$$

In the continuous time setting, the notation $P^t(x,y)$ refers to the Markov *semigroup* associated of the chain, that is a operator such the following composition rule holds: for every $t,s>0$ it holds $P^{t+s}(x,y)=\sum_{z\in V}P^t(x,z)P^s(z,y)$.

A Markov chain X with transition matrix P is

- *aperiodic* if there is no partition $V=C_1\cup\ldots\cup C_k$, s.t., $\forall x\in C_k$, it holds $P(x,C_{k+1})=1$,
- *irreducible* if for every $x,y\in V$, there exist $t>0$ such that $P^t(x,y)>0$, i.e., it is possible to reach y starting from x, with positive probability.

Example (Simple Random Walk) Consider a graph G, i.e., a couple (V,E) with $E\subseteq V\times V$. The Simple Random Walk (SRW in the following) on G is a Markov chain X with transition matrix

$$P(x,y)=\begin{cases}\frac{1}{D_x^+} & \text{if } (x,y)\in E\\ 0 & \text{otherwise}\end{cases},\qquad x,y\in V,$$

where $D_x^+ \coloneqq |\{y\in V:(x,y)\in E\}|$ is the (out-)degree of x. Depending on the specific choice of graph G its properties may vary. For instance, the SRW on $\mathbb{Z}$ is periodic: starting from 0, for even times it will always be in even positions and vice-versa. On the other hand, the SRW on $\mathbb{Z}$ is irreducible, since this happens every time that the underlying graph is connected.

2.1 Long-run evolution

We are interested in considering and characterizing the convergence towards the equilibrium for a Markov chain X with transition matrix P. To do so, we first have to identify a notion of equilibrium and a distance between probability distributions. Intuitively, an equilibrium has to be linked to a notion of invariance under the action of the transition operator P. For a probability distribution π, this writes $\pi P=\pi$. A such distribution is called *invariant* or *stationary*. It is possible to show that irreducible Markov chains admit a unique invariant distribution.

It turns out that invariant distributions provide the only possible target for the *long-run evolution* of a Markov Chain. A very basic theorem states that if X is an aperiodic and irreducible Markov chain with transition matrix P, then, for every probability distribution μ, supported on V, it holds

$$\frac{1}{2}\sum_{y\in V}\left|\mu P^t(y)-\pi(y)\right|\xrightarrow[t\to+\infty]{}0,$$

where π is the unique stationary distribution. The object on the left takes the name of *total variation distance* and in what follows it will be denoted by

$$d_{\mathrm{TV}}(\nu_1,\nu_2):=\frac{1}{2}\sum_{y\in V}|\nu_1(y)-\nu_2(y)|,$$

for any two probability distributions ν_1, ν_2 supported on V. This constitutes the most natural notion of distance between probability measure, corresponding to the L^1 distance. It is possible to consider other kind of metrics by passing to the density ν_1/ν_2 of ν_1 with respect to ν_2 and taking, for instance, suitable L^p distances with $p > 1$, or the relative entropy.

Example If $n \in \mathbb{N}$ is odd, the SRW on the integer torus $\mathbb{Z}/n\mathbb{Z}$, is such that

$$\forall x, y \in \mathbb{Z}/n\mathbb{Z}, \quad P^t(x,y) \xrightarrow[t\to+\infty]{} \frac{1}{n}.$$

However, if n is even, the chain is periodic and the above convergence does not hold.

In general, it can be easily verified that for the SRW on an undirected (and aperiodic) connected random graph $G = (V, E)$, the stationary distribution π is proportional to the degrees of the graph, and hence is given, for every $y \in V$, by $\pi(y) = D_y^+/2|E|$. However, if the graph is directed, the stationary distribution has no explicit form in terms of the degrees.

2.2 Monte Carlo simulations

A Markov chain benefiting of a sufficiently fast convergence to the stationary equilibrium, can be used to compute averages in an approximated way, when explicit computations become numerically unfeasible. We provide here a motivating example.

Given a graph $G = (V, E)$, a positive *energy functional* H on V, and a parameter $\beta > 0$ (called *inverse temperature*), consider the *Gibbs density*

$$\pi_\beta(x) \coloneqq \frac{e^{\beta H(x)}}{\sum_{y\in V} e^{\beta H(y)}}, \qquad x \in V.$$

Gibbs densities frequently arise in the context of equilibrium statistical mechanics. In this setting, V is usually taken to be the set of of ± 1-spin configuration on a bounded domain Λ of $\mathbb{Z}^d$ of size $N \in \mathbb{N}$. If N is sent to $+\infty$, the size of V is 2^N and diverges exponentially fast.

Given a bounded function $f : V \mapsto \mathbb{R}$, consider now the problem of computing numerically

$$\mathbb{E}_{\pi_\beta}[f] \coloneqq \sum_{x\in V} f(x)\pi_\beta(x).$$

Due to the presence of the normalization constant in the Gibbs density, if N is large, the computation may be extremely lengthy. A way to overcome the problem is to construct an irreducible Markov Chain with stationary distribution π_β and simulate it until it is *"well mixed"*. Roughly, this will give an approximated sample of π_β and an empirical average of some iterated samples will give an approximation of the average desired average. In the case of the Gibbs density above, a candidate is the Markov chain X with transition matrix

P_β defined as,

$$P_\beta(x,y) = \begin{cases} \frac{1}{D_x^+} \min\left\{\frac{\pi_\beta(y)}{\pi_\beta(x)}, 1\right\} & \text{if } (x,y) \in E \\ 0 & \text{otherwise} \end{cases}, \qquad \text{for } x \neq y,$$

and $P_\beta(x,x) = 1 - \sum_{y \neq x} P_\beta(x,y)$. It holds $\pi_\beta P_\beta = \pi_\beta$, and hence π_β is invariant under P_β. X is defined through an acceptance-rejection scheme, which is commonly referred to as *Metropolis–Hastings algorithm*. Of course, to make this approximated average precise, one needs to characterize the speed of convergence of X to π_β, when the latter is the unique invariant distribution of X. This motivates the study of *Mixing times* for Markov chains.

3 Mixing time

We introduce the notation $d(t) := \max_{x \in V} d_{\mathrm{TV}}(P^t(x,\cdot), \pi)$, and, for $0 < \epsilon < 1$, we define the *worst-case mixing time* by

$$t_{\mathrm{mix}}(\epsilon) :=\inf\{t \geq 0 : d(t) \leq \epsilon\}.$$

Notice that the maximum, can be taken over all initial distributions (not only Dirac deltas), still giving the same number, by convexity. Moreover, this definition depends on the choice of the distance. One of the ones cited above, will still provide a meaningful and interesting object.

In what follows, we will take V to be a set of $n \in \mathbb{N}$ labelled elements. We will later send $n \to +\infty$, being interested in estimating the order of $t_{\mathrm{mix}}(\epsilon)$ as n grows.

In some cases, a careful analysis of the spectrum of the transition matrix P, can provide upper and lower bounds on $t_{\mathrm{mix}}(\epsilon)$. This is the case for a reversible dynamics. A transition matrix P is *time-reversible* w.r.t. π, if

$$\pi(x)P(x,y) = \pi(y)P(y,x), \qquad \forall x, y \in V.$$

This is equivalent to say that P is self-adjoint in the Hilbert space $L^2(\pi)$ of π-square-integrable functions, endowed with the scalar product defined by

$$\langle f, g\rangle_\pi \coloneqq \sum_{x \in V} f(x)g(x)\pi(x), \quad f, g \in L^2(\pi).$$

Then, the matrix $\left(\sqrt{\frac{\pi(y)}{\pi(x)}}P(x,y)\right)_{x,y}$, which has the same spectrum as P, is symmetric.

Example It is easy to check that the SRW on a undirected connected graph is reversible w.r.t. to its unique stationary distribution.

If P is irreducible and reversible, by the *Perron-Frobenius theorem*, it has eigenvalue 1 with multiplicity 1, and the other eigenvalues (they are real) can be written in increasing order:

$$1 = \lambda_1 > \lambda_2 \geq \ldots \geq \lambda_n > -1.$$

Then, the absolute spectral radius $\lambda_\star$ of X and the relaxation time $t_{\rm rel}$ of X are defined as

$$\lambda_\star := \max\{|\lambda_2|, |\lambda_n|\} \qquad \text{and} \qquad t_{\rm rel} := \frac{1}{1-\lambda_\star}.$$

The relaxation time $t_{\rm rel}$ provides an upper and a lower bound on $t_{\rm mix}$.

Theorem *If P is reversible and irreducible it holds $d(t) \le \frac{\lambda_\star^t}{2\sqrt{\min_{x\in V}\pi(x)}}$. Moreover,*

$$(t_{\rm rel}-1)\log\left(\frac{1}{2\epsilon}\right) \le t_{\rm mix}(\epsilon) \le t_{\rm rel}\log\left(\frac{1}{\epsilon\cdot\min_{x\in V}\pi(x)}\right).$$

The relaxation time, can be estimated by direct computations, Poincaré inequalities, Cheeger bounds, Log-Sobolev inequalities and other ways.

Example We provide some example of bounds for the mixing time of the SRW on networks.

- SRW on the d-dimensional torus $\mathbb{Z}^d/n\mathbb{Z}^d$ has mixing time
$$t_{\rm mix}(\epsilon) \le dn^2\lceil\log_4(d/\epsilon)\rceil,$$
- Specifically for $d=1$ (and n odd!) there exists a decreasing function function $\Psi(\cdot)$ such that $t_{\rm mix}(\epsilon) = \Psi(\epsilon)n^2$.
- SRW on the Aldous' dog has relaxation time $t_{\rm rel}$ of order $n^2\log n$, and hence the mixing time $t_{\rm mix}(\epsilon)$ has at least order $n^2\log n\log(\frac{1}{2\epsilon})$

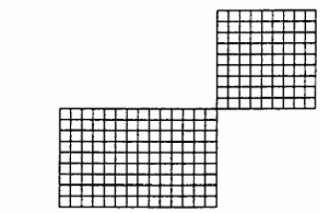


Figure 1: This is Aldous' dog.

3.1 Cutoff phenomenon

The cutoff phenomenon, corresponds to an abrupt decay of the TV-distance. It was discovered in the '80s in the context of random permutations and card shuffling [1], [11] and since then has become object of many investigations [12] and has given birth to a very wide research activity.

Sometimes the dependence on ε of $t_{\rm mix}(\varepsilon)$ is very weak, as the following example shows.

Example (Card shuffling) There are many ways to shuffle n cards. A natural one consists in: splitting the deck in two parts and deciding a (random) procedure to interleave cards from the two halves to form a new deck. It has been shown in [2] that, repeating the one of these procedures t times, where

$$t = \frac{3}{2}\log_2 n + \theta, \qquad \theta > 0,$$

the distance to the equilibrium (uniform distribution on the $n!$ permutations of n elements), is

$$1 - \frac{2}{\sqrt{2\pi}} \int_{-\infty}^{\frac{-2^{-\theta}}{4\sqrt{3}}} e^{-u^2/2}\, du + O\left(\frac{1}{n^{1/4}}\right).$$

This implies that fixing $\gamma > 0$ and repeating the procedure $(\frac{3}{2} - \gamma)\log_2 n$ times, the distance converges to 1 as $n \to \infty$, while repeating the procedure $(\frac{3}{2} + \gamma)\log_2 n$ times, the distance converges to 0 as $n \to \infty$. Then, for every $\varepsilon > 0$, $t_{\text{mix}}(\varepsilon) = \frac{3}{2}\log_2 n(1 + o(1))$.

This is an occurrence of the so–called cutoff phenomenon.

m	1	2	3	4	5	6	7	8	9	10
25	1.000	1.000	0.999	0.775	0.437	0.231	0.114	0.056	0.028	0.014
32	1.000	1.000	1.000	0.929	0.597	0.322	0.164	0.084	0.042	0.021
52	1.000	1.000	1.000	1.000	0.924	0.614	0.334	0.167	0.085	0.043
78	1.000	1.000	1.000	1.000	1.000	0.893	0.571	0.307	0.153	0.078
104	1.000	1.000	1.000	1.000	1.000	0.988	0.772	0.454	0.237	0.119
208	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.914	0.603	0.329
312	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.883	0.565

Figure 2: Total variation distance for m shuffles of 25, 32, 52, 104, 208 or 312 distinct cards ([2]).

Definition (Cutoff) A sequence of transition matrices exhibits a *cutoff* if there exists a time–scale $t_n^\star$ such that, for every $\epsilon > 0$, $\lim_{n\to+\infty} t_{\text{mix}}(\epsilon)/t_n^\star = 1$. This is equivalent to say that, for every $\gamma > 0$,

$$\lim_{n\to+\infty} d((1-\gamma)t_n^\star) = 1 \quad \& \quad \lim_{n\to+\infty} d((1+\gamma)t_n^\star) = 0.$$

This means that the limit shape of total variation profile, namely the plot of the distance $d(\cdot)$ shows an abrupt decay of the TV-distance at the critical timescale $t_n^\star$. This provides a first order description for the mixing time, which can refined in precence of a cutoff *window*.

Definition (Cutoff with window) A sequence of transition matrices exhibits a cutoff with *window* if there exist two timescales $t_n^\star$ and $\mathrm{w}_n^\star = o(t_n^\star)$ such that,

$$\lim_{c\to-\infty}\lim_{n\to+\infty} d(t_n^\star + c\mathrm{w}_n^\star) = 1 \quad \& \quad \lim_{c\to+\infty}\lim_{n\to+\infty} d(t_n^\star + c\mathrm{w}_n^\star) = 0.$$

Proving the occurrence of a cutoff is a very delicate matter, which requires the knowledge of the diffusion properties of the chain. A necessary, but not sufficient, condition is that

$$t_{\text{rel}} = o(t_{\text{mix}}).$$

Despite an increasing amount of work on the subject, the cutoff phenomenon is still far from be- ing completely understood, and the research of simple conditions (i.e., easy-to-check and model independent) guaranteeing the presence of a cutoff is still very active.

4 Random Walks on Random Graphs

In this section, we present the setting of directed graphs and recent results for the mixing time of the SRW in this environment.

A random graph is a sequence of random variables $(G_n)_{n\in\mathbb{N}}$, defined on a common probability space $(\Omega', \mathcal{F}', \mathbb{P})$, where, for each $n \in \mathbb{N}$, G_n is a graph with n vertices. Equivalently, a random graph model constitutes in a sequence of probability distributions $\mathbb{P}(G_n \in \cdot)$ on the set of graphs with n vertices, which satisfy proper consistence conditions. In what follows, the dependence on n will be hidden into the vertex set V of the graph, and we will say that an event A_n happens *with high probability* if its probability $\mathbb{P}(A_n)$ converges to 1 as $n \to +\infty$.

Example (Erdős-Rényi random graph) The easiest example is the Erdős-Rényi random graph. It can be obtained by fixing, for every n, a parameter $p_n \in (0, 1)$, and for each couple $x, y \in V$, including the edge (x, y) in the graph independently of the others with probability p_n. This results in a simple graph (see Fig. 3). Even though usually edges are undirected, the same can be done in the oriented setting.

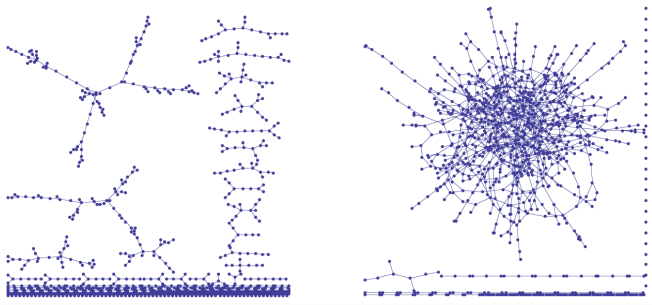


Figure 3: Realizations of Erdős-Rényi random graphs with $n = 1000$ vertices parameter $p_n = \frac{1.1}{n}$ and $p_n = \frac{2}{n}$ respectively ([14]).

Example (Directed Configuration Model) For each n, fix a deterministic bi-degree sequence $(d_x^-, d_x^+)_{x\in V}$ such that $d_x^{\pm} < +\infty$ and

$$\sum_{x\in V} d_x^- = \sum_{x\in V} d_x^+.$$

The directed configuration model is obtain by attaching, to every vertex $x \in V$, d_x^- half in-edges (heads) and d_x^+ half out-edges (tails), matching them according to a uniform random bijection (see Fig. 4). This results in a directed multigraph (where multiple edges can match the same ordered couple of vertices), but erasing the multiple connections, we can obtain a simple directed graph. It can be shown that, with high probability as $n \to +\infty$, if $d_x^\pm \geq 2$, the digraph is strongly connected (i.e., the SRW on this graph is irreducible). Moreover in can be shown that with uniformly (in n) positive probability, the digraph is simple (i.e., with no multiple edges).

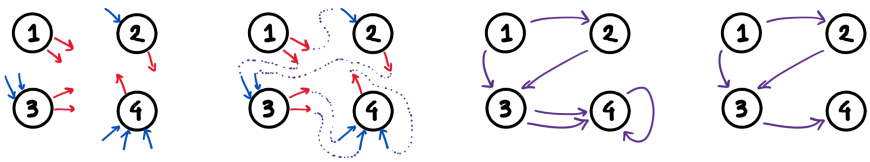


Figure 4: Construction procedure of the directed configuration model.

The next model is a sort of mixture between the previous two examples, and takes the name of Chung-Lu model.

Example (Chung-Lu digraph) For each n, consider a sequence of weights $(w_x^-, w_x^+)_{x \in V}$ with

$$\sum_{x \in V} w_x^- = \sum_{x \in V} w_x^+ \asymp n.$$

For each couple $x, y \in V$, include the oriented edge (x, y) in the digraph, independently of the others, with probability

$$p_{xy} = w_x^+ w_y^- \frac{\log n}{n} \wedge 1.$$

This implies that the order of average degree is $\asymp w_x^\pm \log n$, for large n. It can be shown that if $\min_{x \in V} w_x^\pm > 1$ uniformly (in n), then, with high probability, the digraph is strongly connected and hence the SRW on it is irreducible and admits a unique stationary distribution.

4.1 Results

We take G to be a Chung-Lu directed random graph and we consider the SRW on it. Let us point out that there are two different probability measure to be considered:

- $\mathbb{P}$: probability measure encoding the randomness of the graph
- $\mathbf{P}^G$: probability measure encoding the randomness of the SRW (for every fixed graph G)

In this setting, $t_{\text{mix}}(\epsilon)$ is a random variable depending on G. However, we can still try to show that a cutoff takes place *with high probability*, i.e., that there exists timescale $t_n^\star$ such

that, for each $0 < \epsilon < 1$,

$$\forall \delta > 0, \qquad \mathbb{P}\left(\left|\frac{t_{\text{mix}}(\epsilon)}{t^\star} - 1\right| > \delta\right) \xrightarrow[n\to+\infty]{} 0.$$

Cutoffs *with high probability* have been proved for random walks on several graph models with bounded degree: the SRW on d-regular RG [17], the non-backtracking RW on Configuration Model [4], the SRW on directed configuration model [7, 9]. The expression of the mixing time in these contexts is

$$t_n^\star \;=\; \frac{\log n}{\nu \cdot \mathrm{H}},$$

where the two quantities ν and H depend on the specific model. The first, ν, is the speed of the random walk. For sparse random graphs, represents, roughly, the almost sure limit of the ratio $d(Y_t, \rho)/t$, where $d(\cdot)$ is the graph distance and Y_t is a random walk on a suitable random ρ-rooted tree, coupled with the original graph. the second one, H, represents a notion of *row entropy* for the transition matrix P:

$$\mathrm{H} = -\frac{1}{n}\sum_{x,y\in V} P(x,y)\log P(x,y).$$

In our setting, being the graph directed and locally tree like, we have $\nu = 1$, and we can define a similar notion of entropy by

$$\mathrm{H} = \sum_{x\in V} \frac{w_x^-}{\sum_{y\in V} w_y^-}\mathbb{E}[\log(D_x^+ \vee 1)],$$

which is asymptotically $\log\log(n)$. Let $d^{(x)}(t) = d_{\text{TV}}(P^t(x,), \pi)$. Following the approach of [7, 8, 9], we have the following theorem, proved in [5].

Theorem ([5]) *Let G be a realization of the Chung-Lu digraph, and $t_n^\star = \frac{\log n}{\log\log n}$. Assume that there exist $\lambda, C > 1$ and $0 < \eta < 1$ such that*

$$\sqrt{\lambda} \le w_x^+ \le C, \quad \forall x \in V, \qquad\qquad \sum_{x\in V} (w_x^-)^{2+\eta} \le Cn.$$

Then for every $\gamma > 0$,

$$\min_{x\in V} d^{(x)}((1-\gamma)t_n^\star) \xrightarrow[n\to+\infty]{\mathbb{P}} 1, \qquad\qquad \max_{x\in V} d^{(x)}((1+\gamma)t_n^\star) \xrightarrow[n\to+\infty]{\mathbb{P}} 0.$$

It shows that cutoff takes place *with high probability* uniformly in the starting point of the SRW. In particular this holds also for the worst-case distance defined in the previous sections. Moreover, it is possible to identify the cutoff window, under an additional condition on the following quantity, which can be interpreted as a variance

$$\sigma^2 \coloneqq \sum_{x\in V} \frac{w_x^-}{\sum_{y\in V} w_y^-}\mathbb{E}[\log^2(D_x^+ \vee 1)] - \mathrm{H}^2,$$

We have the following theorem, proved in [5], in the same hypothesis of the previous one. This theorem states that the distance to equilibrium has a smooth decay, given by a Gaussian profile, independent of the parameters and visible at the critical time–scale $t^\star + c\mathrm{w}^\star_n$, for $c > 0$.

Theorem ([5]) *Let* $\mathbf{w} := \frac{\sigma}{\mathrm{H}}\sqrt{t^\star}$ *. Assume that there exists* $\xi > 0$ *such that* $\sigma^2 \gg \frac{(\log\log n)^{2+\frac{\xi}{\xi+2}}}{(\log n)^{\frac{\xi}{\xi+2}}}$. *Then*

$$\max_{x\in V}\left|d^{(x)}(t^\star_n + c\mathrm{w}^\star_n) - \overline{\Phi}(c)\right| \xrightarrow[n\to+\infty]{\mathbb{P}} 0,$$

where $\overline{\Phi}(c) := \frac{1}{\sqrt{2\pi}}\int_c^{+\infty} e^{-\frac{u^2}{2}}\,du$.

We conclude by stating our last result, valid for the SRW on a directed graph exhibiting a community structure. For an integer $m > 1$, and two reals $\lambda > 1$, and $\alpha \equiv \alpha_n \le \frac{m-1}{m}$, take m communities $(V_j)_{j\le m}$ with the same set $[n]$ of vertex-labels, and let $V = \bigcup_{j\le m} V_j$. Consider the directed graph $\mathrm{DBM}(n, m, p, \alpha)$ defined as follows:

(a) for each $j = 1, \dots, m$ let G_j be Erdős-Rényi digraph on V_j with $p = \lambda\frac{\log(n)}{n}$;

(b) for each edge (x, y) in some G_j with probability α: *(i)* erase (x, y), *(ii)* select u.a.r. in $[m]\setminus\{j\}$ a number k, and *(iii)* draw (x, z), where $z \in V_k$ has the same vertex-label as y.

The following theorem, proved in [6], describes a mixing trichotomy, namely three distinct mixing behaviours. It shows that, depending on the order of the parameter α, the SRW exhibits a cutoff at the time $t^\star_n = \frac{\log n}{\log\log n}$ or an abrupt partial relaxation at $t^\star_n = \frac{\log n}{\log\log n}$, followed by an exponential relaxation to the equilibrium on a possibly larger time–scale. This analysis enriches the results given in [3] and [13] for random walks on undirected graphs.

Theorem ([6]) *Let* G *be a realization of the random digraph* $\mathrm{DBM}(n, m, p, \alpha)$*, and* $t^\star_n = \frac{\log n}{\log\log n}$*. The following mixing trichotomy takes place.*

- ***Subcritical case (Fig. 5):*** *if* $\alpha^{-1} \ll t^\star_n$ *and* $\alpha \le \frac{m-1}{m}$*, then, for all* $\beta > 0$ *with* $\beta \ne 1$*,*

$$\max_{x\in V} |\|\mathbf{P}^G{}_x(X_{\beta t^\star_n} \in \cdot) - \pi\|_{\mathrm{TV}} - \mathbf{1}_{\{\beta<1\}}| \xrightarrow[n\to+\infty]{\mathbb{P}} 0\,.$$

- ***Critical case (Fig. 5):*** *if* $\alpha^{-1} \sim Ct^\star_n$ *for* $C > 0$*, then, for all* $\beta > 0$ *with* $\beta \ne 1$*,*

$$\max_{x\in V}\left|\|\mathbf{P}^G{}_x(X_{\beta t^\star_n} \in \cdot) - \pi\|_{\mathrm{TV}} - \mathbf{1}_{\{\beta<1\}} - \tfrac{m-1}{m}\mathrm{e}^{-\frac{\beta}{C}\frac{m}{m-1}}\mathbf{1}_{\{\beta>1\}}\right| \xrightarrow[n\to+\infty]{\mathbb{P}} 0\,.$$

- ***Supercritical case (Fig. 6):*** *if* $\alpha^{-1} \gg t^\star_n$ *and* $\alpha^{-1} \ll \lambda n\log(n)$*, then*
  - *(local equilibrium at* $t^\star_n$*) for any* $\beta \ne 1$*,*

$$\max_{x\in V}\left|\|\mathbf{P}^G{}_x(X_{\beta t^\star_n} \in \cdot) - \pi\|_{\mathrm{TV}} - \mathbf{1}_{\{\beta<1\}} - \tfrac{m-1}{m}\mathbf{1}_{\{\beta>1\}}\right| \xrightarrow[n\to+\infty]{\mathbb{P}} 0\,,$$

– *(whole mixing at α^{-1}) for any $\beta > 0$,*

$$\max_{x\in V}\left|\|\mathbf{P}^G{}_x(X_{\beta\alpha^{-1}}\in\cdot)-\pi\|_{\mathrm{TV}}-\tfrac{m-1}{m}\mathrm{e}^{-\frac{\beta m}{m-1}}\right|\xrightarrow[n\to+\infty]{\mathbb{P}}0\,.$$

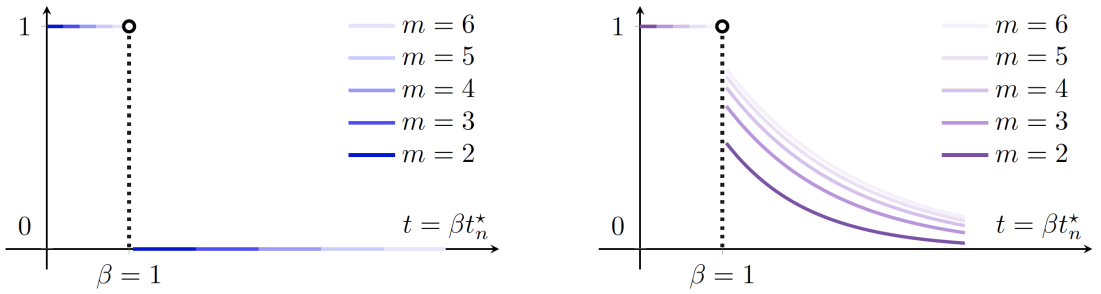


Figure 5: Plot of the (theoretical) limiting mixing profile in the subcritical case (left) and critical case (right) with $C=2$ and $m=2,3,4,5,6$.

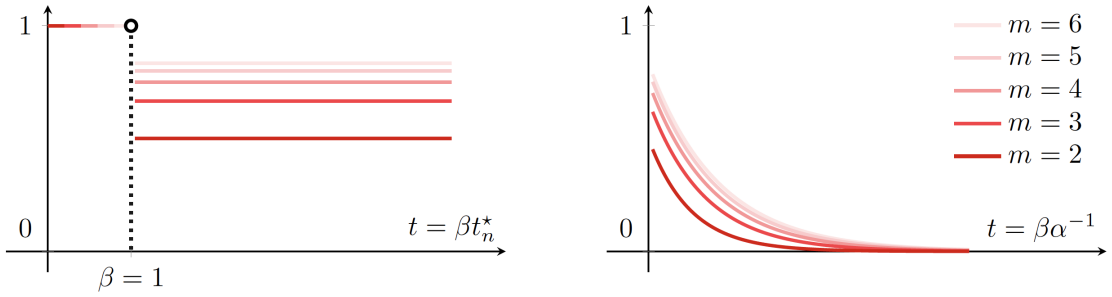


Figure 6: Plot of the (theoretical) limiting mixing profile in the supercritical case, with $m=2,3,4,5,6$, in the two timescales $t\asymp t_n^\star$ (left) and $t\asymp\alpha^{-1}$ (right).

References

[1] D. J. Aldous and P. W. Diaconis, *Shuffling cards and stopping times.* Amer. Math. Monthly 93(5): 333–348, 1986.

[2] D. Bayer and P. W. Diaconis, *Trailing the dovetail shuffle to its lair.* Ann. Appl. Probab. 2(2): 294–313, 1992.

[3] A. Ben-Hamou, *A threshold for cutoff in two-community random graphs.* Ann. Appl. Probab. 30(4): 1824–1846, 2020.

[4] A. Ben-Hamou, J. Salez, *Cutoff for nonbacktracking random walks on sparse random graphs.* Ann. Probab. 45(3): 1752–1770, 2017.

[5] A. Bianchi, G. Passuello, *Mixing cutoff for simple random walks on the Chung-Lu digraph.* Random Structures Algorithms 66(1): 1–20, 2025.

[6] A. Bianchi, G. Passuello, M. Quattropani, *Mixing Trichotomy for random walks on directed stochastic block models.* Preprint 2025. *arXiv:2504.06851*.

[7] C. Bordenave, P. Caputo, J. Salez, *Random walk on sparse random digraphs.* Probab. Theory Relat. Fields 170(3): 933–960, 2018.

[8] C. Bordenave, P. Caputo, J. Salez, *Cutoff at the "entropic time" for sparse markov chains.* Probab. Theory Relat. Fields 173(1): 261–292, 2019.

[9] X. S. Cai, P. Caputo, G. Perarnau, M. Quattropani, *Rankings in directed configuration models with heavy tailed in-degrees.* Ann. Appl. Probab. 33 (6B): 5613–5667, 2023.

[10] C. Cooper, A. Frieze, *Stationary distribution and cover time of random walks on random di-graphs.* J. Comb. Theory Ser. B 102(2): 329–362, 2012.

[11] P. W. Diaconis and M. M. Shahshahani, *Generating a random permutation with random transpositions.* Z. Wahrsch. Verw. Gebiete 57(2): 159–179, 1981.

[12] P. Diaconis, *The cutoff phenomenon in finite Markov chains.* Proc. Natl. Acad. Sci. USA 93(4): 1659–1664, 1996.

[13] J. Hermon, A. Šarković, P. Sousi, *Cutoff for random walk on random graphs with a community structure.* To appear in Ann. Appl. Probab. .

[14] R. van der Hofstad, "Random Graphs and Complex Networks". Cambridge University Press, 2016.

[15] Achim Klenke, "Probability Theory: A Comprehensive Course". Springer, London, 2014.

[16] D. A. Levin and Yuval Peres, "Markov Chains and Mixing Times". American Mathematical Society, Providence, RI, 2017. Second edition. With contributions by Elizabeth L. Wilmer, With a chapter on "Coupling from the past" by James G. Propp and David B. Wilson.

[17] E. Lubetzky, A. Sly, *Cutoff phenomena for random walks on random regular graphs.* Duke Math. J. 153(3): 475–510, 2010.

Meanfield Turnpike Theorems

Denis Shishmintsev (*)

1 Context and Motivation

Mean Field Games (MFG) study systems with a large number of agents interacting indirectly through the overall distribution of states. Instead of modeling each interaction, MFG theory focuses on the behaviour of a representative agent influenced by a "mean field". One key phenomenon in such systems is the turnpike property: when the time horizon is large, optimal trajectories and controls stay close to a steady state for most of the interval. The problem itself is well-studied and investigated, see for example, classical papers of Trélat et. al [4] and [3]. But, these works were developed only for local type of optimal control problems, i.e., problems which does not involve distributions to the dynamic. The most difficult to make a step further was in lac of extension of standard local Pontryagin's Maximum Principle (PMP) to the non-local case. This question was closed by Averboukh et. al in [1]. In this work, based on the observed literature, we:

- Generalizes the turnpike property to **nonlocal dynamics and cost functions**.
- Considers both **Eulerian** (distributional) and **Lagrangian** (trajectory-based) formulations.
- Proves **exponential turnpike estimates** for both formulations.

2 Problem formulation

2.1 Dynamic Lagrangian and Eulerian settings

In this section we will give two different formulations of mean field optimal control problems. Following the terminology introduced in [2], we describe, first, the so-called Lagrangian formulation. To this end, we consider a standard atomless probability space

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: shishmin@math.unipd.it . Seminar held on 20 February 2025.

$(\Omega, \mathcal{F}, \mathbb{P})$. Then, the optimal control problem is given by

$$(\mathcal{P}_L) \begin{cases} \min\limits_{u \in \mathcal{U}_L} \displaystyle\int_\Omega \left(\int_0^T L\Big(X(t,\omega), X(t)_\sharp \mathbb{P}, u(t,\omega)\Big) \mathrm{d}t + \varphi\Big(X(T,\omega), X(T)_\sharp \mathbb{P}\Big) \right) d\mathbb{P}(\omega) \\ \text{s.t.} \quad \begin{cases} \dot{X}(t,\omega) = v(X(t,\omega), X(t)_\sharp \mathbb{P}, u(t,\omega)), \\ X(0) = X_0. \end{cases} \end{cases}$$

where $u \in \mathcal{U}_L$ and $\mathcal{U}_L := \big\{[0,T] \ni t \mapsto u(t) \in L^2(\Omega, U) : u(t) \text{ is measurable}\big\}$, $v : \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d) \times U \to \mathbb{R}^d$, $L : \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d) \times U \to \mathbb{R}$ and $\varphi : \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$. The dependence of the data on the measure $X(t)_\sharp \mathbb{P}$ makes the problem nonlocal and models the interaction among particles and/or the interaction of the mass with the surrounding environment and it is usually referred to as a mean field interaction [2]. The couple (X, u) will be referred to as a Lagrangian process.

Next we describe the Eulerian form of our optimal control problem. Here, the system evolves according to a curve of probability measures, which evolution is driven by a nonlocal continuity equation. The problem reads as:

$$(\mathcal{P}_E) \begin{cases} \min\limits_{u \in \mathcal{U}_E} \Big[\displaystyle\int_0^T \int_{\mathbb{R}^d} L(x, \mu(t), u(t,x)) d\mu(t)(x) dt + \int_{\mathbb{R}^d} \varphi(x, \mu(T)) d\mu(T)(x) \Big], \\ \text{s.t.} \quad \begin{cases} \partial_t \mu(t) + \mathrm{div}_x \Big(v(\mu(t), u(t,x)) \mu(t) \Big) = 0, \\ \mu(0) = \mu^0. \end{cases} \end{cases}$$

where $u \in \mathcal{U}_E$ and $\mathcal{U}_E := \big\{[0,T] \ni t \mapsto u(t) \in L^2(\mathbb{R}^d, U) : u(t) \text{ is measurable}\big\}$, $L : \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d) \times U \to \mathbb{R}$, $\varphi : \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ and $v : \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d) \times U \to \mathbb{R}^d$. The couple (μ, u) will be referred to as an Eulerian process.

2.2 Static Problem and Turnpike Setting

As $T \to \infty$, the dynamic problems approximate the corresponding dynamic problems, under suitable regularity assumptions, could be approximated their static analogous:

$$(\mathcal{P}_{L_S}) \begin{cases} \min\limits_{(X,u) \in L^2(\Omega, \mathbb{R}^d) \times L^2(\Omega, U)} \displaystyle\int_\Omega L(X, X_\sharp \mathbb{P}, u) d\mathbb{P}(\omega) \\ \text{s.t.} \quad v(X, X_\sharp \mathbb{P}, u) = 0 \quad \mathbb{P}\text{-a.e.} \end{cases}$$

$$(\mathcal{P}_{E_S}) \begin{cases} \min\limits_{(\mu,u) \in \mathcal{P}_2(\mathbb{R}^d) \times L^2(\mathbb{R}^d, U)} \displaystyle\int_{\mathbb{R}^d} L(x, \mu, u) d\mu(x) \\ \text{s.t.} \quad v(\mu, u) = 0 \quad \mu\text{-a.e.} \end{cases}$$

3 Assumptions

The following assumptions are used throughout the analysis:

(H1) The control set U is a Hilbert space.

(H2) There exists a constant C_∞ such that for all $x \in \mathbb{R}^d$, $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $u \in U$:

$$|L(x,\mu,u)| \leq C_\infty \left(1 + \|x\|^2 + M_2^2(\mu) + \|u\|^2\right)$$

(H3) L is continuously differentiable in x and μ, and

$$\|\nabla_x L(x,\mu,u)\|^2 \leq C_{ox}(1 + \|x\|^2 + M_2^2(\mu) + \|u\|^2)$$

$$\|\nabla_\mu L(x,y,\mu,u)\|^2 \leq C_{o\mu}(1 + \|x\|^2 + \|y\|^2 + M_2^2(\mu) + \|u\|^2)$$

(H4) The functional L is convex in u.

(H5) The terminal cost ϕ is continuously differentiable in x, Fréchet differentiable in μ, and the derivatives satisfy:

$$\|\nabla_x \phi(x,\mu)\|^2 \leq C_{\phi x}(1 + \|x\|^2 + M_2^2(\mu))$$

$$\|\nabla_\mu \phi\|^2 \leq C_{\phi\mu}(1 + \|x\|^2 + \|y\|^2 + M_2^2(\mu))$$

(H6) The velocity field $v(x,\mu,u)$ is affine in u.

(H7) v is continuously differentiable in x and Fréchet differentiable in μ, with bounded derivatives:

$$\|D_x v\|_{L^2} \leq C_x, \quad \|D_\mu v\|_{L^2} \leq C_\mu$$

4 Main Results

Lemma 4.1 (Lipschitz continuity of a nonlocal velocity field) *Let $v : L^2(\Omega,\mathbb{R}^d) \times \mathcal{P}_2(\mathbb{R}^d) \times L^2(\Omega,U) \to \mathbb{R}^d$ satisfy (H7). Then v is Lipschitz with respect to $X \in L^2(\Omega,\mathbb{R}^d)$, that is:*

$$\|v(X', X'_\# P, u(t)) - v(X, X_\# P, u(t))\| \leq L\|X' - X\|_{L^2(\Omega,\mathbb{R}^d)}$$

Proof. By assumption (H7), the map $x \mapsto v(x,\mu,u)$ is differentiable with bounded derivatives, and similarly for $\mu \mapsto v$. Since the map $X \mapsto X_\# P$ is Lipschitz from $L^2(\Omega) \to \mathcal{P}_2(\mathbb{R}^d)$, the overall composition is Lipschitz in X. Estimate is obtained by applying chain rule and bounding each term by its respective Lipschitz constant. □

Lemma 4.2 (Gronwall's Lemma in nonlocal setting) *Let $X(t), X'(t)$ be solutions of the ODE:*

$$\dot{X}(t) = v(X(t), X(t)_\# P, u(t)), \quad X(0) = X_0, \quad X'(0) = X_0 + \delta x(0)$$

Then:

$$\|X'(t) - X(t)\| \leq e^{Ct}\|\delta x(0)\|_{L^2(\Omega,\mathbb{R}^d)}$$

Proof. Let $\delta X(t) = X'(t) - X(t)$. Then:

$$\frac{d}{dt}\delta x(t) = v(X'(t), X'(t)_{\#}P, u(t)) - v(X(t), X(t)_{\#}P, u(t))$$

By Lemma 4.1, the r.h.s. is Lipschitz in $\delta X(t)$, yielding:

$$\|\delta x(t)\| \leq \|\delta x(0)\| \cdot e^{Lt}$$

by Gronwall's inequality. □

Theorem 4.3 (Exponential turnpike property in nonlocal Lagrangian formulation) *Assume (H1)-(H3) and (H5). Let (X^*, u^*) be a solution of the static problem and let Ψ^* be a multiplier. Assume:*

- *H_{uu} is positive definite at (X^*, Ψ^*, u^*),*
- *Hessian components of the Hamiltonian are bounded,*
- *The pair $(A, H_{\Psi u})$ is exponentially stabilizable,*
- *The pair (A, C) is exponentially detectable.*

Then there exist $\varepsilon, \mu, c > 0$ such that for any $T > 0$, if:

$$\|X(0) - X^*\| + \|\Psi(T) - \Psi^*\| \leq \varepsilon$$

then:

$$\|X(t) - X^*\| + \|\Psi(t) - \Psi^*\| + \|u(t) - u^*\| \leq c(e^{-\mu t} + e^{-\mu(T-t)})$$

Proof.

1. **Pontryagin System:** Write the necessary optimality conditions as a forward-backward system using the Hamiltonian H.
2. **Linearization:** Linearize the Pontryagin system around the steady state (X^*, Ψ^*, u^*) and express the system as:

$$\frac{d}{dt}\begin{pmatrix}\delta x \\ \delta \psi\end{pmatrix} = \begin{pmatrix}A & B \\ C^*C & -A^*\end{pmatrix}\begin{pmatrix}\delta x \\ \delta \psi\end{pmatrix}$$

   where $B = -H_{\Psi u}H_{uu}^{-1}H_{u\Psi}$, and $C^*C = H_{Xu}H_{uu}^{-1}H_{uX} - H_{XX}$.

3. **Riccati Equation:** Solve the algebraic Riccati equation associated to this system. Under the stabilizability and detectability assumptions, there exists a bounded positive definite operator P such that the feedback law stabilizes the forward component.
4. **Decoupling Transformation:** Use a transformation based on the Riccati solution to decouple the system into a stable forward and a stable backward system.

5. **Estimate Deviations:** Apply exponential semigroup estimates and a Gronwall-type lemma to bound the deviations from the steady state, concluding exponential proximity.

□

Theorem 4.4 (Exponential turnpike property in Eulerian formulation) *Under the same assumptions as above, and using the equivalence between Eulerian and Lagrangian problems, for any optimal solution* $(\mu(t), u(t))$ *with steady state* (μ^*, u^*)*, there exist constants* $c, \mu > 0$ *such that:*

$$W_2(\mu(t), \mu^*) + \|u(t) - u^*\| \leq c(e^{-\mu t} + e^{-\mu(T-t)}), \quad \forall t \in [0, T]$$

Proof. 1. **Lift to Lagrangian:** Use the transformation from Eulerian to Lagrangian formulations to convert the Eulerian problem into a Lagrangian one.

2. **Apply Theorem 4.3:** Apply the exponential turnpike result already proven in the Lagrangian setting.

3. **Pushforward Estimate:** Pushing the result back to the Eulerian framework using the stability of the pushforward operation in Wasserstein space, conclude that turnpike property holds.

□

5 Example

Let us consider the model problem of nonlocal linear-quadratic (LQ) regulator in Lagrangian formulation. Precisely, take a standard probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We formulate the problem on $\mathbb{R}^d$ and the dynamic of an agent is given by the equation

$$\begin{aligned} \frac{d}{dt}X(t,\omega) &= A(\omega)X(t,\omega) + B(\omega)u(t,\omega), \\ X(0,\omega) &= X_0(\omega) \end{aligned} \tag{1}$$

and the payoff functional is defined as:

$$\begin{aligned} &\frac{1}{2}\mathbb{E}_{\sim\mathbb{P}}\left(\int_0^T \left[X^T(t)Q_xX(t) + u^T(t)Ru(t)\right]dt + X^T(T)G_xX(T)\right) \\ &+\frac{1}{2}\int_0^T \mathbb{E}_{\sim\mathbb{P}}\left[(X(t) - \mathbb{E}_{\sim\mathbb{P}}X(t))^T Q_\mu (X(t) - \mathbb{E}_{\sim\mathbb{P}}X(t))\right]dt \\ &\frac{1}{2}\mathbb{E}_{\sim\mathbb{P}}\left[(X(T) - \mathbb{E}X(T))^T G_\mu (X(T) - \mathbb{E}_{\sim\mathbb{P}}X(T))\right], \end{aligned} \tag{2}$$

where $u(t,\omega) \in U = \mathbb{R}^m$, for all $t \in [0,T]$ and for a.e. $\omega \in \Omega$. Let, $A, Q_x, Q_\mu, G_x, G_\mu$ be $(d \times d)$ bounded matrices, B be a $(d \times m)$ bounded matrix and R be a $(m \times m)$ symmetric bounded matrix. Moreover, the matrices Q_x, Q_μ, G_x, G_μ are symmetric positive semi-definite, while R is positive definite. Let $X(t)$ be the minimizer for the problem (1)-(2) and Ψ_L be the adjoint vector with the transversality condition

$$\Psi(T,\omega) = -(G_x + G_\mu)X(T,\omega) + G_\mu \left(\int_\Omega X(T,\eta) d\mathbb{P}(\eta) \right).$$

Then, the corresponding static problem reads as

(3)
$$\min_{\substack{(X,u)\in L^2(\Omega,\mathbb{R}^d)\times L^2(\Omega,U) \\ \text{s.t. } AX+Bu=0}} \frac{1}{2}\mathbb{E}_{\sim\mathbb{P}}\left(X^T Q_x X + u^T R u + X^T Q_\mu X - \left(\int_\Omega X(\omega)^T d\mathbb{P}(\omega) \right) Q_\mu \left(\int_\Omega X(\omega) d\mathbb{P}(\omega) \right) \right).$$

Using convex nature of the corresponding static problem, we obtain global turnpike property result in LQ type problems.

References

[1] Yurii Averboukh and Dmitry Khlopin, *Pontryagin maximum principle for the deterministic mean field type optimal control problem via the lagrangian approach.* ArXiv preprint `arXiv:2207.01892` (2022).

[2] Giulia Cavagnari, Stefano Lisini, Carlo Orrieri, and Giuseppe Savaré, *Lagrangian, eulerian and Kantorovich formulations of multi-agent optimal control problems: equivalence and gamma-convergence.* Journal of Differential Equations 322 (2022), 268–364.

[3] Emmanuel Trélat and Can Zhang, *Integral and measure-turnpike properties for infinite-dimensional optimal control systems.* Mathematics of Control, Signals, and Systems, 30 (2018), 1–34.

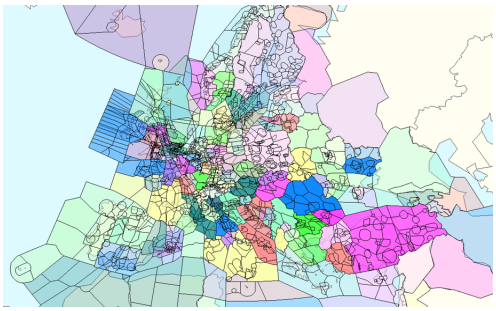
[4] Emmanuel Trélat, Can Zhang, and Enrique Zuazua, *Steady-state and periodic exponential turnpike property for optimal control problems in Hilbert spaces.* SIAM Journal on Control and Optimization 56/2 (2018), 1222–1252.

An Integer Linear Programming Model for the Dynamic Airspace Configuration problem

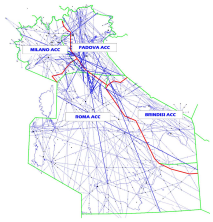
Martina Galeazzo (*)

1 Introduction and Literature review

Aviation is one of the most global industries, because of its power to connect people, cultures, and businesses across continents. In fact, it provides the only rapid worldwide transportation network, making it essential for global business. Aviation generates economic growth, creates jobs, and is a key factor in international trade and tourism. According to recent estimates by the Air Transport Action Group (ATAG), the total economic impact (direct, indirect, induced, and tourism catalytic) of the European aviation industry has reached USD 823 billion [8]. In order for Air Navigation Service (ANS) providers to monitor traffic safely and efficiently, airspace has to be functionally partitioned into control units, with respect to traffic density.



(a) European Airspace



(b) Italian Airspace and its ACCs

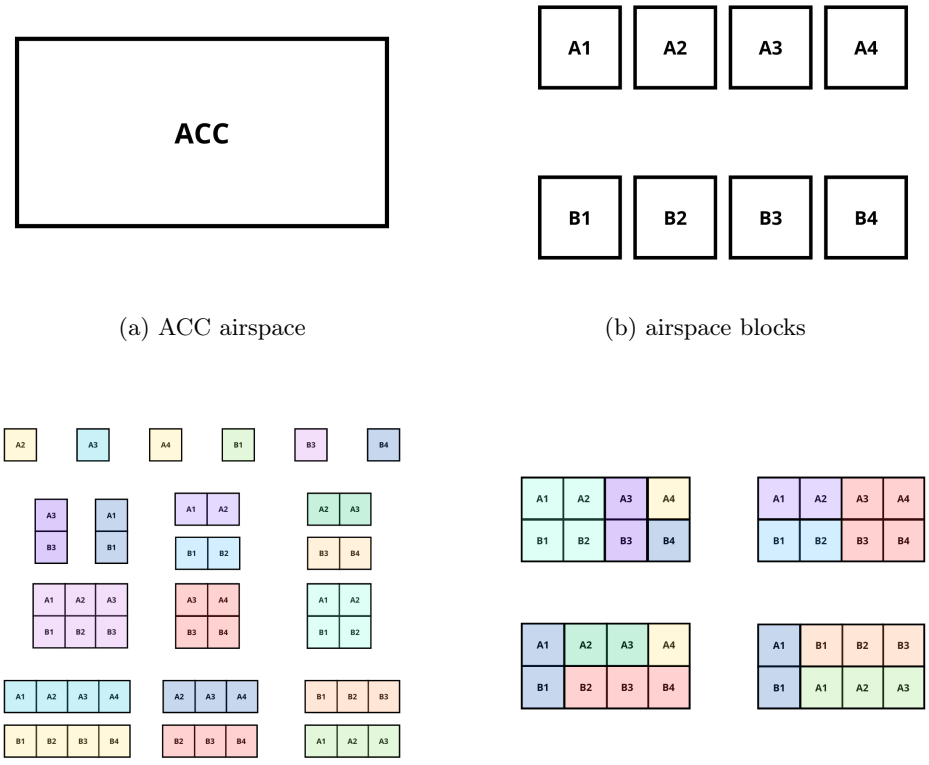
Figure 1

As a first-level airspace partitioning, we consider Area Control Centers (ACCs), which are autonomous with respect to traffic management; each ACC is divided into the aforementioned control units, called sectors, whose shapes and sizes vary in order to accommo-

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: galeazzo@math.unipd.it . Seminar held on 6 March 2025.

date the air traffic evolution; in Figure 1, as an example, we can see the partition of the European airspace into sectors and the four ACCs of the Italian airspace.

Since air traffic is human-managed, a congested area has to be assigned to multiple controllers and, therefore, is split into different sectors, as higher traffic concentrations require more control resources. The same area, under milder traffic conditions, would be partitioned into fewer sectors, or constitute a sector itself. To each sector corresponds a quantity called capacity, that provides a measure of the traffic volume that can be handled in that sector while maintaining a high level of safety. As sector capacity is limited, it is essential to efficiently manage airspace in order to maximise the total traffic volume it can absorb. For the purpose of our study, we consider the airspace model based on the concept of airspace block: an airspace block is a 3D portion of the airspace, and, in this perspective, a sector is a 3D connected union of one or more airspace blocks; moreover, an airspace configuration is a partition of the airspace into disjoint sectors; in Figure 2 we provide a 2D example of the structure we just described.



(a) ACC airspace

(b) airspace blocks

(c) sectors

(d) configurations

Figure 2: Example of airspace structure and partitioning.

In the last decades, a considerable effort has been put into the study and development of methods to effectively manage airspace; therefore, the literature on this subject is wide and multifaceted. Although, at this point, the specific terminology is quite established, some terms appear to be used interchangeably; for this reason, we begin by providing the definitions we refer to from now on. In particular, we focus on the difference between Airspace Configuration (AC) and Airspace Sectorization (AS), following the definitions provided in [7]. In this perspective, AC involves rearranging predefined portions of the airspace (airspace blocks) into sectors to obtain a configuration, whereas AS does not rely on these established and commonly used airspace elements, allowing a more flexible partition of the airspace. Our primary focus lies in the dynamic counterpart of AC (Dynamic Airspace Configuration, or DAC), which, adhering to the principles of AC, generates a sequence of configurations (configuration plan) to be deployed in a given time-frame.

The literature on DAC is quite rich in terms of proposed approaches and techniques; for instance, in [11], DAC is tackled as a graph partitioning problem: the authors consider a graph where nodes represent airspace blocks and arcs connect spatially adjacent blocks, and solve the problem using a genetic algorithm. In [4], the authors propose a recursive greedy algorithm that, starting from the current airspace sectors, aims to combine under-utilized ones, based on a measure of the predicted traffic excess. Reference [1] also relies on a graph representation, where nodes correspond to airspace blocks, and arcs connect nodes if the associated blocks are connected by trajectories. The aim is to provide a smooth configuration plan that avoids abrupt sector changes, and it is achieved by employing two successive simulated annealing algorithms and running a shortest path algorithm. In [3], three algorithms are presented to find a configuration plan that minimizes workload cost: a myopic heuristic, an exact dynamic programming algorithm, and a rollout approximate dynamic programming algorithm. In [9], an Integer Programming model is introduced to assign each airspace block to a sector from a predefined list, therefore obtaining a configuration.

The foundation of this work is the assumption of the availability of a set of commonly used configurations in a given airspace. This approach has many advantages: the objects of controllers' training are usually configurations, therefore we can rely on a high degree of familiarity; we do not need to check for configuration feasibility, since it has already been practically proven; and we can easily impose operational constraints on the configurations' dynamics, such as the need to avoid the frequent configuration changes that can occur when trying to accommodate traffic peaks. We propose an Integer Programming model for DAC that, encompassing the aforementioned constraints on the dynamics, provides an optimal configuration plan based on traffic forecasts. In light of this, it is suitable for application during the tactical phase of operations to obtain a starting point for the configuration plan to be deployed the following day.

2 Problem description

Starting from a given family of airspace configurations, our goal is to compute a sequence of configurations that meets the (expected) air traffic demand as much as possible; in other words, we aim to minimize the total excess of air traffic demand in the time horizon we consider, which is finite and discretized into time periods. Configuration changes are only allowed at discrete times corresponding to such time periods; we refer to this partition of the time horizon as *decision discretisation* (d), since, in each of the time periods we consider, we decide which configuration to implement (active configuration). We also work under the assumption that the air traffic demand for each sector over time is known. As for the dynamics of the configuration plan we aim to construct, we require it to be "smooth", meaning that: (i) transitions between configurations that are very different from one another (e.g. they have few sectors in common) should be avoided, (ii) configurations must remain active for at least an "operational" time interval, to allow configuration setup and to ease the monitoring operations (permanence requirement), and (iii) the configuration plan should not oscillate between the same configurations in response to demand fluctuations (quiescence requirement). In fact, if left unchecked, the configuration plan may present frequent configuration switching and schedule the same configuration after a short time, to follow traffic variations; both these occurrences are very impractical from an operational point of view, so we implement ad hoc constraints to prevent them from happening. Towards (i), practitioners rely on the concept of "configuration compatibility" and only allow transitions between compatible configurations; moreover, they use long decision time intervals, therefore configuration permanence (and quiescence) is guaranteed as well. Since we aim to work with short decision time intervals, towards (ii) and (iii), we impose permanence constraints to ensure that an active configuration remains active for at least t_p time periods (permanence interval), and we constrain a minimum time interval (longer than the permanence one) between consecutive reactivation of a same configuration, that will be called quiescence interval and will consist of t_q consecutive decision time periods. The notation we will use is as follows:

- T is the set of decision time periods in which the time horizon is discretised, indexed from 0 to $|T| - 1$,
- C is the family of available airspace configurations,
- E_t^c is a parameter that measures the traffic overload, or excess, in configuration c at decision time period t.

Furthermore, with C^t we denote the set of configurations available in time period t. This models the fact that, due to operational requirements, a given configuration may be operated only at specific time intervals.

3 Modeling DAC on a directed graph

It is possible to model the DAC problem on a suitable directed, weighted graph $G = (V, A)$ where:

- $V = \{(c,t) \mid t \in T, c \in C^t\} \cup \{(\alpha,0),(\omega,|T|+1)\}$ is the set of nodes, corresponding to the configuration-time pairs, plus two dummy nodes acting as source and sink;
- $A = A^P \cup A^T \cup A^\alpha \cup A^\omega$ is the set of arcs, which includes permanence arcs (A^P), connecting nodes associated to the same configuration, transition arcs (A^T), between nodes corresponding to different configurations, and artificial arcs (A^α and A^ω), for which one of the endpoints is a dummy node; in this graph, arcs represent the feasible transitions between configurations over time.

In particular, A^P and A^T are defined as:

$$A^P = \{((c,t),(c,t+1)) \mid t \in T \setminus \{|T|-1\}, c \in C^t \cap C^{t+1}\}$$
$$A^T = \{((c,t),(d,t+1)) \mid t \in T \setminus \{|T|-1\}, c \in C^t, d \in C_c^{t+1}\}$$

Permanence arcs represent the possibility of maintaining a configuration active for consecutive time periods, while transition arcs correspond to transitions between different configurations. As for A^α and A^ω, they are defined as:

$$A^\alpha = \{((\alpha,-1),(c,0)) \mid c \in C^0\}$$
$$A^\omega = \{((c,|T|-1),(\omega,|T|)) \mid c \in C^0\}$$

We also define the weight function $w : A \to \mathbb{N}$ as follows:

$$w((b,t-1),(c,t)) = \begin{cases} E_t^c & \text{if} \quad c \neq \omega \\ 0 & \text{otherwise} \end{cases}$$

i.e., the weight of an arc corresponds to the excess of air traffic demand associated to its head; let us notice that all arcs entering one node have the same weight.

Remark 1 If we neglect permanence and quiescence constraints, solving the DAC problem is equivalent to finding the weighted shortest path from $(\alpha,-1)$ to $(\omega,|T|)$ on the directed graph we just described. Therefore, the DAC problem can be solved in polynomial time, with respect to the size of the graph, using well-known shortest path algorithms (e.g., the Bellman-Ford algorithm [2], which has a time complexity of $O(|V| \cdot |A|)$).

On the other hand, the inclusion of permanence and/or quiescence constraints requires the implementation of ad hoc label-setting algorithms to solve the problem; in [10], such an algorithm has been efficiently implemented and used in a robust optimization framework.

4 Mathematical Programming

In order to solve an optimization problem, one of the possible frameworks is Mathematical Programming, which relies on mathematical models to describe the features of the optimal solution by means of mathematical relations. The constitutive elements of a model are the following:

- sets, that group the elements of the problem,

- parameters, the known quantities representing the data of the problem,
- decision variables, whose optimal value has to be determined,
- constraints, mathematical expressions describing the conditions that feasible solutions have to satisfy,
- objective function, that is the function of the decision variables that has to be minimized or maximized.

Whenever the objective function is linear and the constraints are given by a linear system of equations and inequalities, we speak of *Linear Programming*; depending on the domain of the decision variables, we distinguish three cases:

- Linear Programming (LP), if all the variables can take real values,
- Integer Linear Programming (ILP), if all the variables can only take integer values,
- Mixed Integer Linear Programming (MILP), if some variables are constrained to be integer and the others can take real values.

The general structure of a MILP model is as follows:

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

where $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ describe the constraints coefficient, $c \in \mathbb{R}^n$ is the vector of the objective function coefficients, and $I \subset \{1, 2, \ldots, n\}$ is the index set of the integer variables.

4.1 Integer Linear Programming model for DAC

We now present an Integer Linear Programming model for the DAC problem under permanence and quiescence constraints, and we begin by introducing the two families of binary variables used in the formulation:

- x_t^c, for every $t \in T$ and $c \in C^t$, taking value 1 if configuration c is active at time period t, and 0 otherwise,
- s_t^c, for every $t \in T \setminus \{|T|-t_p+1, \ldots, |T|-1\}$ and $c \in C^t$, taking value 1 if configuration c is activated at time period t (i.e. c is active at time t, but not at time $t-1$), and 0 otherwise.

In the following, whenever variables x and s are not defined, they can be replaced by 0.

ILP model for DAC

$$\min \sum_{t\in T}\sum_{c\in C^t} E_t^c \cdot x_t^c + \varepsilon \sum_{t\in T}\sum_{c\in C^t} n^c \cdot x_t^c$$

s.t.

$$(1) \quad \sum_{c\in C^t} x_t^c = 1 \qquad \forall\, t \in T$$

$$(2) \quad x_t^c - \sum_{c'\in C_c^{t+1}} x_{t+1}^{c'} \le 0 \qquad \forall\, t \in T \setminus \{|T|-1\}, \quad \forall\, c \in C^t$$

$$(3) \quad \sum_{\tau=t}^{t+t_p-1}\sum_{c\in C^\tau} s_\tau^c \le 1 \qquad \forall\, t \in T \setminus \{|T|-t_p+1,\dots,|T|-1\}$$

$$(4) \quad x_t^c + \sum_{\tau=t+1}^{t+t_q} s_\tau^c \le 1 \qquad \forall\, t \in T \setminus \{|T|-1\}, \quad \forall\, c \in C^t$$

$$(5) \quad x_t^c - x_{t-1}^c \le s_t^c \qquad \forall\, t \in T \setminus \{0\}, \quad \forall\, c \in C^t$$

$$(6) \quad x_0^c \le s_0^c \qquad \forall\, c \in C^0$$

$$(7) \quad x_t^c \in \{0,1\} \qquad \forall\, t \in T, \quad \forall\, c \in C^t$$

$$(8) \quad s_t^c \in \{0,1\} \qquad \forall\, t \in T \setminus \{|T|-t_p+1,\dots,|T|-1\}, \quad \forall\, c \in C^t$$

The objective function consists of two terms: the first is the total traffic overload during the time frame, while the second is a penalization term (that will always be smaller than 1, thanks to the proper choice of ε) which ensures that, the traffic overload being the same, the configuration consisting of the smallest number of sectors is chosen; here n^c denotes the cardinality of configuration c. Constraint (1) imposes that exactly one configuration is active at each time interval, while constraint (2) takes care of the compatibility between configurations that are active in consecutive time periods; to this end, we introduce the set C_c^{t+1} comprising all configurations that can be implemented at time period $t+1$ if configuration c is active at time t. In order to avoid too frequent switching between different configurations, as we discussed before, constraints (3) (permanence constraints) impose that any active configuration remains active for at least t_p consecutive time periods. On the other hand, constraints (4) (quiescence constraint) are concerned with the amount of time that has to elapse before a configuration that has been deactivated can be activated again. In this sense, (4) imposes that a configuration c that is deactivated at time t cannot be reactivated within the consecutive t_q time periods. In light of the previous considerations on variable s, configuration permanence at the end of the time frame is guaranteed; in fact, the s variables are replaced by 0 in the last $t_p - 1$ time periods, hence no configuration can

be activated. Finally, constraints (5) and (6) link the values of variables x_t^c and s_t^c for every $t \in T$, ensuring that s_t^c takes value 1 if c is activated in t. In the case of a configuration $c \in C^t \setminus C^{t-1}$, variable x_{t-1}^c is not defined; therefore constraint (5) reads $x_t^c \leq s_t^c$.

5 Polyhedral Study

In this section, we present an analysis of the structure of the polytope P^{PQ}, corresponding to the linear relaxation of the set defined by constraints (1)-(8). In particular, our main goal is to prove that permanence and quiescence constraints are facet-defining for conv $\left(P_I^{PQ}\right)$, i.e. the convex hull of the integer points in P^{PQ}. To achieve this, we begin by computing the dimension of P^{PQ}; we then identify redundant permanence and quiescence constraints, and we conclude by presenting the proofs for the facet-defining inequalities.

5.1 Dimension of the polytope P^{PQ}

For the sake of clarity, we will work under the assumption that $C^t = C$ for every $t \in T$ and $C_c^{t+1} = C$ for every $t \in T \setminus \{|T| - 1\}$ and every $c \in C$, i.e. all configurations are always available and all possible transitions are feasible. In our simplified setting, the dimension n of the space in which variables x_t^c and s_t^c live is $|T| \cdot |C| + (|T| - t_p + 1) \cdot |C| = 2|T| \cdot |C| - (t_p - 1)\,|C|$. We begin by recalling the following result from literature.

Theorem 1 ([5]) *Let $P = \{x \in \mathbb{R}^n : Ax \leq b\}$ be a non-empty polyhedron. Then* $\dim(P) = n - \operatorname{rank}(A^=)$*, where $A^= x \leq b^=$ is the system comprising all implicit equalities of $Ax \leq b$, i.e. the inequalities that hold as equalities for every point in P.*

In light of Theorem 1, in order to determine the dimension of the polytope P^{PQ}, we need to identify the implicit equalities in our formulation and compute the rank of the corresponding coefficients matrix. The following Lemma will exhibit all the implicit inequalities.

Lemma 1 *The following results hold true:*

(a) *Constraint (3) is an implicit equality for $t = 0$;*

(b) *$s_t^c = 0$ for every $t \in \{1, \ldots, t_p - 1\}$ and every $c \in C$,*

(c) *constraint (5) is an implicit equality for $t \in \{1, \ldots, t_p - 1\} \cup \{|T| - t_p + 1, \ldots, |T| - 1\}$,*

(d) *constraint (6) is an implicit equality for every $c \in C$.*

We now move on to the computation of the rank of matrix $A^=$, whose entries are the coefficients of the implicit equalities. Each column of $A^=$ is associated with either an x variable or an s variable, thus there are $2|T| \cdot |C| - (t_p - 1)\,|C|$ columns. As for the rows, each one corresponding to an implicit equality, by summing up the previous considerations, we have:

- $|T|$ equalities coming from constraints (1),

- 1 implicit equality from Lemma 1a,
- $(t_p - 1)\,|C|$ implicit equalities from Lemma 1b,
- $2\,(t_p - 1)\,|C|$ implicit equalities from Lemma 1c
- $|C|$ implicit equalities from Lemma 1d.

Thus, $A^=$ has $|T| + (3t_p - 2)\,|C| + 1$ rows, and its entries are either 0 or 1.

Remark 2 By ordering the columns of $A^=$ according to ascending values of index t for the corresponding variables, we obtain the following block matrix:

$$A^= = \begin{pmatrix} B_{11} & 0 & 0 \\ 0 & B_{22} & 0 \\ 0 & 0 & B_{33} \end{pmatrix}$$

The only blocks with entries different from 0 are the ones along the diagonal, B_{11}, B_{22}, and B_{33}.

Example 1 We now show matrix $A^=$ for $|T| = 7$, $|C| = 2$ and $t_p = 3$.

	x_0^A	x_0^B	s_0^A	s_0^B	x_1^A	x_1^B	s_1^A	s_1^B	x_2^A	x_2^B	s_2^A	s_2^B	x_3^A	x_3^B	s_3^A	s_3^B	x_4^A	x_4^B	s_4^A	s_4^B	x_5^A	x_5^B	x_6^A	x_6^B	
α_0	1	1	0									0													
	0			0	1	1	0					0													
	0							0	1	1	0	0													
β_1^A	0							0	1	0		0													
β_1^B	0								0	1	0														
β_2^A	0										0	1	0												
β_2^B	0											0	1												
γ_0^A	1	0	−1	0								0													
γ_0^B	0	1	0	−1	0							0													
γ_1^A	−1	0	0	0	1	0	−1	0				0													
γ_1^B	0	−1	0	0	0	1	0	−1	0			0													
γ_2^A	0			0	−1	0	0	0	1	0	−1	0													
γ_2^B	0				0	−1	0	0	0	1	0	−1													
	0	0	1	1	0	0	1	1	0	0	1	1													
	0											0	1	1	0	0	0								
																	1	1	0						α_0
																	0			0	1	1	0	0	
																	0				0	1	1		
																	−1	0	0	0	1	0	0	0	β_1^A
																	0	−1	0	0	0	1	0	0	β_1^B
																	0			0	−1	0	1	0	β_2^A
																	0				0	−1	0	1	β_2^B

Remark 3 B_{22} has full rank, which is equal to $|T| - 2t_p$.

This can be easily verified by noticing that the rows in B_{22} are linearly independent, given that the sets of columns in which each of them presents a non-zero value are disjoint. Lemmas 2 and 3, on the other hand, are concerned with the rank of blocks B_{11} and B_{33}, respectively, and show that neither of them has full rank.

Lemma 2 $\operatorname{rank} B_{11} = 1 + (2t_p - 1)\,|C|$.

Lemma 3 $\operatorname{rank} B_{33} = 1 + (t_p - 1)\,|C|$.

Lemma 4 *In light of Remark 3 and Lemmas 2 and 3,* $\operatorname{rank} A^{=} = |T| + (3t_p - 2)\,|C| - 2\,(t_p - 1)$. *By Theorem 1, we can state:*

$$\dim(P) = 2|T| \cdot |C| - |T| + (3 - 4t_p)\,|C| + 2\,(t_p - 1) \tag{9}$$

We now move on to the identification of redundant permanence and quiescence constraints; we recall that a constraint is redundant if removing it from the system does not change the feasible region.

Lemma 5 *For* $t \in \{1, \dots, t_p - 1\} \cup \{|T| - t_p, \dots, |T| - 1\}$, *constraint (3) is redundant.*

Lemma 6 *For* $t \in \{0, \dots, t_p - 2\} \cup \{|T| - t_p - t_q + 1, \dots, |T| - 1\}$ *and for all* $c \in C$ *constraints (4) are redundant.*

5.2 Facet defining inequalities

The aim of this subsection is to show that, whenever they are not redundant, permanence and quiescence constraints are facet defining for $\operatorname{conv}\left(P_I^{PQ}\right)$; in the following we will assume $|C| \geq 2$. We begin by recalling the definition of face and facet of a polyhedron and a well-know characterization of facets.

Definition 1 ([5]) A face of a given polyhedron P is a set of the form:

$$F := P \cap \{x \in \mathbb{R}^n : cx = d\}$$

where $cx \leq d$ is a valid inequality for P. Inclusion-wise maximal proper faces of P are called facets.

Theorem 2 ([5]) *A face F of a polyhedron P is a facet if and only if F is nonempty and* $\dim(F) = \dim(P) - 1$.

In light of these definitions and of the results presented in the previous subsection, we can state the following theorems.

Theorem 3 *Whenever inequalities (4) are not redundant, they are facet defining for* $\operatorname{conv}\left(P_I^{PQ}\right)$.

Theorem 4 *Whenever inequalities (3) are not redundant, they are facet defining for* $\text{conv}\left(P_I^{PQ}\right)$.

6 Numerical Study

In this section, we present some numerical results obtained by testing our model on five days of study, in summer 2019; we consider real data on 166 available airspace configurations, built using 99 sectors, covering a central-southern region of Madrid ACC and obtained with the DAC framework presented in [9]. The number of sectors that each configuration consists of varies from 2 to 10. Moreover, we discretized the day in 288 five-minute intervals, indexed from 0 to 287, and considered traffic data sampled every five minutes; further details on time discretisation and the computation of parameter E_t^c can be found in [6], here we just highlight the importance of having a fine decision discretization (we choose to operate with five-minute intervals, while current practice usually relies on 20-minute intervals), and a *data discretization* δ (a measure of the frequency with which traffic data is sampled) that is finer than the decision discretization.

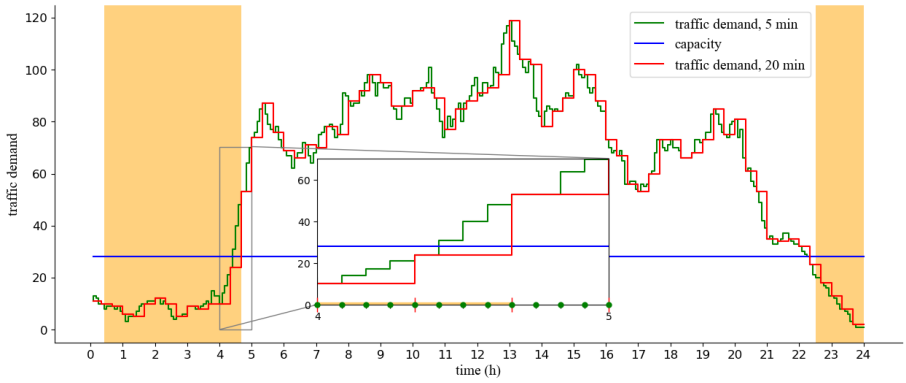


Figure 3: Comparison between the traffic demand and capacity of a sector for July 20$^{\text{th}}$ [6]

Figure 3 provides an example of the practical advantages of a finer decision discretization: it depicts the traffic demand with a 5 and 20-minute data discretization, and the capacity of a sector activated in the optimal solution for July 20$^{\text{th}}$, with a 20-minute decision discretization; the shaded orange portion of the graphic marks the time intervals in which the sector was active. For the most part of these intervals, the traffic demand of the sector is lower than its capacity; however, in the hour between 4:00 a.m. and 5:00 a.m., when the traffic demand (for both data discretizations) first exceeds the capacity, we notice a difference (which is magnified in the inset). Indeed, we observe a traffic peak at 4:25 a.m. for $\delta = 5$ that the coarser discretization does not capture until 4:40 a.m. As a result, at 4:20 a.m., which is a decision time for the 20-minute decision discretization (displayed with red bars in the inset), the sector remains active with an excess that is zero for $\delta = 20$, but the green line clearly shows that there are three 5-minute intervals in which

a traffic excess is registered. This preliminary analysis goes to show that low values of traffic excess obtained with a coarse data discretization do not necessarily correspond to non-challenging traffic conditions, underscoring the advantages of more refined approaches in capturing a more realistic image of the actual traffic conditions.

In the following, we assume that the minimum time that a configuration has to remain active to obtain a sequence of configurations that is compatible with the operational necessities of air traffic controllers is 20 minutes; therefore, for $d = 5$ parameter t_p is always set to a value of at least 4, meaning that a configuration has to remain active for four consecutive 5-minute intervals. Table 1 provides further insight into the advantages of fine decision and data discretization, comparing the result obtained with $d = \delta = 5$ and $d = \delta = 20$. We remark that we are working in the simplified setting with $C = C^t$ for every $t \in T$ and full compatibility between configurations.

	$d = 5$, $\delta = 5$			$d = 20$, $\delta = 20$		
Day	o.f.	time	s avg (max)	ovl	time	AP s avg (max)
20/07	0.0	7.13	4.38 (8)	179.0	1.05	4.75 (6)
21/07	0.0	6.38	4.33 (8)	158.0	1.08	4.71 (6)
22/07	0.0	7.31	4.36 (7)	213.0	1.06	4.83 (6)
25/07	0.0	5.67	4.14 (7)	111.0	1.19	4.77 (6)
04/08	0.0	7.86	4.23 (8)	212.0	1.20	4.79 (6)

Table 1

Columns *time* report on the computational time, while column *o.f.* presents the value of the objective function for $d = \delta = 5$; such column is replaced by column *ovl* (overload) for $d = \delta = 20$, where by overload we mean the total traffic excess obtained by considering the optimal solution for the rougher discretization and computing the corresponding excess for $\delta = 5$. As we can notice, this leads to overloads that are much higher than the results obtained with finer discretization and the nominal zero excess given by the objective function for $d = \delta = 20$ (not reported in the table). This phenomenon is consistent with what we observed in Figure 3, and is further displayed in Figure 4, that clearly shows multiple time intervals in which the traffic demand for $\delta = 5$ (red line) exceeds the total capacity of the active configuration (blue line). As for columns *s avg (max)*, they show the average and maximum cardinality of the active configurations obtained by our model (plain column) and associated with the plans that were actually deployed in the days we consider (column AP); by comparing them we notice that the average value provided by our model is lower, proving the effectiveness of the cardinality penalization term in the objective function. As far as the maximum values are concerned, let us remark that the cardinality of the configurations considered for the actual plans ranged from 2 to 6, while we also considered configurations with up to ten sectors.

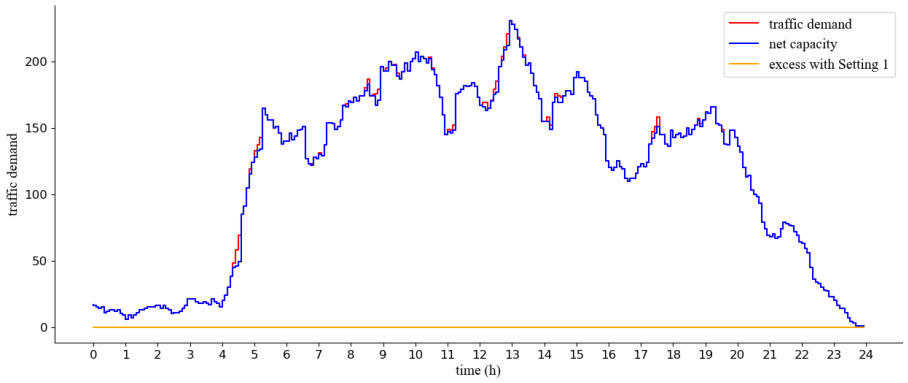


Figure 4: Comparison between traffic demand and capacity for the optimal solution with $d = \delta = 20$ on July 20^{th} [6]

At this point, with a view to increasing the degree of realism of our instances, we deem as feasible only the transitions between similar configurations, i.e., those that have at least 50% of sectors in common and have close cardinalities, and divide the day into time slots during which different sets of configurations are available. Tables 2 and 3 present results obtained by setting a permanence interval of 20 minutes and quiescence intervals of 45 and 60 minutes, respectively. Columns marked with *permanence* show the average, maximum, and minimum number of consecutive time intervals of activation for the configurations deployed in the optimal solution, while columns *cardinality* report the size of the active configurations.

			permanence			cardinality		
Day	o.f.	time	avg	max	min	avg	max	min
20/07/2019	4.0	22.28	10.67	56	4	5.85	10	2
21/07/2019	9.0	28.50	10.67	39	4	5.69	10	2
22/07/2019	5.0	39.19	7.78	26	4	5.72	9	2
25/07/2019	0.0	22.69	9.29	55	4	5.78	9	2
04/08/2019	0.0	20.22	9.29	28	4	5.59	9	2

Table 2: Results with $t_p = 4$ and $t_q = 9$, with limitations on the configurations cardinality

			permanence			cardinality		
Day	o.f.	time	avg	max	min	avg	max	min
20/07/2019	4.0	30.00	13.09	48	4	5.58	10	2
21/07/2019	9.0	34.42	7.78	38	4	5.29	10	2
22/07/2019	5.0	26.95	13.09	44	4	5.51	10	2
25/07/2019	0.0	21.47	9.60	46	4	5.57	9	2
04/08/2019	0.0	18.45	9.93	51	4	5.62	10	2

Table 3: Results with $t_p = 4$ and $t_q = 12$, with limitations on the configurations cardinality

By comparing the two tables, we notice that increasing the quiescence interval does not seem to have an impact on the total traffic excess, while it results in a mild increase of the maximum cardinality of an active configuration. As for the minimum cardinality, it remains equal to 2 in all cases; such small configurations are active during the night and early morning, when traffic demand is low, and they tend to remain active for longer periods. The opposite behavior can be observed during traffic peaks, that typically occur in the middle of the day; during these periods, larger configurations are deployed to accommodate the traffic volume, and such configurations are used for shorter periods of time, as we notice by looking at columns *permanence min*. The minimum duration of an activation interval always corresponds to 20 minutes, i.e. the minimum duration we impose by means of the permanence constraints.

7 Conclusions

After providing a general introduction on the airspace structure and on the different approaches to airspace management, we presented the Dynamic Airspace Configuration (DAC) problem and a graph representation encompassing its basic features and further regularity requirements on the configuration dynamics (permanence and quiescence constraints). We then moved on to Mathematical Programming, describing the features of an optimization problem, and proposed an Integer Linear Programming model for DAC.

We investigated the structure of the feasible region of such a model, computing the dimension of the convex hull of the integer points it contains and proving that permanence and quiescence constraints are facet defining for the convex hull.

We tested our model on the historical traffic data of five days in the summer of 2019 and on a set of configurations built using 99 sectors of Madrid ACC. We compared the performances of different data and decision discretization, highlighting the fact that using the finest data discretization, even when considering a rougher decision discretization, results in a configuration sequence that can better absorb the traffic peaks occurring in the middle of a decision interval.

Moreover, we increased the degree of realism of our instances by reducing the availability of the configurations during the day and by imposing stricter criteria on configurations compatibility. This resulted in a moderate increase in the total traffic excess registered during the day. We also remarked that increasing the duration of the quiescence interval does not affect the quality of the optimal configuration plan in any noticeable way.

References

[1] Judicaël Bedouet, Thomas Dubot, and Luis Basora, *Towards an operational sectorisation based on deterministic and stochastic partitioning algorithms.* In SESAR Innovation Days, 2016.

[2] Richard Bellman, *On a routing problem.* Quarterly of Applied Mathematics, 16: 87–90, 1958.

[3] Michael Bloem and Pramod Gupta, *Configuring airspace sectors with approximate dynamic programming.* In 27th International Congress of the Aeronautical Sciences (ICAS), 2010.

[4] Michael Bloem and Parimal Kopardekar, *Combining airspace sectors for the efficient use of air traffic control resources.* In AIAA Guidance, Navigation and Control Conference and Exhibit, 2008.

[5] Michele Conforti, Gérard Cornuéjols, and Giacomo Zambelli, "Integer Programming". Graduate Texts in Mathematics. Springer Cham, 1 edition, 2014.

[6] Martina Galeazzo, Luigi De Giovanni, M. Florencia Lema-Esposto, and Guglielmo Lulli, *An integer programming approach to dynamic airspace configuration.* Presented at the International Conference on Research in Air Transportation (ICRAT 2024), Singapore, July 2024.

[7] Ingrid Gerdes, Annette Temme, and Michael Schultz, *Dynamic airspace sectorisation for flight-centric operations.* Transportation Research Part C: Emerging Technologies, 95: 460–480, 2018.

[8] Air Transport Action Group, *Aviation benefits beyond borders.* 2018. (`https://atag.org/media/lggnx00h/abbb18_full-report_web-2.pdf`).

[9] M. Florencia Lema-Esposto, Manule Ángel Amaro-Carmona, Natividad Valle-Fernández, Enrique Iglesias-Martínez, and Adrián Fabio-Bracero, *Optimal dynamic airspace configuration (dac) based on state-task networks (stn).* In SESAR Innovation Days, 2021.

[10] Go Nam Lui, Guglielmo Lulli, Luigi De Giovanni, Martina Galeazzo, Iciar Garcia-Ovies Carro, and Rebeca Llorente Martinez, *A robust optimization approach for dynamic airspace configuration.* To be presented at the upcoming Air Transportation Research & Development Symposium (ATRD 2025), Prague, Czech Republic, June 2025.

[11] Marina Sergeeva, Daniel Delahaye, Catherine Mancel, and Andrija Vidosavljevic, *Dynamic airspace configuration by genetic algorithm.* Journal of Traffic and Transportation Engineering, 4(3): 300–314, 2017.

Resonances and Quasi-Collisions in the Three-Body Problem

Xiang Liu (*)

Abstract. Mean motion resonance, a phenomenon occurring when two celestial bodies have orbital periods in a commensurable ratio, plays a pivotal role in both stabilizing and destabilizing motions within our Solar System. For highly eccentric orbits, quasi-collisions become a significant factor. When such eccentric orbits are trapped in resonance, perturbations can induce chaotic motions, leading to rapid changes in orbital elements and transitions of different dynamical states. This presentation will begin by introducing the concept of mean motion resonance within the framework of the restricted three-body problem. Subsequently, we will explore the application of Hamiltonian perturbation theory for low-eccentricity orbits. Finally, we will demonstrate the limitations of this theory when applied to highly eccentric orbits.

1 Introduction

1.1 Restricted Three-Body Problem

The restricted three-body problem is composed of the primary body P_0 (the Sun), the secondary body P_1 (a planet) and a massless body P, where the primary P_0 and the secondary P_1 comprise a two-body system and the motion of P is affected by the gravitational force of P_0 and P_1. In our case we only consider the circular restricted three-body problem (CR3BP), of which both P_0 and P_1 rotate circularly with respect to the barycenter of P_0 and P_1.

To study the motion of P we introduce the heliocentric coordinates

$$\begin{aligned}\mathbf{r}_1 &= P_1 - P_0,\\ \mathbf{r} &= P - P_0.\end{aligned}$$

Then the equations of motion of P in heliocentric coordinates are

$$\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} = -Gm_0\frac{\mathbf{r}}{\|\mathbf{r}\|^3} + Gm_1\left(\frac{\mathbf{r}_1-\mathbf{r}}{\|\mathbf{r}_1-\mathbf{r}\|^3} - \frac{\mathbf{r}_1}{\|\mathbf{r}_1\|^3}\right),$$

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: liu@math.unipd.it . Seminar held on 20 March 2025.

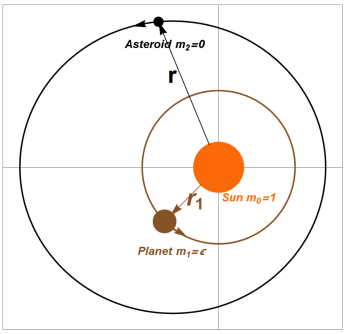


Figure 1: Restricted three-body problem.

where $G = 1$ is the gravitational constant, $m_0 = 1, m_1 = \varepsilon$ are the masses of P_0 and P_1, and $\mathbf{r}_1 := (\cos(n_1 t), \sin(n_1 t))$. In the document we also assume that the motion of the asteroid P is located in the orbital plane of P_0 and P_1, which means we are studying the planar circular restricted three-body problem (PCR3BP). In the heliocentric reference frame, the Hamiltonian of the planar version of the problem is represented by

$$\begin{aligned}\mathcal{H}(\mathbf{r}, \mathbf{p}, t) &:= H_0(\mathbf{r}, \mathbf{p}) + \varepsilon H_1(\mathbf{r}, \mathbf{r}_1) \\ &= \frac{\|\mathbf{p}\|^2}{2} - \frac{1}{\|\mathbf{r}\|} + \varepsilon \left(-\frac{1}{\|\mathbf{r} - \mathbf{r}_1\|} + \frac{\mathbf{r} \cdot \mathbf{r}_1}{\|\mathbf{r}_1\|^3} \right)\end{aligned}$$

where $\mathbf{r}_1$ is dependent of the time t, therefore the Hamiltonian is not autonomous. The Hamilton's equations are described by the following system of $2n$ first order differential equations

$$\dot{\mathbf{r}} = \frac{\partial \mathcal{H}}{\partial \mathbf{p}} = \mathbf{p}, \tag{1}$$

$$\dot{\mathbf{p}} = -\frac{\partial \mathcal{H}}{\partial \mathbf{r}}. \tag{2}$$

Here the frequency of the circular orbit n_1 is called the mean motion of P_1

$$n_1 = \sqrt{\frac{G(m_0 + m_1)}{r_1^3}} = \frac{2\pi}{T_1}, \tag{3}$$

where T_1 is the period of the circular orbit. By denoting $\lambda_1 := n_1 t$, we can also introduce its conjugate momentum Λ_1, which leads to the autonomous Hamiltonian representation of the problem:

$$\mathcal{H}(\mathbf{r}, \mathbf{p}) = \frac{\|\mathbf{p}\|^2}{2} - \frac{1}{\|\mathbf{r}\|} + n_1 \Lambda_1 + \varepsilon \left(-\frac{1}{\|\mathbf{r} - \mathbf{r}_1(\lambda_1)\|} + \frac{\mathbf{r} \cdot \mathbf{r}_1(\lambda_1)}{\|\mathbf{r}_1(\lambda_1)\|^3} \right)$$

where the perturbation part H_1 is dependent of the new angle variable λ_1. The mass ratio ε characterizes the strength of the perturbation, for instance, $\varepsilon \simeq 0.001$ for Jupiter problem and $\varepsilon \simeq 0.000051$ for Neptune problem. According to the value of ε and the distance between P and P_1 we have the following situations:

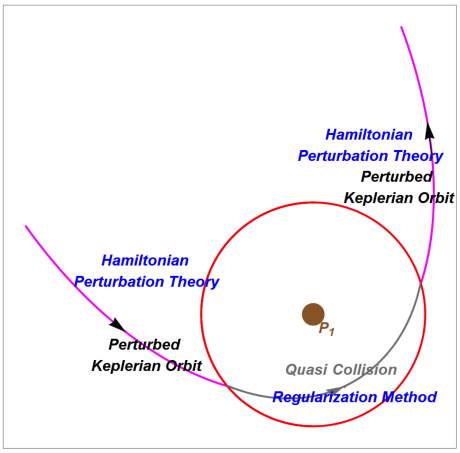


Figure 2: Illustration of application of Hamiltonian perturbation theory (see [5] for some analytical description). Note the red circle around P_1 denotes close encounter region.

- If $\varepsilon = 0$ we have $\mathcal{H} = \frac{\|\mathbf{p}\|^2}{2} - \frac{1}{\|\mathbf{r}\|}$, which represents that P is attracted only by the Sun. In this case we have the Kepler problem, which is integrable. The motion of P can be computed by its initial condition.
- If $0 < \varepsilon << 1$ and the asteroid P is far from the planet P_1, we have that the perturbation H_1 is small. In this situation Hamiltonian perturbation theory can be employed.
- If P is close to P_1, H_1 is not a perturbation: during the quasi-collision strong modifications to the Keplerian motion can occur in short time, which could lead to a very large value of the perturbation. In such case the dynamics is difficult to predicted, classical perturbation theory fails, and numerical integration methods would be unstable.

1.2 Mean motion resonances and quasi collisions

As we defined before, $n_1 = \frac{2\pi}{T_1}$ is the mean motion of the planet P_1, which is the orbital frequency of its motion. Similarly, we could introduce also the orbital frequency T and the mean motion $n = \frac{2\pi}{T}$ of the asteroid P. When the asteroid is in $p : q$ resonance with the planet if

$$\frac{T_1}{T} = \frac{p}{q}, \qquad p, q \in \mathbb{N},$$

where p and q are coprime positive integers. Equivalently, the resonance condition can be expressed with the mean motions

$$\frac{n}{n_1} = \frac{p}{q},$$

which is the so called $p : q$ mean motion resonance.

For the periodic solutions of the problem, we have the following equations

$$\|\mathbf{r}_1(t+T_1)\| = \|\mathbf{r}_1(t)\|, \quad \forall t$$
$$\|\mathbf{r}(t+T)\| = \|\mathbf{r}(t)\|, \quad \forall t.$$

Therefore if for some $t \in \mathbb{R}$, $\|\mathbf{r}(t)\| \simeq \|\mathbf{r}_1(t)\|$ we have a quasi-collision, and this quasi-collision repeats after

$$\Delta t = qT_1 = pT. \tag{4}$$

In the Solar system we have orbits satisfying such resonance conditions. For instance, the Asteroid 2024 YR24, which is in $1:4$ resonance with the Earth, was deemed to have chance of Earth impact in the near future.

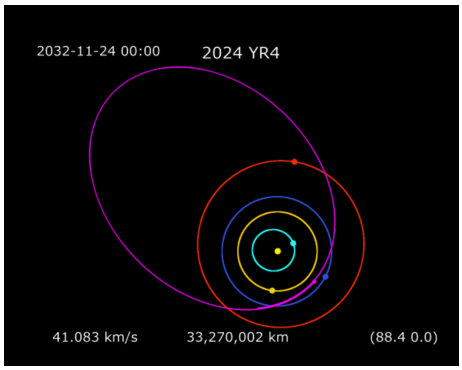


Figure 3: Predicted motion of the Asteroid 2024 YR24. The yellow dot in the middle is the Sun, the blue circle is the trajectory of the Earth, and the pink elliptical trajectory is the predicted orbit of the asteroid. The asteroid and the Earth are predicted to have an close encounter on November 24th, 2032 (see the pink dot and the blue dot).

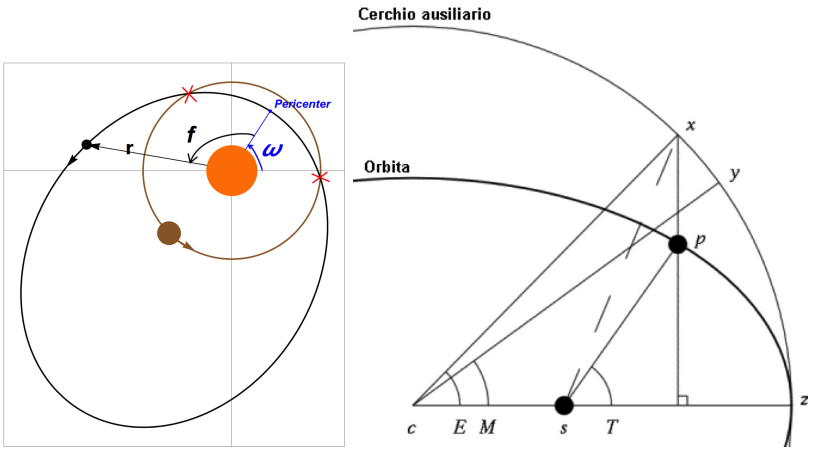


Figure 4: Illustration of the orbital elements.

2 Variables for representation of the Hamiltonian

To suitably represent the Hamiltonian we need to utilize suitable variables (for instance, action-angle variables). Before introducing them, we first introducing some orbital variables, which are commonly used by mathematicians and astronomists (see [4]), to describe the shape of the orbit and the position of the asteroid.

2.1 Orbital elements

In the left picture of Fig. 4 we show the basic elements to describe the shape of the orbit and the position of the asteroid, which are called orbital elements:

- a: semi-major axis of the elliptic orbit of the asteroid;
- e: eccentricity of the elliptic orbit of the asteroid;
- f: true anomaly, the angle between the current position of asteroid and the pericenter (or perihelion when the central body is the Sun);
- ω: argument of pericenter (or perihelion), the angle between the pericenter and the reference x-axis, which indicates how the elliptic orbit of the asteroid is rotated in the reference frame.

With these variables, according to the Kepler's first law we have the following

$$\|\mathbf{r}\| = \frac{a(1-e^2)}{1+e\cos(f)},$$

which can be regarded as polar expression of the length of the radius vector. When $f = 0$, we have $\|r\| = a(1-e)$ which means the asteroid is located at the pericenter. When $f = \pi$, $\|r\| = a(1+e)$ which means the asteroid is located at the apocenter. Besides the true anomaly f, two additional angles are frequently used to describe the motion, of which one is called the eccentric anomaly E, corresponding to the position of the asteroid in the auxiliary circle with radius equal to a, see the left picture of Fig. 4. There are several equations describing their relation

$$\sin f = \frac{\cos E - e}{1-e\cos E}, \quad \cos f = \frac{\sqrt{1-e^2}\sin E}{1-e\cos E},$$
$$\tan\frac{f}{2} = \sqrt{\frac{1+e}{1-e}}\tan\frac{E}{2}.$$

Note they are one-to-one correspondent.

To define the other angle we first recall the Kepler's third law:

$$\frac{2\pi}{T} = \frac{1}{a^{3/2}},$$

where T is period of the elliptic orbit of the asteroid. The left-hand side is the frequency of the orbit, also called mean motion of the asteroid, denoted by n. Then we could introduce the other angle, called the mean anomaly M

$$M = n(t - t_0),$$

where t is the time and t_0 is the time of passage at pericenter. The relationship between the eccentric anomaly and the mean anomaly is the Kepler equation:

$$M = E - e \sin E.$$

2.2 Delaunay variables

The aforedefined orbital elements are not action-angle variables we need. However, we could introduce the well-known Delaunay variables based on the orbital elements:

$$\begin{aligned} L &= \sqrt{a}, & l &= M, \\ G &= \sqrt{a}\sqrt{1-e^2}, & g &= \omega, \end{aligned}$$

or, equivalently the modified Delaunay variables

$$\begin{aligned} \Lambda &= \sqrt{a}, & \lambda &= l + g = M + \omega \\ \Phi &= L - G = \sqrt{a}(1 - \sqrt{1-e^2}), & \varphi &= -g = -\omega. \end{aligned}$$

In our case we use the modified Delaunay variables, and the integrable part of the Hamiltonian will be conjugated to

$$H_0 = -\frac{1}{2\Lambda^2} + n_1 \Lambda_1.$$

To calculate the perturbation part H_1 with Delaunay variables one has to express orbital elements with Delaunay variables and substitute them into the following expression of H_1

$$H_1 = -\left(\frac{1}{\sqrt{r^2 + 1 - 2r\cos(\psi)}} - r\cos(\psi) \right),$$

where $\psi := (\omega + f) - \lambda_1$ is the difference between the longitudes of P and P_1.

Finally the original Hamiltonian is canonically conjugate to

$$\mathcal{H}(\Lambda, \Phi, \Lambda_1, \lambda, \varphi, \lambda_1) = -\frac{1}{2\Lambda^2} + n_1 \Lambda_1 + \varepsilon H_1(\Lambda, \Phi, \lambda, \varphi, \lambda_1).$$

Note the unperturbed problem has 2 frequencies (or, 2 mean motions)

$$\begin{aligned} n_1 &= \frac{\partial H_0}{\partial \Lambda_1}, \\ n &= \frac{\partial H_0}{\partial \Lambda} = \frac{1}{\Lambda^3} = \frac{1}{a^{3/2}}. \end{aligned}$$

We recall that $p : q$ resonance is equivalent to

$$\frac{n}{n_1} = \frac{p}{q}.$$

3 Hamiltonian Perturbation Theory

In the section we briefly introduce the Hamiltonian perturbation theory, since our purpose is know the limit of the application of the Hamiltonian perturbation theory (see [4]) for quasi-collision problems. A quasi-integrable Hamiltonian system is defined by the following Hamiltonian function

$$\mathcal{H}(I,\phi) = H_0(I) + \varepsilon H_1(I,\phi)$$

where $(I,\phi) \in \Omega \times \mathbb{T}^n \subset \mathbb{R}^n \times \mathbb{T}^n$ are action angle variables (in previous section we introduced the action-angle variables for the circular restricted three-body problem) and $\varepsilon \in \mathbb{R}$ is a small parameter. The Hamilton's equations are

$$\begin{aligned} \dot{\phi} &= \frac{\partial \mathcal{H}}{\partial I} &&= \frac{\partial H_0}{\partial I} + \varepsilon \frac{\partial H_1}{\partial I} \\ \dot{I} &= -\frac{\partial \mathcal{H}}{\partial \phi} &&= -\varepsilon \frac{\partial H_1}{\partial \phi}. \end{aligned}$$

The Hamiltonian $\mathcal{H}$ is obtained by slightly perturbing the integrable Hamiltonian H_0, therefore the flow is correspondingly perturbed. The effect of perturbation on the flow is studied by Hamiltonian perturbation theory (for instance, KAM theory and Nekhoroshev theory). The purpose of Hamiltonian perturbation theory is to find a canonical transformation $\mathcal{C} : (J,\psi) \to (I,\phi)$ to conjugate the Hamiltonian to a more integrable one (this process is also called averaging)

$$\tilde{\mathcal{H}}(J,\psi) := \mathcal{H}(\mathcal{C}(J,\psi)).$$

Based on the problem we try to consider, there are two cases: non-resonant and resonant cases. For non-resonant case, the canonical transformation could give us

$$\tilde{\mathcal{H}}(J,\psi) := \tilde{H}_0(J) + \varepsilon \tilde{H}_1(J) + \varepsilon^2 \tilde{H}_2(J,\psi),$$

where $\tilde{H}_2$, the new perturbation part, is of order $\mathcal{O}(\varepsilon^2)$. If there exists $k_* \in \mathbb{Z}^n$ such that

$$k_* \cdot \frac{\partial \tilde{H}_0}{\partial J} = 0,$$

we say that the frequencies are resonant. For this case, the canonical transformation will give us a different averaged Hamiltonian

$$\tilde{\mathcal{H}}(J,\psi) := \tilde{H}_0(J) + \varepsilon \tilde{H}_1(J, k_* \cdot \psi) + \varepsilon^2 \tilde{H}_2(J,\psi),$$

from which one realize that some linear combination of angle variables cannot be eliminated.

To construct the canonical transformation one need to find a suitable Hamiltonian (called generating Hamiltonian or function), whose flow is and will be used as the canonical transformation we are looking for. To see that we first define the Lie Series.

3.1 Lie Series

The averaging process can be realized by Lie series operator associated to a generating Hamiltonian χ. Be defining the Lie derivative $\mathcal{L}_\chi F := \{F, \chi\}$ and the exponential Lie operator is defined as

$$\exp(\varepsilon\mathcal{L}_\chi) = \sum_{s\geq 0} \frac{\varepsilon^s}{s!}\mathcal{L}_\chi^s = id + \varepsilon\{\cdot, \chi\} + \frac{\varepsilon^2}{2}\{\{\cdot, \chi\}, \chi\} + ...$$

Then the Hamiltonian flow of $\chi(J, \psi)$ is expressed by

$$\begin{aligned}\exp(\varepsilon\mathcal{L}_\chi)\, J(t) &= J(t) + \varepsilon\{J(t), \chi\} + \frac{\varepsilon^2}{2}\{\{J(t), \chi\}, \chi\} + ... = J(t+\varepsilon),\\ \exp(\varepsilon\mathcal{L}_\chi)\, \psi(t) &= \psi(t) + \varepsilon\{\psi(t), \chi\} + \frac{\varepsilon^2}{2}\{\{\psi(t), \chi\}, \chi\} + ... = \psi(t+\varepsilon),\end{aligned}$$

from which one could realized that the exponential Lie operator applied to the action-angle variables (J, ψ) are just the expansion of the flow $(J(t), \psi(t))$ of the generating Hamiltonian χ at time $t = \varepsilon$ with (J, ψ) as the initial condition.

Instead of applying the canonical transformation directly to the variables, one could just apply the Lie operator to the original Hamiltonian and replace the old variables with the new variables.

Theorem 1 (Exchange theorem)

$$\mathcal{H}(I, \phi)\Big|_{I=\exp(\varepsilon\mathcal{L}_\chi)J, \phi=\exp(\varepsilon\mathcal{L}_\chi)\psi} = \exp(\varepsilon\mathcal{L}_\chi)\,\mathcal{H}\Big|_{I=J,\phi=\psi}.$$

Therefore the averaging process is simplified, and the averaged Hamiltonian has the form

$$\begin{aligned}\tilde{\mathcal{H}} &= \exp(\varepsilon\mathcal{L}_\chi)\,\mathcal{H}\\ &= H_0 + \varepsilon H_1 + \varepsilon\{H_0, \chi\} + \mathcal{O}(\varepsilon^2).\end{aligned}$$

If we consider the *non-resonant case*, we would finally obtain the following expression

$$\tilde{\mathcal{H}}(I, \phi) = H_0(I) + \varepsilon c_0(I) + \mathcal{O}(\varepsilon^2).$$

Define the frequency map $\omega_0(I) := \frac{\partial H_0}{\partial I}$ we have the generating function χ defined by

$$\chi(I, \phi) = \sum_{k\in\mathbb{Z}^n} d_k(I)\exp(ik\cdot\phi), \text{with} \begin{cases} d_0 = 0,\\ d_k(I) = -i\frac{c_k(I)}{k\cdot\omega_0(I)}, \end{cases}$$

where c_k are Fourier coefficients of H_1

$$H_1(I, \phi) = \sum_{k\in\mathbb{Z}^n} c_k(I)\exp(ik\cdot\phi).$$

As we mentioned before, if $\exists k_* \in \mathbb{N}^n$ such that $k_* \cdot \omega_0 \simeq 0$, we have resonant frequencies. In PCR3BP, frequencies are mean motions, correspondingly we have mean motion resonance. The mean motion resonance would lead to small or zero divisors problem since the generating Hamiltonian is defined by using $k_* \cdot \frac{\partial H_0}{\partial I}$ as denominators. Therefore to find the suitable generating Hamiltonian, we first define a resonant set

$$\mathcal{K} = \{k \in \mathbb{Z}^n : k \cdot \omega_0 = 0\}.$$

Correspondingly we have the generating Hamiltonian as follows

$$\chi(I, \phi) = \sum_{k \in \mathbb{Z}^n} d_k(I) \exp(ik \cdot \phi), \text{with} \begin{cases} d_k = 0, \text{if } k \in \mathcal{K} \\ d_k(I) = -i \frac{c_k(I)}{k \cdot \omega_0(I)}, \end{cases}$$

and the averaged Hamiltonian is, for $k_* \in \mathcal{K}$,

$$\tilde{\mathcal{H}}(I, \phi) = H_0(I) + \varepsilon \sum_{n \in \mathbb{Z}, k_* \in \mathcal{K}} c_{nk_*}(I) \exp(i\ n(k_* \cdot \phi)) + \mathcal{O}(\varepsilon^2).$$

4 Mean motion resonance dynamics in PCR3BP

Classical Fourier expansion of the perturbing function H_1 in (modified) Delaunay variables is

$$H_1 = \sum_{(k,m) \in \mathbb{Z}^2} c_{k,m}(\Lambda, \Phi) \exp[i(k\lambda + (k+m)\varphi + m\lambda_1)].$$

For $p : q$ resonance, we introduce the following action angle variables

$$\begin{pmatrix} \sigma \\ \nu \\ \theta \end{pmatrix} = \begin{pmatrix} q & q-p & -p \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \lambda \\ \varphi \\ \lambda_1 \end{pmatrix}, \begin{pmatrix} S \\ N \\ \Theta \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ p-q & q & 0 \\ p-q+1 & q-1 & 1 \end{pmatrix} \begin{pmatrix} \Lambda \\ \Phi \\ \Lambda_1 \end{pmatrix},$$

where $\sigma = q\lambda + (q-p)\varphi - p\lambda_1$ is called the resonant angle (which indeed is the linear combination of angles variables in previous section, i.e. $k_* \cdot \phi$). In these variables, the integrable part is

$$H_0 = -\frac{1}{2(qS - N)^2} + n_1(-pS + \Theta),$$

and the perturbation part would depend on the summation $\nu + \theta$

$$H_1 = H_1(S, N, \sigma, \nu + \theta).$$

Classical expansion fails if we have collisions. To represent the collision property we would introduce the "singular set". First we consider the system of equations:

$$\begin{aligned} \|\mathbf{r}\| &= \|\mathbf{r}_1\| + d, \\ \omega + f &= \lambda_1 + \alpha, \end{aligned}$$

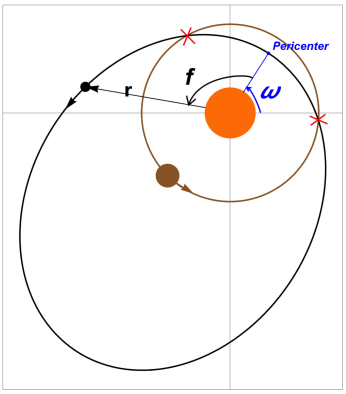


Figure 5: Illustration of the crossing orbits.

when $d = 0$ and $\alpha = 0$ a collision happens, and when the values of $|d|$ and $|\alpha|$ are small we observe close encounters. The solution expressed in resonant variables is

$$\begin{pmatrix} \sigma \\ \nu + \theta \end{pmatrix} = \begin{pmatrix} q & -p \\ -1 & 1 \end{pmatrix} \begin{pmatrix} M_*^i(S, N, d) \\ f_*^i(S, N, d) - \alpha \end{pmatrix},$$

where $i = 1$ means that tangent orbits occur; $i = 2$ we have crossing orbits. Since we are going to study the resonant dynamics in $\sigma - S$ plane, we define a set parameterized by d and α:

$$\mathcal{S}_{d,\alpha} = \left\{ (S, N, \sigma) : \sigma = qM_*^i(S, N, d) - p(f_*^i(S, N, d) - \alpha), i = 1, 2 \right\}.$$

The set $\mathcal{S}_{0,0}$ is called "singular set". If $(S, N, \sigma) \in \mathcal{S}_{0,0}$ then there exists one value of $\nu + \theta$ corresponding to a collision of P with P_1.

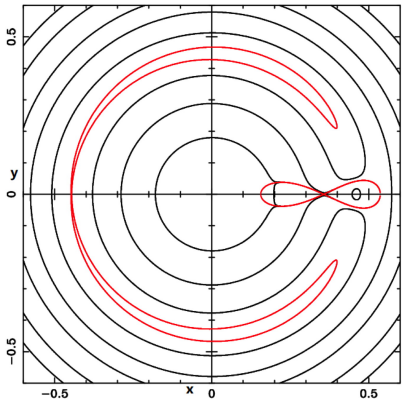


Figure 6: Illustration of the collision set for $N = -1.06$. The axes are $x = \sqrt{2S}\cos\sigma, y = \sqrt{2S}\sin\sigma$. The red curve near the right equilibrium is the collision curve. The other red curve is the separatrix of the equilibrium at $\sigma = \pi$.

Fig. 6 illustrates the singular set projected into $x-y$ plane for $N=-1.06>N_c$. One could clearly notice that the existence of the singular indeed generates a new stable region around the equilibrium at $\sigma=0$.

4.1 Resonant Norma Form Hamiltonian and its dynamics

Partially expanding H_1 with respect to $\nu+\theta$, which is well defined for crossing orbits

$$H_1=c_0(S,N,\sigma)+\sum_{k\geq 1}c_k(S,N,\sigma)\cos(k(\nu+\theta))+\sum_{k\geq 1}s_k(S,N,\sigma)\sin(k(\nu+\theta))$$

with the generating function defined by

$$\chi=\sum_{k\geq 1}\left(b_k\sin(k(\nu+\theta))+d_k\cos(k(\nu+\theta))\right)$$

where

$$b_k=\frac{c_k}{k\left(\frac{\partial H_0}{\partial N}+\frac{\partial H_0}{\partial \Theta}\right)},\ d_k=-\frac{s_k}{k\left(\frac{\partial H_0}{\partial N}+\frac{\partial H_0}{\partial \Theta}\right)}$$

The resonant normal form Hamiltonian is obtained by ignoring higher order terms

$$\mathcal{H}^{Res}(N,S,\sigma)=H_0(S,N)+\varepsilon H_1^{Res}(S,N,\sigma)=H_0(S,N)+\varepsilon c_0(S,N,\sigma)$$

Clear for the averaged Hamiltonian $\mathcal{H}^{Res}$, N is a first integral

$$N=\sqrt{a}\left[p-q\sqrt{1-e^2}\right],$$

whose contour curves could be represented in $a-e$ plane.

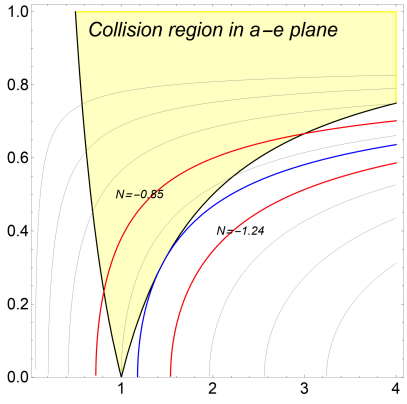


Figure 7: Contour curves of the first integral N.

For values of a and e satisfying

$$a(1-e)\leq 1\leq a(1+e),$$

there is a collision region. Correspondingly there exists a critical value N_c, the blue curve in the contour curves, such if $N < N_c$ no collision shows (but close encounters are possible). For $1:2$ resonance, the critical value is $N_c \simeq -1.08524....$

From now on we would choose the following coordinates to represent the resonant dynamics (see [6] for more examples and details)

$$x = \sqrt{2S}\cos\sigma, \quad y = \sqrt{2S}\sin\sigma$$

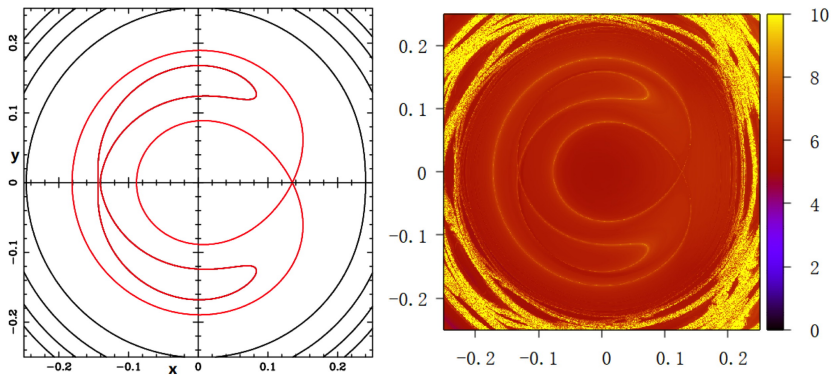


Figure 8: The phase portrait and Fast Lyapunov Indicators (FLI, see [3]) for $N = -1.24$ for Jupiter problem.

In Fig. 8 one can clearly see the resonant dynamics for $N = -1.24$. To compute the FLI we first fix the value of N and $\nu + \theta = \pi$ and select one point (σ, S) then we could integrate the regularized equations of motion and obtain the value of FLI. We could also notice that the FLI representation of separatrices (yewllow curve inside the resonant region) is in accordance with resonant phase portraits.

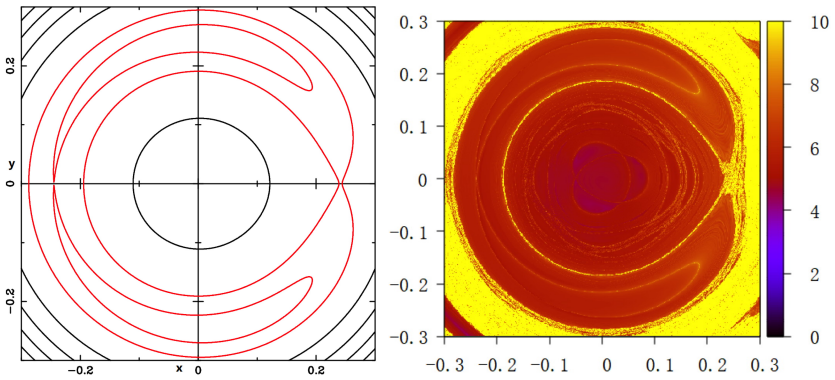


Figure 9: The phase portrait and Fast Lyapunov Indicators (FLI) for $N = -1.20$ for Jupiter problem.

As $N = -1.20$ increase a little bit, we notice similar resonant dynamics in the resonant phase portraits but the outer separatrix becomes chaotic.

As N increases to a value larger than N_c we have singular set in the phase portrait.

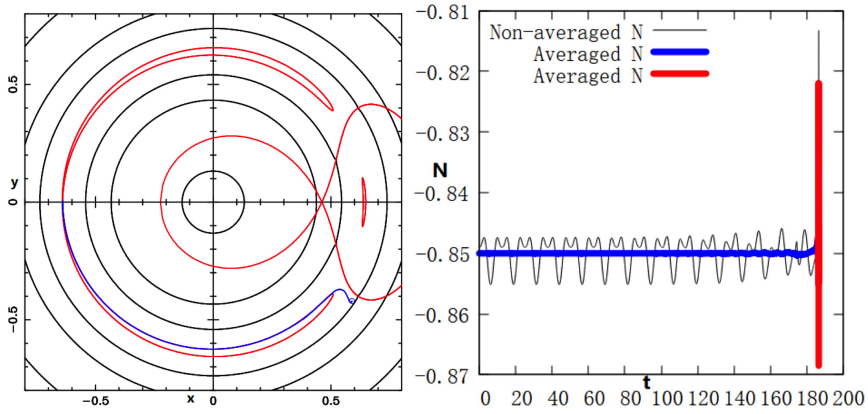


Figure 10: The phase portrait and an orbit for $N = -0.85$ for Jupiter problem.

As $N = -0.85$ we have resonant dynamics with singular set, which seems generating a new stable region around $\sigma = 0$. Besides, an orbit is also projected into (σ, S) plane, whose averaged and non-averaged values of N are also depicted in the right figure, from which one could clearly see that averaging works when the orbit is far away from the singular set and the dynamics is difficult to predict when the orbit is close to the singular set.

In the following we also show some results for Neptune mass ratio $\varepsilon = 0.000051....$

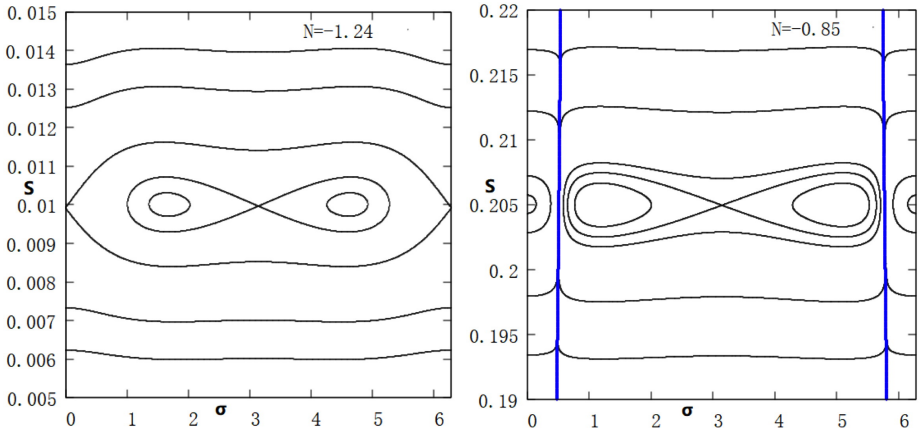


Figure 11: Resonant phase portrait in (σ, S) plane for $N = -1.24$ (left) and $N = -0.85$ (right) for Neptune problem.

From Fig. 11, one can see that the dynamics of resonant normal form Hamiltonian are similar with Jupiter cases: for $N = -1.24 < N_c$ we have very regular dynamics and if

$N = -0.85 > N_c$ there are singular set.

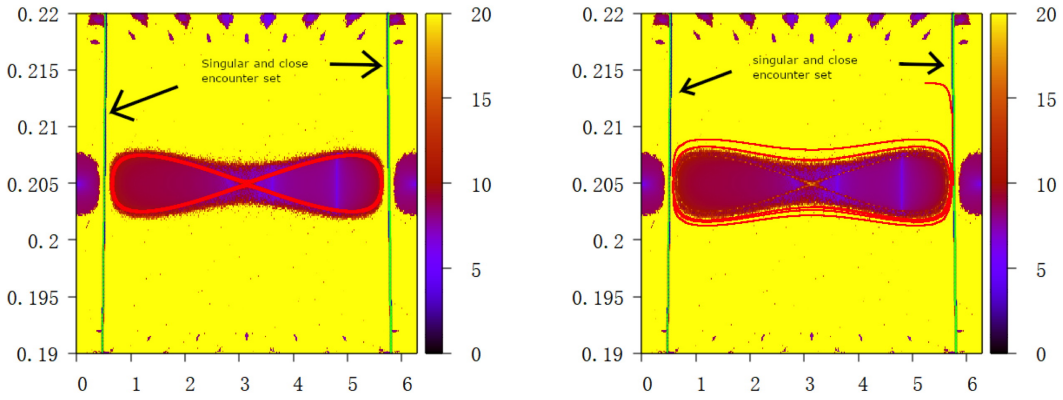


Figure 12: Representation of FLI values in (σ, S) plane and projection of regular and chaotic orbits for $N = -0.85$.

In Fig. 12, with FLI we could identify the regular and chaotic regions. Besides, we could clearly notice that when the initial condition of the orbit is close to the boundary of the regular region, the close encounters effect would accumulate and the orbit would gradually become chaotic.

References

[1] Beaugé, C., *Asymmetric Librations in Exterior Resonances.* Celestial Mechanics and Dynamical Astronomy 60 (1994), 225–248.

[2] Ferraz-Mello, S. and Sato, M., *The very-high-eccentricity asymmetric expansion of the disturbing function near resonances of any order.* Astronomy and Astrophysics, 225 (1989), 541–547.

[3] Guzzo, M., Lega, E., *Theory and applications of fast Lyapunov indicators to model problems of celestial mechanics.* Celestial Mechanics and Dynamical Astronomy 135 (2023), article 37.

[4] Morbidelli, A., "Modern celestial mechanics: aspects of solar system dynamics". Advances in Astronomy and Astrophysics, 2002.

[5] Mastroianni, R. and Efthymiopoulos, C., *The phase-space architecture in extrasolar systems with two planets in orbits of high mutual inclination.* Celestial Mechanics and Dynamical Astronomy 135 (2023), article 22.

[6] Liu, X., Guzzo, M., *On the limits of application of mean motion resonant normal forms of the three-body problem for crossing orbits and close encounters.* Celestial Mechanics and Dynamical Astronomy 137/1 (2025). `https://doi.org/10.1007/s10569-024-10232-0`.

Lavrentiev Phenomenon and semicontinuous envelope for integral functionals

Tommaso Bertin (*)

In these notes we present recent results concerning the non-occurrence of the Lavrentiev Phenomenon for integral functionals with Lagrangians that are non-convex and non-continuous with respect to the last variable. This problem has been studied for decades, with significant progress made in recent years.

To avoid the Lavrentiev Phenomenon, we examine the integral representation of the lower semicontinuous envelope, allowing us to apply results from the literature regarding convex Lagrangians.

Before defining the Lavrentiev Phenomenon, we first introduce some basic concepts from the Calculus of Variations. The initial definitions concern special continuity and compactness notions.

Definition 1 A function $F : X \to \mathbb{R}$ is sequentially continuous if

$$\lim_n x_n = x \Rightarrow \lim_n F(x_n) = F(x)\,.$$

Definition 2 A topological space X is sequentially compact if every sequence x_n has a convergent subsequence.

The classical notions of continuity and compactness imply the sequential ones. The reverse is true if X has a countable basis of open sets at every point.

We recall a sequential version of a classical result: the Weierstrass Theorem.

Theorem 3 *If X is sequentially compact and $F : X \to \mathbb{R}$ is sequentially continuous then F attains maximum and minimum on X.*

This fundamental theorem has profound implications in the Calculus of Variations. In fact, it involves two concepts, continuity and compactness, that, in a certain sense, are in competition with each other. If the topology of X is rich (meaning there are many open sets), there are more continuous functions but fewer compact sets. On the other hand, if

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: bertin@math.unipd.it . Seminar held on 2 April 2025.

the topology is poor, there are fewer continuous functions but more compact sets. Our goal is now to find a suitable topology that preserves a sufficiently large family of continuous functions while still ensuring an adequate family of compact sets.

We observe that if F is continuous then for every $Y \subset X$ such that $\overline{Y} = X$

$$\inf_Y F = \inf_X F, \qquad \sup_Y F = \sup_X F\,.$$

In the Calculus of Variations, the focus is on the study of minima. Therefore, we do not require the continuity of F, but rather only its lower semicontinuity.

Definition 4 A function $F : X \to \mathbb{R}$ is lower sequentially semicontinuous if

$$\lim_n x_n = x \Rightarrow F(x) \leq \liminf_n F(x_n)$$

where

$$\liminf_n a_n = \sup_n \inf_{m \geq n} a_m\,.$$

Theorem 5 *If X is sequentially compact and $F : X \to \mathbb{R}$ is lower sequentially semicontinuous then F attains minimum on X.*

We note that in this case we can only say that if $Y \subset X$ then

$$\inf_Y F \geq \inf_X F\,.$$

So far, we have considered a generic topological space X. Now, given $\Omega \subset \mathbb{R}^N$, let us introduce some classical spaces commonly used in the Calculus of Variations.

Definition 6 A function $u \in L^P(\Omega)$, with $p \in [1, +\infty[$, if

$$\int_\Omega |u(x)|^p dx < +\infty\,.$$

A function $u \in L^\infty(\Omega)$ if

$$\text{esssup}_\Omega |u| < +\infty\,.$$

These spaces, equipped with the standard topology induced by the norm, have a limitation: they contain few compact sets. For example, the unit ball is not compact. To obtain more compact sets while preserving a large family of continuous functions, we introduce the concept of the weak topology.

Definition 7 A sequence $u_n \rightharpoonup u$ in $L^p(\Omega)$ if for every $v \in L^{p'}(\Omega)$ with $p' = \frac{p}{p-1}$

$$\int_\Omega u_n(x)v(x)dx \to \int_\Omega u(x)v(x)dx\,.$$

A sequence $u_n \rightharpoonup^* u$ in $L^\infty(\Omega)$ if for every $v \in L^1(\Omega)$

$$\int_\Omega u_n(x)v(x)dx \to \int_\Omega u(x)v(x)dx\,.$$

We also need a notion of derivative. In particular, we require the validity of the integration by parts formula. To address this, we introduce the definition of Sobolev spaces.

Definition 8 A function $u \in W^{1,p}(\Omega)$ if $u \in L^p(\Omega)$ and for every $1 \leq \alpha \leq n$ there exists $\partial^\alpha u \in L^p(\Omega)$ such that for every $\varphi \in C_C^\infty(\Omega)$

$$\int_\Omega \partial^\alpha u(x)\varphi(x)dx = -\int_\Omega u(x)\partial^\alpha\varphi(x)dx\,.$$

For these spaces, we can also introduce the concept of weak convergence.

Definition 9 A sequence $u_n \rightharpoonup u$ in $W^{1,p}(\Omega)$ if $u_n \rightharpoonup u$ in $L^p(\Omega)$ and $\partial^\alpha u_n \rightharpoonup \partial^\alpha u$ in $L^p(\Omega)$ for every $1 \leq \alpha \leq n$.
A sequence $u_n \rightharpoonup^* u$ in $W^{1,\infty}(\Omega)$ if $u_n \rightharpoonup^* u$ in $L^\infty(\Omega)$ and $\partial^\alpha u_n \rightharpoonup^* \partial^\alpha u$ in $L^\infty(\Omega)$ for every $1 \leq \alpha \leq n$.

Now we are ready to state one of the main theorems in the Calculus of Variations.

Theorem 10 *Let $\Omega \subset \mathbb{R}^N$ be bounded Lip open, $f : \Omega \times \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^+$, C^2, convex with respect to the last variable and with $a \in L^1(\Omega)$ and $b > 0$ such that*

$$f(x, u, \xi) \geq b\|\xi\| + a(x)\,.$$

Then the functional

$$F(u) := \int_\Omega f(x, u(x), \nabla u(x))dx$$

is weakly lower semicontinuous in $W^{1,1}(\Omega)$.

The idea of a connection between the weak lower semicontinuity of F and the convexity of the Lagrangian with respect to its last variable originates from Tonelli.

This theorem is particularly useful for studying the existence of minimizers. Indeed, if the functional is also coercive (meaning if its sublevel sets are bounded) then any minimizing sequence, possibly after extracting a subsequence, converges to a limit, which is a minimizer. However, starting from a regular sequence, we obtain no information about the regularity of the minimizer. Moreover, it may happen that the value of the minimizer is strictly less than the $\liminf$ of the functional values along the minimizing sequence. This situation is known as the Lavrentiev Phenomenon.

Definition 11 Let be $Y \subset X$, $\overline{Y} = X$ and $F : X \to \mathbb{R}$ we say the Lavrentiev Phenomenon occurs if

$$\inf_Y F > \inf_X F\,.$$

In our case $X = W^{1,1}(\Omega)$ and $Y = W^{1,\infty}(\Omega)$.

This phenomenon can occur even with regular Lagrangians; for instance, Manià showed in [12] that

$$\min_{\mathrm{id}+W_0^{1,1}([0,1])} \int_0^1 (x - u^3(x))^2 |u'(x)|^6 \, dx \tag{0.2}$$
$$< \inf_{\mathrm{id}+W_0^{1,\infty}([0,1])} \int_0^1 (x - u^3(x))^2 |u'(x)|^6 \, dx \, .$$

In [15], Zhikov presented another example involving a Lagrangian that depends only on the gradient and the spatial variable. Considering

$$f(x,\xi) = |\xi|^p + a(x)|\xi|^q \quad a(x) := \begin{cases} \frac{x_1 x_2}{\sqrt{x_1^2 + x_2^2}} & x_1 x_2 > 0 \\ 0 & x_1 x_2 \leq 0 \end{cases}$$

where $\Omega = B_1(0) \subset \mathbb{R}^2$, $1 \leq p < 2 < 3 < q$; there is a boundary condition φ such that

$$\inf_{W^{1,p}_\varphi(\Omega)} \int_\Omega f(x, \nabla u(x)) dx < \inf_{W^{1,q}_\varphi(\Omega)} \int_\Omega f(x, \nabla u(x)) dx \, .$$

Up to this point, we have focused on the minimizer. In [8], Buttazzo and Mizel proposed interpreting the Lavrentiev Phenomenon as a problem involving relaxed functionals, introducing the concept of the Lavrentiev gap.

Definition 12 If F is sequential lower semicontinuous on X we define

$$F_Y(u) := \begin{cases} F(u) \text{ in } Y \\ +\infty \text{ in } X \setminus Y \end{cases}$$

and

$$scF_Y := \sup\{G \quad \text{l.s.c. on } X | G \leq F \quad \text{on} \quad Y\} \, .$$

We define the Lavrentiev gap as

$$L(u) = scF_Y(u) - F(u) \, .$$

In our case

$$F(u) := \int_\Omega f(x, u(x), \nabla u(x)) dx$$

$$scF_Y(u) = \inf \left\{ \liminf \int_\Omega f(x, u_n(x), \nabla u_n(x) dx \Big| u_n \rightharpoonup u \quad \text{in} \quad W^{1,1}(\Omega) \right\}$$

with $u \in W^{1,1}(\Omega)$ and $u_n \in W^{1,\infty}(\Omega)$.
Clearly, if $L(u) = 0$ for every $u \in W^{1,1}(\Omega)$, then the Lavrentiev Phenomenon does not occur.

We now present some results in which $L(u) \equiv 0$, and consequently, the Lavrentiev Phenomenon does not occur.

The first theorem, due to Alberti and Serra Cassano ([1]), concerns the one-dimensional autonomous case. Note that the approximating sequence may not preserve the boundary data.

Theorem 13 *Assume for every $r > 0$ there exists $c > 0$ such that $f : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R} \cup \{+\infty\}$ is bounded on $B_r \times B_c$.*
Then, $\forall p \in [1, \infty[$ and $\forall u \in W^{1,p}(I, \mathbb{R}^M)$ $\exists u_n \in W^{1,\infty}(I, \mathbb{R}^M)$ such that

$$\|u_n - u\|_{W^{1,p}} \to 0$$

$$\int_\Omega f(u_n(x), Du_n(x))dx \to \int_\Omega f(u(x), Du(x))dx\,.$$

This result was generalized by Mariconda in [14] to the non autonomous case with arbitrary boundary values, under the following condition:
For every K there are $k, \beta \geq 0$, $\gamma \in L^1(I)$, $\varepsilon^* > 0$ such that $\forall x \in I$

$$|f(x_2, u, \xi) - f(x_1, u, \xi)| \leq (kf(x, u, \xi) + \beta|\xi|^p + \gamma(x))|x_2 - x_1|$$

$\varepsilon^* > 0$ for $x_1, x_2 \in [x - \varepsilon^*, x + \varepsilon^*] \cap I$, $u \in B_K$, $\xi \in \mathbb{R}^n$.
In this paper the approximating sequence preserves the boundary datum.

We now turn our attention to the multidimensional scalar case and present a recent result by Bousquet, published in 2023 ([4]).

Theorem 14 *Let $\Omega \subset \mathbb{R}^N$ be bounded Lipschitz open, $f : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}$ be continuous in both variable and convex w.r.t. the last variable and $\varphi \in W^{1,\infty}(\Omega)$. Then $\forall u \in \varphi + W_0^{1,1}(\Omega)$ $\exists u_n \in \varphi + W_0^{1,\infty}(\Omega)$ such that*

$$u_n \to u \quad in \quad W^{1,1}(\Omega)$$

$$\lim_n \int_\Omega f(u_n(x), \nabla u(x))dx = \int_\Omega f(u(x), \nabla u(x))dx\,.$$

If f is not convex with respect to the last variable, we can consider the so-called bipolar f^{**}. Actually, the bipolar in the sense of the convex analysis is the greatest function convex and lower semicontinuous with respect to the last variable lower or equal than f. We can express f^{**} as

$$f^{**}(x, u, \xi) = \sup\{h(\xi)\,\text{affine}\,|h(\xi) \leq f(x, u, \xi) \quad \forall \xi \in \mathbb{R}^n\}$$

or

$$f^{**}(x,u,\xi) = \inf\Big\{ \sum_{i=1}^{n+1} \alpha_i f(x,u,\xi_i) \Big| \sum_{i=1}^{n+1} \alpha_i \xi_i = \xi \Big\}$$

where $0 \le \alpha_i \le 1$, $\sum_{i=1}^{n+1} \alpha_i = 1$.

To apply the result by Bousquet, we now study an integral representation formula for semicontinuous envelopes. Specifically, we seek conditions under which

$$\inf\Big\{ \liminf \int_\Omega f(x,u_n(x),\nabla u_n(x))dx \Big| u_n \rightharpoonup^* u \Big\} = \int_\Omega f^{**}(x,u(x),\nabla u(x))dx\,. \tag{0.3}$$

The first results were obtained by Ekeland and Temam ([10]) and Marcellini and Sbordone ([13]), assuming continuity with respect to (u,ξ).
Further developments were made by Buttazzo, Dal Maso, De Giorgi, and Leaci ([6], [7], [9], [5]) during the 1980s and 1990s. Many of these works focus on relaxing the continuity assumption with respect to u.

Our contribution is to avoid the continuity assumption with respect to ξ, under a suitable set of conditions.

Hypothesis 15 The function $f(x,u,\xi) : \Omega \times \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}$ satisfies

a) $f(x,u,\xi) : \Omega \times \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}$ is a borelian function;

b) $\forall B \subset \mathbb{R} \times \mathbb{R}^N$ bounded $\exists a \in L^1(\Omega)$ such that $|f(x,u,\xi)| \le a(x)$ $\forall (u,\xi) \in B$,

c) $\forall u \in W^{1,\infty}(\Omega)$, $\forall \tilde{B} \subset \mathbb{R}^N$ bounded and $\forall \delta > 0$ $\exists T \subset \Omega$ compact such that $|\Omega \setminus T| < \delta$ and $f(x,u(x),\xi)$ is continuous w.r.t. $x \in T$ uniformly as ξ varies in $\tilde{B}$,

d) for almost every x the function $f(x,u,\xi)$ is continuous w.r.t. u uniformly as ξ varies in bounded sets.

Assumption *b*) ensures the integrability of the Lagrangian. Assumption *c*) is a uniform version of Lusin's Theorem, which allows us to avoid requiring continuity with respect to ξ.

We are now ready to present our main result:

Theorem 16 *Let Ω be an open bounded Lipschitz subset of $\mathbb{R}^N$ and let $f : \Omega \times \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}$ satisfy Hypothesis 15. For every $\overline{u} \in W^{1,\infty}(\Omega)$ exists a sequence $u_n \in \overline{u} + W_0^{1,\infty}(\Omega)$ such that*

$$\lim_{n\to\infty} \|u_n - \overline{u}\|_\infty = 0$$

and

$$\lim_{n\to\infty} \int_\Omega f(x,u_n(x),\nabla u_n(x))dx = \int_\Omega f^{**}(x,\overline{u}(x),\nabla \overline{u}(x))dx\,.$$

Furthermore

$$sc(F_{W^{1,\infty}})(\overline{u}) = \int_\Omega f^{**}(x,\overline{u}(x),\nabla \overline{u}(x))dx\,.$$

This theorem can be used to investigate the relationship between the occurrence of the Lavrentiev Phenomenon for integral functionals with non-convex Lagrangians and its non-occurrence for the corresponding relaxed functionals.

Theorem 17 *Let $\Omega \subset \mathbb{R}^N$ be bounded Lip open, $f : \Omega \times \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}$ satisfy Hypothesis 15, $\varphi \in W^{1,\infty}(\Omega)$ and $1 \leq p < +\infty$. If*

$$\inf_{\varphi + W_0^{1,p}(\Omega)} \int_\Omega f^{**}(x, u(x), \nabla u(x))dx = \inf_{\varphi + W_0^{1,\infty}(\Omega)} \int_\Omega f^{**}(x, u(x), \nabla u(x))dx$$

then

$$\inf_{\varphi + W_0^{1,p}(\Omega)} \int_\Omega f(x, u(x), \nabla u(x))dx = \inf_{\varphi + W_0^{1,\infty}(\Omega)} \int_\Omega f(x, u(x), \nabla u(x))dx\,.$$

We can now apply this result to Bousquet's work in the autonomous case, under a simplified set of assumptions.

Hypothesis 18 The function $f : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}$ satisfies

(a) $f(u, \xi) : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}$ is borelian,

(b) $f(u, \xi)$ is continuous w.r.t. u uniformly as ξ varies on each bounded set of $\mathbb{R}^N$,

(c) $f(u, \cdot)$ is bounded on bounded sets of $\mathbb{R}^N$ for every $u \in \mathbb{R}$.

Under these assumptions, we can state the following theorem concerning the absence of Lavrentiev gap, and thus non occurrence of the Lavrentiev Phenomenon, between $W^{1,1}(\Omega)$ and $W^{1,\infty}(\Omega)$ in the autonomous multidimensional scalar case.

Theorem 19 *Let $\Omega \subset \mathbb{R}^N$ be bounded Lipschitz open, $\varphi \in W^{1,\infty}(\Omega)$, $f : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^+$ satisfy Hypothesis 18, be uniformly superlinear and f^{**} be continuous.*
Then for every $u \in \varphi + W_0^{1,1}(\Omega)$

$$sc(F_{W^{1,\infty}})(u) = \int_\Omega f^{**}(u(x), \nabla u(x))dx$$

where $sc(F_{W^{1,\infty}})$ is the lower semicontinuous envelope of

$$F_{W^{1,\infty}}(u) = \begin{cases} \int_\Omega f(u(x), \nabla u(x))dx & \text{if} \quad u \in \varphi + W_0^{1,\infty}(\Omega) \\ +\infty \quad \text{if} & u \in \varphi + W_0^{1,1}(\Omega) \setminus W_0^{1,\infty}(\Omega) \end{cases}$$

with respect to the weak topology of $W^{1,1}(\Omega)$.

References

[1] G. Alberti, F. Serra Cassano. Ser. Adv. Math. Appl. Sci. 18, 1994.

[2] J.M. Ball, V.J. Mizel. Arch. Rational Mech. Anal. 90, 1985.

[3] M. Belloni, G. Buttazzo. Kluwer Ac. Pub. Group, Dordrecht, 1995.

[4] P. Bousquet. Ann. Sc. Norm. Super. Pisa Cl. Sci. (5) 24, 2023.

[5] G. Buttazzo. Longman Scientific & Technical, 1989.

[6] G. Buttazzo, G. Dal Maso, E. De Giorgi. Atti della Accademia Nazionale dei Lincei, 1983.

[7] G. Buttazzo, A. Leaci. Journal of Functional Analysis 61, 1985.

[8] G. Buttazzo, V.J. Mizel. J. Funct. Anal. 110, 1992.

[9] G. Dal Maso, nn. Univ. Ferrara - Sez. VII - 8c. Mat. Vol. XXXIV, 1988.

[10] I. Ekeland, R. Témam, "Convex Analysis and Variational Problems". 1999 (First edition 1976).

[11] M. Lavrentiev. Ann. Matem. Pura Appl. 4, 1926.

[12] B. Manià. Boll. Un. Matem. Ital. 13, pp 147–153. 1934.

[13] P. Marcellini, C. Sbordone, "Non linear Analysis, Theory, Methods & Applications". Vol. 4, 1980.

[14] C. Mariconda. Calc. Var. Partial Differential Equations 62, no. 2, 2023. Vol. 43, No. 4, pp 1298–1312, 2005.

[15] V.V. Zhikov. Russ. J. Math. Phys. 3(2), 1995.

Let's play symplectic billiards!

Alessandra Nardi (*)

Abstract. In these notes, we will provide an overview of mathematical billiards, with a particular focus on *symplectic billiards*. The goal is to present some recent integrability and spectral rigidity results for this recent class of dynamical systems.

1 Birkhoff billiards

A mathematical billiard consists of a planar domain $\Omega \subset \mathbb{R}^2$ (billiard table) and a point mass (billiard ball) moving freely inside Ω, without friction and following a rectilinear path. For our aim, we assume Ω to be strictly convex.
The usual billiards are the so-called *Birkhoff billiards* in which the law of motion is precisely the standard reflection law –see Figure 1– that is

$$\text{angle of incidence} \;=\; \text{angle of reflection}$$

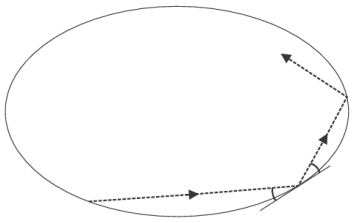


Figure 1: The Birkhoff billiard map.

The billiard map is

$$\begin{aligned} f : \partial\Omega \times (0,\pi) &\to \partial\Omega \times (0,\pi) \\ (p,\varphi) &\mapsto (p',\varphi'). \end{aligned} \tag{1.1}$$

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: nardi@math.unipd.it . Seminar held on 15 April 2025.

Given a parametrization $\gamma : \mathbb{S}^1 := \mathbb{R}/2\pi\mathbb{Z} \to \partial\Omega$, sending $t \mapsto \gamma(t)$, we can write

$$\begin{aligned} p &= \gamma(t), & r &:= -\cos\varphi, \\ p' &= \gamma(t'), & r' &:= -\cos\varphi' \end{aligned}$$

and rewrite the map (1.1) as

$$\begin{aligned} f : \mathbb{S}^1 \times (-1,1) =: \mathbb{A} &\to \mathbb{A} \\ (t,r) &\mapsto (t',r'). \end{aligned}$$

Such a map admits a generating function:

$$L(t,t') := \|\gamma(t) - \gamma(t')\|, \tag{1.2}$$

in fact $r'dt' - rdt = dL(t,t')$.

Example 1.1 Let $\partial\Omega$ be a circle. In such a case, the angle is constant along the motion, and this implies that the phase space $\mathbb{A}$ admits a continuous foliation into continuous closed invariant curves that are not null-homotopic (see Figure 2). The existence of such a foliation gives precisely what is called a *totally integrable* billiard table.

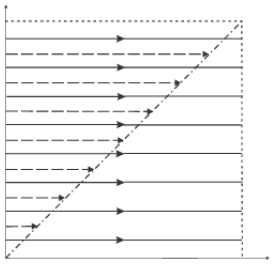


Figure 2: Phase portrait of the Birkhoff billiard map in the circle.

Example 1.2 Let $\partial\Omega$ be an ellipse. In such a case, the phase portrait is well-known in literature (see, e.g., [8]) and is represented in Figure 3. The phase cylinder $\mathbb{A}$ admits a continuous foliation into closed invariant curves. The existence of such a foliation gives precisely what is called an *integrable* billiard table. We stress that in this case, we don't ask for the non-null-homotopic assumption.

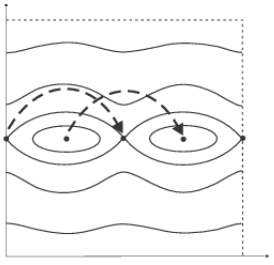


Figure 3: Phase portrait of the Birkhoff billiard map in the ellipse.

The main open problem in this field is the following conjecture, due to Birkhoff [7].

Conjecture 1.3 (Birkhoff) *The only integrable billiard tables are circles and ellipses.*

The conjecture remains open; however, there are many interesting results in this direction, including local and perturbative ones. We will mention two of them: the first is due to M. Bialy [5], while the second is quite recent and is a joint work of M. Bialy and A. Mironov [6].

Theorem 1.4 (M. Bialy, 1993) *The only totally integrable Birkhoff billiard tables are circles.*

Theorem 1.5 (M. Bialy and A. Mironov, 2022) *Let Ω be a centrally symmetric C^2 strongly-convex domain with boundary $\partial\Omega$. Assume that the Birkhoff billiard map of $\partial\Omega$ has a (simple) continuous invariant curve of rotation number $1/4$ (winding once around the phase-space) and consisting only of 4-periodic orbits. Moreover, suppose that the domain between this invariant curve and the lower boundary of the phase-space cylinder is entirely foliated by continuous closed invariant curves that are not null-homotopic. Then $\partial\Omega$ is an ellipse*

2 Symplectic billiards

In 2018, P. Albers and S. Tabachnikov [1] introduced a new class of billiards with the intent of constructing a billiard dynamics having the *area* as generating function, instead of the length (see (1.2)).

Consider $\Omega \subset \mathbb{R}^2$ a strictly convex domain whose boundary is positively oriented counterclockwise, and assume the origin O to be in the interior of Ω. Since Ω is strictly convex, for every $x \in \partial\Omega$ there exists a unique point $x^* \in \partial\Omega$ such that $T_x\partial\Omega = T_{x^*}\partial\Omega$, that is, whose tangent line to the boundary of Ω is parallel to that of x. The symplectic billiard map is defined as

$$T : \mathcal{P} \longrightarrow \mathcal{P}, \qquad (x, y) \mapsto (y, z) \iff x - z \in T_y\partial\Omega$$

where $\mathcal{P} = \{(x, y) \in \partial\Omega \times \partial\Omega : x < y < x^*\}$, according to the orientation chosen, see Figure 4.

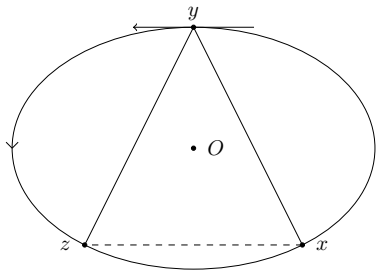


Figure 4: The symplectic billiard map.

Let ω be the standard area form, $\omega(x,y) := \det(x,y)$. This turns out to be a generating function for the symplectic billiard map, in fact

$$\begin{aligned} 0 = \frac{d}{dy}\Big[\omega(x,y) + \omega(y,z)\Big] &\iff \det(x,v) + \det(v,z) = 0 \quad v \in T_y\partial\Omega \\ &\iff \det(x-z,v) = 0 \\ &\iff T(x,y) = (y,z). \end{aligned}$$

Note also that $\omega(x,y)$ is precisely twice the area of the triangle xOy.
By means of a parametrization $\gamma : \mathbb{S}^1 \to \partial\Omega$, $t \mapsto \gamma(t)$, we can define the variables

$$\begin{cases} s_1 = -\frac{\partial\omega(t_1,t_2)}{\partial t_1} \\ s_2 = \frac{\partial\omega(t_1,t_2)}{\partial t_2}. \end{cases}$$

Then (t_1, s_1) turns out to be coordinates on $\mathcal{P}$, and we can look at the symplectic billiard map on the corresponding annulus, $T(t_1, s_1) = (t_2, s_2)$.

Remark 2.1 The symplectic billiard dynamics is invariant up to affine transformations of the plane. This occurs because the parallelism condition is invariant under affinities.

Example 2.2 Again, the easiest example to consider is circles. The symplectic billiard map acts as follows:

$$T(x,y) = (y,z) \quad \iff \quad x - z \in T_y\partial\Omega.$$

By a simple proof of synthetic geometry, this occurs if and only if the angle of incidence is equal to the angle of reflection, see Figure 5.

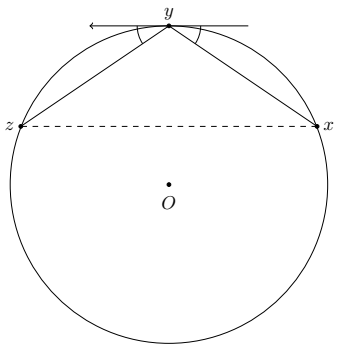


Figure 5: The symplectic billiard map in the circles.

This means that in circles

$$\text{Birkhoff billiard dynamics} = \text{Symplectic billiard dynamics}$$

and, in particular, that circular symplectic billiard tables are *totally integrable*. Moreover, by Remark 2.1, the invariance of the symplectic billiard map up to affinities directly implies that elliptic symplectic billiard tables are *totally integrable*.

Such an example leads to a natural question: are ellipses the only totally integrable billiard tables?
The answer was given by L. Baracco and O. Bernardi in [2].

Theorem 2.3 (L. Baracco, O. Bernardi, 2024) *The only totally integrable symplectic billiards are ellipses.*

The further question is: what happens if we ask for less restrictive hypotheses, i.e., if we assume the foliation for a smaller part of the phase space?
In this direction, in a joint work with L. Baracco and O. Bernardi [3], we were able to obtain the analogue of the Bialy-Mironov result.

Theorem 2.4 (L.B, O.B., A.N., 2024) *Let Ω be a centrally symmetric C^2 strongly-convex domain with boundary $\partial\Omega$. Assume that the symplectic billiard map $T : \mathcal{P} \to \mathcal{P}$ of $\partial\Omega$ has a (simple) continuous invariant curve $\delta \subset \mathcal{P}$ of rotation number $1/4$ (winding once around $\partial\Omega$) and consisting only of 4-periodic orbits. If one of the parts between δ and each boundary of the phase-space $\mathcal{P}$ is entirely foliated by continuous invariant closed (not null-homotopic) curves, then $\partial\Omega$ is an ellipse.*

These are *rigidity* results, because asking for the foliations as in the hypotheses of both previous theorems, fixes the domain to be an ellipse. Another way to look for rigidity is by means of the *area spectrum*.

Let $\{x_j\}_{j=0}^{q}$ be a periodic trajectory for the symplectic billiard map, that is $T(x_{j-1}, x_j) = (x_j, x_{j+1})$ for every $j = 1, \dots q-1$, and $x_0 = x_q$. Its *action* is defined as

$$\sum_{j=0}^{q-1} \omega(x_j, x_{j+1})$$

and if the orbit winds once around the boundary $\partial\Omega$, it is precisely twice the area of the polygon of vertices $\{x_j\}_{j=0}^{q-1}$. The area spectrum is defined as

$$\mathcal{A}(\Omega) = \mathbb{N}\{\text{action of all closed trajectories of } \Phi\} \cup \mathbb{N}\{A_\Omega\},$$

where A_Ω is the area of Ω. It is now clear that, from Remark 2.1, given two strictly convex domains Ω and Ω' with the same area, if the corresponding symplectic billiard maps T_Ω and $T_{\Omega'}$ are conjugated by an equi-affine transformation of the plane (that is affine unitary map of $\mathbb{R}^2$), then $\mathcal{A}(\Omega) = \mathcal{A}(\Omega')$.

Then a natural question arises: is it true that if $\mathcal{A}(\Omega) = \mathcal{A}(\Omega')$, then Ω and Ω' are necessarily equi-affine?
A partial answer to this question was given in a joint work with L. Baracco and O. Bernardi, [4], for two different classes of domains.

Theorem 2.5 (L.B., O.B., A.N., 2024) *Let $\mathcal{M}$ be the set of strictly convex domains with sufficiently (finitely) smooth boundary, everywhere positive curvature, axial symmetry, and sufficiently close to an ellipse. Then any $\Omega, \Omega' \in \mathcal{M}$ with the same area spectrum are necessarily equi-affine.*

Theorem 2.5 (L.B., O.B., A.N., 2024) *Let $\mathcal{M}$ be the set of strictly convex domains with sufficiently (finitely) smooth boundary, everywhere positive curvature, central symmetry, sufficiently close to an ellipse, and even-rationally integrable. Then any $\Omega, \Omega' \in \mathcal{M}$ with the same area spectrum are necessarily equi-affine.*

References

[1] Albers P.; Tabachnikov, S., *Introducing symplectic billiards.* Adv. Math. 333 (2018): 822–867.

[2] Baracco, L.; Bernardi, O., *Totally integrable symplectic billiards are ellipses.* Advances in Mathematics, Volume 454, (2024), 109873.

[3] Baracco, L.; Bernardi, O.; Nardi, A., *Bialy-Mironov type rigidity for centrally symmetric symplectic billiards.* Nonlinearity 37, (2024), 125025.

[4] Baracco, L.; Bernardi, O.; Nardi, A., *Area spectral rigidity for axially symmetric and Radon domains.* ArXiv:2410.12644, 2024.

[5] Bialy, M., *Convex billiards and a theorem by E. Hopf.* Math. Z. 214 (1993), no. 1, 147–154.

[6] Bialy, M.; Mironov. A.E., *The Birkhoff-Poritsky conjecture for centrally symmetric billiard tables.* Ann. of Math. 2, 196/1 (2022): 389–413.

[7] G.D. Birkhoff, *On the periodic motions of dynamical systems.* Acta Math. 50 (1927): 359–379.

[8] Siburg, K.F., "The principle of least action in geometry and dynamics". Lecture Notes in Mathematics, vol. 1844, xiii+128 pp. Berlin, Germany: Springer, 2004.

Inductive Methods in the Representation Theory of Finite Groups of Lie Type: An Introduction via $GL_n(q)$

Elena Collacciani (*)

Abstract. These notes aim to offer a glimpse into the techniques used to construct and classify irreducible representations of finite groups of Lie type. To convey the core ideas while avoiding heavy technicalities, we focus on the case of the General Linear group over a finite field.

The main goal of this note is to give an introduction to some of the main ideas underlying the representation theory of finite groups of Lie type.

In order to explore the fundamental ideas avoiding heavy technicalities, we will actually focus on just one family of finite groups of Lie type, namely the one of finite general linear groups (see Example 3). We aim to privilege intuition over rigorousness, and therefore plenty of space will be devoted to concrete examples, while we will often hide the abstract definition of the objects involved. Nevertheless, the main ideas and techniques developed here are the same ones underlying the general case. In particular, we place special emphasis on the role of inductive methods - a fundamental paradigm in representation theory - which allow to reduce complex algebraic problems to more manageable combinatorial ones.

The interested reader can refer to [6, 2, 1] for rigorous and self-contained expositions on the subject.

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: elena.collacciani@studenti.unipd.it . Seminar held on 8 May 2025.

1 Preliminaries

We give a quick recollection about the basic objects involved in these notes, with no claim of being exhaustive.

1.1 Groups

Definition 1.1 A group is a set G endowed with a binary operation

$$\begin{aligned} - \cdot - \ : G \times G &\to G \\ (g_1, g_2) &\mapsto g_1 \cdot g_2 \end{aligned}$$

satisfying the following properties:

- associative: for any $g_1, g_2, g_3 \in G$ it holds $(g_1 \cdot g_2) \cdot g_3 = g_1 \cdot (g_2 \cdot g_3)$
- existence of the neutral element: there exists an element $id \in G$ such that $id \cdot g = g = g \cdot id$ for any $g \in G$
- existence of the inverses: for any $g \in G$ there exists an element $h \in G$ such that $h \cdot g = e = g \cdot h$. We denote such h by g^{-1}.

A group is finite if $|G| < \infty$.
A group is abelian if it satisfies the following additional property:

- for any $g_1, g_2 \in G$ it holds $g_1 \cdot g_2 = g_2 \cdot g_1$.

Example 1 For any $n \in \mathbb{N}$, let $[1, n] := \{i \in \mathbb{N} | 1 \leq i \leq n\}$. Then

$$S_n = \{\sigma : [1, n] \to [1, n] \mid \sigma \text{ is bijective}\}$$

endowed with the composition of maps is a finite group, called the symmetric group of n elements. The group S_1, S_2 is commutative, while S_n is not commutative for any $n > 2$.

A subgroup $H \leq G$ of G is a subset of G containing the neutral element and such that the group operation $(- \cdot -)$ of G restricted to $H \times H$ lands in H.

Given a group G, we say that a subset S generates G if any element of G can be expressed as a product of elements contained in S and their inverses. We write $G = \langle S | R \rangle$, and say that $\langle S | R \rangle$ is a presentation for G, if S is a set of generators for G and R is a complete set of relations to which the generators in S are subject. Formally, $\langle S | R \rangle$ is a presentation for G if G is isomorphic to the quotient of the free group over S by the normal subgroup generated by R.

Definition 2 Let H, G be groups. A group homomorphism is a map $f : H \to G$ mapping the neutral element of H to the neutral element of G and respecting the group operations, i.e. satisfying $f(h_1 \cdot h_2) = f(h_1) \cdot f(h_2)$ for any $h_1, h_2 \in H$.

1.2 Finite fields

Definition 3 A field is a set $\mathbb{F}$ endowed with two binary operations

$$\begin{aligned} -+- &: \mathbb{F} \times \mathbb{F} \to \mathbb{F} \\ &(f_1, f_2) \mapsto f_1 + f_2 \\ -\cdot- &: \mathbb{F} \times \mathbb{F} \to \mathbb{F} \\ &(f_1, f_2) \mapsto f_1 \cdot f_2 \end{aligned}$$

satisfying the following properties:

- The set $\mathbb{F}$ endowed with the operation $+$ is an abelian group. The identity element with respect to $+$ is called 0;
- The set $\mathbb{F}^* = \mathbb{F} \setminus \{0\}$ endowed with the operation $\cdot$ is an abelian group. The identity element element with respect to $\cdot$ is called 1;
- Distributivity: for any for any $f_1, f_2, f_3 \in \mathbb{F}$ it holds $(f_1 + f_2) \cdot f_3 = f_1 \cdot f_3 + f_2 \cdot f_3$.

Theorem 1.1 *Let p be a prime number, let $e \in \mathbb{N}$, let $q = p^e$. There exists a field $\mathbb{F}_q$ such that $|\mathbb{F}_q| = q$, and it is unique up to field isomorphism.*

Example 2 If $e = 1$, i.e. $q = p$, the field of p elements $\mathbb{F}_p$ can be described as

$$\mathbb{F}_p = \{0, 1, \ldots, p-1\}$$

with operations given by the usual sum and product on integers reduced modulo p.

1.3 Finite groups of Lie type

For the sake of completeness, we give in this section a formal definition of "finite groups of Lie type" and present some examples, in order to give a sense of the range of applicability of the techniques (appropriately generalized) presented here. This definition requires more background in algebra than the one we are able to present in these notes. The reader not familiar with it can read Example 3 and skip ahead: the understanding of the remaining part is independent of this section.

Definition 4 An affine algebraic group $\mathbf{G}$ over $\overline{\mathbb{F}_p}$, the algebraic closure of $\mathbb{F}_p$, is an affine variety over $\overline{\mathbb{F}_p}$ endowed with a group structure, such that the multiplication map $(- \cdot -)$ and the inversion map defined by the assignment $g \mapsto g^{-1}$ are morphisms.
The affine algebraic group $\mathbf{G}$ is said to be reductive if it has a trivial unipotent radical, i.e. it does not contain non-trivial closed connected normal subgroups consisting of unipotent elements.

Definition 5 An affine algebraic group $\mathbf{G}$ over $\overline{\mathbb{F}_p}$ has an $\mathbb{F}_q$-rational structure, for some power q of p, if there exist:

- an affine variety $X \subseteq \overline{\mathbb{F}_q}^n$ and an isomorphism of affine varieties $\iota : X \to \mathbf{G}$;
- an algebraic group endomorphism $F_q : \mathbf{G} \to \mathbf{G}$ such that

$$F_q \circ \iota(x_1, \dots, x_n) = \iota(x_1^q, \dots, x_n^q)$$

for any $(x_1, \dots x_n)$.

The algebraic group endomorphism F_q is called Frobenius morphism for $\mathbf{G}$.
An endomorphism $F : \mathbf{G} \to \mathbf{G}$ is said to be a Steinberg map if there exists an $m \in \mathbb{N}$ such that F^m is a Frobenius map with respect to some $\mathbb{F}_q$-rational structure.

Definition 6 Let $\mathbf{G}$ be an affine algebraic group that is connected. Let F be a Steinberg map for $\mathbf{G}$. Then

$$\mathbf{G}^F := \{g \in \mathbb{G} \mid g = F(g)\}$$

is called a finite group of Lie type.

Example 3 We give some examples of finite groups of Lie type. Let $n \in \mathbb{N}$.

- Finite general linear group: the group

$$GL_n(q) = \{A \in M_n(\mathbb{F}_q) \mid det(A) \in \mathbb{F}_q^*\}$$

of invertible $n \times n$ matrices with entries in the finite field $\mathbb{F}_q$.

- Finite special linear group: the group

$$SL_n(q) = \{A \in M_n(\mathbb{F}_q) \mid det(A) = 1\}$$

of $n \times n$ matrices with entries in the finite field $\mathbb{F}_q$ and determinant 1.

- Finite special orthogonal group: the group

$$SO_n(q) = \{A \in M_n(\mathbb{F}_q) \mid A^T A = I,\ det(A) = 1\}$$

of $n \times n$ matrices with entries in the finite field $\mathbb{F}_q$ and determinant 1 preserving the standard inner product on $\mathbb{F}_q^n$.

2 Standard representation theory of finite groups

Through all this section, G is a finite group.

Definition 7 A (complex) representation of G is a pair (π, V) where

- V is a complex vector space
- $\pi : G \to GL(V)$ is a group homomorphism from G to the group of linear transformations of G

Example 4 For any finite group G, there is a one-dimensional representation $(\mathbf{1}, \mathbb{C})$, called trivial, given by

$$\begin{aligned} \mathbf{1} : G &\to GL_1(\mathbb{C}) = \mathbb{C}^* \\ g &\mapsto 1 \end{aligned}$$

Example 5 Let $G = GL_2(2)$. Then

$$G = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \right\}$$

Set

$$s := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad\qquad t := \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

Then it can be easily checked that $G = \langle s, t | s^2 = 1, t^2 = 1, sts = tst \rangle$ A representation of $G = GL_2(2)$ is $(Stnd, \mathbb{C}^3)$ given by

$$\begin{aligned} Stnd : GL_2(2) &\to GL(\mathbb{C}^3) \\ s &\mapsto \left(\begin{smallmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{smallmatrix}\right) \\ t &\mapsto \left(\begin{smallmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{smallmatrix}\right) \end{aligned}$$

extended to be a group morphism. This is well-defined since

$$\begin{aligned} &Stnd(s))^2 = 1 \\ &Stnd(t))^2 = 1 \\ &Stnd(tst) = Stnd(t)Stnd(s)Stnd(t) = \left(\begin{smallmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{smallmatrix}\right) = Stnd(s)Stnd(t)Stnd(s) = Stnd(sts). \end{aligned}$$

Definition 8 Let (π, V), (π', V') be representations of G. A morphism of representations is a linear map $T : V \to V'$ such that for any $g \in G$ it holds $T \circ \pi(g) = \pi'(g) \circ T$. We denote by $Hom_G(\pi, \pi')$ the set of all the morphisms from (π, V) to (π', V'). A morphism of representations T is an isomorphism if it is bijective.

From now on, we will often consider representations up to isomorphism, identifying isomorphic ones.

Definition 9 Let (π, V) be a representation of G. Then a representation (π', V') is a subrepresentation of (π, V) if

- V' is a linear subspace of V such that for any $g \in G$, it holds $\pi(g)V' = V'$
- for any $g \in G$, it holds $\pi'(g) = \pi(g)|_{GL(V')}$

Example 6 Assume the same notations as in Example 5. The representation $(Stnd, \mathbb{C}^3)$ has a subrepresentation isomorphic to $(\mathbf{1}, \mathbb{C})$. Indeed, we realize $\mathbb{C}$ as a subspace of $\mathbb{C}^3 = \langle e_1, e_2, e_3\rangle_{\mathbb{C}}$ as $V' = \langle e_1 + e_2 + e_3\rangle_{\mathbb{C}} = \{(x, x, x) | x \in \mathbb{C}\} \subset \mathbb{C}^3$ by

$$\begin{aligned} \iota : \mathbb{C} &\hookrightarrow \mathbb{C}^3 \\ x &\mapsto (x, x, x) \end{aligned}$$

Then for any $x \in \mathbb{C}$

$$\begin{aligned} Stnd(s)\iota(x) = Stnd(s)(x, x, x) = (x, x, x) = \iota(\mathbf{1}(s)(x)) \\ Stnd(t)\iota(x) = Stnd(t)(x, x, x) = (x, x, x) = \iota(\mathbf{1}(t)(x)) \end{aligned}$$

Theorem 2.1 *Let (π, V) be a representation of G, and let (π_1, V_1) be subrepresentation of (π, V). Then there exists a subrepresentation (π_2, V_2) of (π, V) such that $V = V_1 \oplus V_2$, and it is unique up to isomorphism.*

In the notation of Theorem 2.1, we call the representation (π_2, V_2) the complement of (π_1, V_1) in (π, V), and we write $(\pi, V) = (\pi_1, V_1) \oplus (\pi_2, V_2)$ or sometimes, omitting the vector spaces, $\pi = \pi_1 \oplus \pi_2$.

Example 7 Assume notation from Example 5. We saw in Example 6 that $(\mathbf{1}, \langle e_1, e_2, e_3\rangle_{\mathbb{C}})$ is a subrepresentation of $(Stnd, \mathbb{C}^3)$. Its complement is given by the subspace $V_2 = \langle e_1 - e_3, e_2 - e_3\rangle_{\mathbb{C}}$. It is G stable, so it defines a subrepresentation:

$$\begin{aligned} &Stnd(s)(e_1 - e_3) = e_2 - e_3 \in V_2 && Stnd(s)(e_2 - e_3) = e_1 - e_3 \in V_2 \\ &Stnd(t)(e_1 - e_3) = e_1 - e_2 = (e_1 - e_3) - (e_2 - e_3) \in V_2 && Stnd(s)(e_2 - e_3) = e_3 - e_2 = e_2 - e_3 \in V_2 \end{aligned}$$

and it holds $\mathbb{C}^3 = \langle e_1, e_2, e_3\rangle_{\mathbb{C}} \oplus \langle e_1 - e_3, e_2 - e_3\rangle_{\mathbb{C}}$.

So considering the inclusion of $\mathbb{C}^2$ as a subspace of $\mathbb{C}^3$ by

$$\begin{aligned} \iota : \mathbb{C}^2 &\to \mathbb{C}^3 \\ e_1 &\mapsto e_1 - e_3 \\ e_2 &\mapsto e_2 - e_3, \end{aligned}$$

we have that the complement in $(Stnd, \mathbb{C}^3)$ of the subrepresentation $(\mathbf{1}, \mathbb{C})$ is (isomorphic to) the Steinberg representation , that is the representation $(St, \mathbb{C}^2)$ defined by

$$\begin{aligned} St : G &\to GL(\mathbb{C}^2) \\ s &\mapsto \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right) \\ t &\mapsto \left(\begin{smallmatrix} 1 & 0 \\ -1 & -1 \end{smallmatrix}\right) \end{aligned}$$

Definition 10 A representation of G is said to be irreducible if it has no non-trivial proper subrepresentations. We write

$$Irr(G) := \{\text{irreducible representations of } G\}/_{\sim}$$

where $\sim$ is the equivalence relation given by isomorphism of representations.

Example 8

- Any one dimensional representation is irreducible. In particular, the trivial representation $(\mathbf{1}, \mathbb{C})$ is irreducible.
- Let $G = GL_2(2)$. The representation $(St, \mathbb{C}^2)$ in Example 7 is irreducible. Indeed a subrepresentation, if existsting, should be one dimensional, and there is no line stabilized by $St(g)$ for any $g \in G$. Indeed the only line stabilized by $St(s)$ is the line $\langle e_1 + e_2 \rangle_{\mathbb{C}}$, but it is not stabilized by t, since $St(t)(e_1 + e_2) = e_1$.

Corollary 2.2 (Maschke Theorem) *Let (π, V) be a representation of G. Then it decomposes as direct sum of irreducible subrepresentations, and the decomposition is unique up to isomorphism of irreducible representations.*

We write $(\pi, V) = \bigoplus_{i=1}^{k} (\pi_i, V_i)^{m_i}$ with $(\pi_i, V_i) \in Irr(G)$ and $m_i \in \mathbb{Z}_{\geq 0}$ for such a decomposition. Sometimes we omit the vector spaces and write just $\pi = \bigoplus_{i=1}^{k} \pi_i^{m_i}$,

Example 9 Let $G = GL_2(2)$. Then $(Stnd, \mathbb{C}^2) = (\mathbf{1}, \mathbb{C}) \oplus (St, \mathbb{C}^2)$.

3 Representation theory of $GL_n(q)$

In this section, let p be a prime and $q = p^e$, with $e \in \mathbb{N}$, be a power of p.

Moreover let $n \in \mathbb{N}$, and let $G = GL_n(q)$, the finite general linear group over the field with q elements (see Example 3.
Our goal is to understand the structure of the representations of G. By Maschke's Theorem, it is enough to understand $Irr(G)$, the irreducible representations of G.
The following remark gives a classification for irreducible representation in the easiest case possible, namely the case of $n = 1$.

Remark 1 We know $Irr(GL_1(\mathbb{F}_q))$. Indeed $GL_1(\mathbb{F}_q) = \mathbb{F}_q^* = \langle \zeta | \zeta^{q-1} = 1 \rangle$, a cyclic group of order $q - 1$. This is an abelian group, and it is known that irreducible representations of abelian groups are one-dimensional. For any $i \in \{1 \dots q-1\}$ we have an irreducible representation defined by

$$\begin{aligned} \chi_i : GL_1(\mathbb{F}_q) &\to GL_1(\mathbb{C}) \\ \zeta &\mapsto e^{\frac{2\pi i}{q-1}} \end{aligned}$$

and it holds

$$Irr(GL_1(\mathbb{F}_q)) = \{\chi_i |\ 1 \leq i \leq q-1\}$$

3.1 Parabolic induction

The guiding principle is to develop methods allowing us to proceed by induction. More precisely, the question would be: if we know the irreducible representations of $G = GL_m(q)$ for $m < n$, are we able to say something about the irreducible representations of $GL_n(q)$? In this section we introduce a tool, parabolic induction, allowing us to bridge knowledge from smaller to bigger groups.

Let $k \in \mathbb{N}$ and $n_1, n_2, \dots, n_k$ be natural numbers such that $\sum_{i=1}^{k} n_i = n$. We denote by

$$(1) \qquad L_{n_1,\dots n_k} := \left\{ \begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_k \end{pmatrix} \in GL_n(q) \;\middle|\; A_i \in GL_{n_i}(q),\ i = 1, \dots, k \right\} \leq G$$

the subgroup of G consisting of block diagonal matrices with k blocks of sizes prescribed by the integers $n_1, \dots n_k$. It holds

$$L_{n_1,\dots n_k} \cong GL_{n_1}(q) \times GL_{n_2}(q) \times \cdots \times GL_{n_k}(q)$$

Proposition 3.1 *There is a bijection*

$$\begin{aligned} Irr(GL_{n_1}(q)) \times \cdots \times Irr(GL_{n_k}(q)) &\to Irr(GL_{n_1}(q) \times \cdots \times GL_{n_k}(q)) \\ ((\pi_1, V_1), \dots, (\pi_k, V_k)) &\mapsto (\pi_1 \otimes \cdots \otimes \pi_k, V_1 \otimes \cdots \otimes V_k) \end{aligned}$$

where for any $(g_1, \dots, g_k) \in GL_{n_1}(q) \times \cdots \times GL_{n_k}(q)$ *and any* $(v_1, \dots, v_k) \in V_1 \otimes \cdots \otimes V_k$ *it holds*

$$\pi_1 \otimes \cdots \otimes \pi_k(g_1, \dots, g_k)(v_1, \dots, v_k) = \pi_1(g_1)v_1 \otimes \cdots \otimes \pi_k(g_k)v_k.$$

Since $L_{n_1,\dots n_k} \cong GL_{n_1}(q) \times GL_{n_2}(q) \times \cdots \times GL_{n_k}(q)$, it follows that there is a bijection

$$Irr(GL_{n_1}(q)) \times \cdots \times Irr(GL_{n_k}(q)) \to Irr(L_{n_1,\dots n_k}).$$

In other words, any irreducible representation of the subgroup $L_{n_1,\dots n_k}$ decomposes as a tensor product of irreducible representations of $GL_{n_i}(q)$. Hence, if we know $Irr(GL_m(q))$ for any $m < n$, we know $Irr(L)$ for any subgroup $L = L_{n_1,\dots n_k}$ of the form (1).

Now we would like to build a representation of G starting from representations of a subgroup L of the form (1).

Generally speaking, there is a way, called induction, to construct a representation of a group starting from a representation of a subgroup. The problem in this case is that L is "too small" to get a representation with a nice structure by induction. The work-around is to use an intermediate subgroup, consisting of block upper triangular matrices.

Let $k \in \mathbb{N}$ and $n_1, n_2, \dots, n_k \in \mathbb{N}$ be such that $\sum_{i=1}^{k} n_i = n$. We denote by

$$P_{n_1,\dots n_k} := \left\{ \begin{pmatrix} A_{11} & A_{1,2} & \cdots & A_{1,4} \\ & A_{22} & \cdots & A_{2,4} \\ & & \ddots & \vdots \\ & & & A_{k,k} \end{pmatrix} \in GL_n(q) \;\middle|\; A_{i,j} \in GL_{n_i}(q),\ 1 \le i \le j \le k \right\} \le G$$

It holds $L_{n_1,\dots n_k} \le P_{n_1,\dots n_k}$. More precisely, denoting

$$U_{n_1,\dots n_k} := \left\{ \begin{pmatrix} I_{n_1} & A_{1,2} & \cdots & A_{1,4} \\ & I_{n_2} & \cdots & A_{2,4} \\ & & \ddots & \vdots \\ & & & I_{n_k} \end{pmatrix} \in GL_n(q) \;\middle|\; A_{i,j} \in GL_{n_i}(q),\ 1 \le i \le j \le k \right\} \le P_{n_1 \dots n_k},$$

where I_{n_i} denotes the $n_i \times n_i$ identity matrix, it holds

$$P_{n_1,\dots n_k} = L_{n_1,\dots n_k} U_{n_1,\dots n_k}$$

Definition 11 Let $k \in \mathbb{N}$ and $n_1, n_2, \dots, n_k \in \mathbb{N}$. Let $L = L_{n_1 \dots n_k}$ and $U = U_{n_1 \dots n_k}$ and $P = P_{n_1 \dots n_k}$. Let (π, V) be a representations of L. The parabolic induction of (π, V) is the representation $(R_L^G \pi, R_L^G V)$ of G given by:

- $R_L^G V := \{ f : G \to V |\ f(lux) = \pi(l) f(x) \text{ for any } l \in L, u \in U, x \in G \}$
- for any $g \in G$, for any $f \in R_L^G V$

  $$(R_L^G \pi(g) f)(x) = f(xg)$$

  for any $x \in G$.

Example 10 Let $k = n$, $n_i = 1$ for any $i = 1 \dots n$. Then

$$L_{1^n} := T = \left\{ \begin{pmatrix} a_1 & & & \\ & a_2 & & \\ & & \ddots & \\ & & & a_n \end{pmatrix} | a_i \in \mathbb{F}_q,\ i = 1, \dots, n \right\} \le GL_n(q) \ \Big\}$$

$$P_{1^n} := B = \left\{ \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ & a_{22} & a_{23} & a_{24} \\ & & \ddots & \vdots \\ & & & a_{nn} \end{pmatrix} | a_{ij} \in \mathbb{F}_q, 1 \le i \le j \le n \right\} \le GL_n(q) \ \Big\}$$

$$U_{1^n} := U = \left\{ \begin{pmatrix} 1 & a_{12} & a_{13} & a_{14} \\ & 1 & a_{23} & a_{24} \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix} | a_{ij} \in \mathbb{F}_q, 1 \le i < j \le n \right\} \le GL_n(q) \ \Big\}$$

We have $Irr(T) \cong Irr(GL_1(q)) \times \cdots \times Irr(GL_1(q)) = Irr(\mathbb{F}_q^*)^n$. In particular, taking the trivial representation on each $\mathbb{F}_q^*$, we get the irreducible representation $(\mathbf{1}, \mathbb{C})$ of T given by $\mathbf{1}(t) = 1$ for any $t \in T$. Then

$$R_T^G(\mathbb{C}) := \{f : G \to \mathbb{C} | f(tux) = f(x) \text{ for any } t \in T, u \in U\}$$

that is the space of functions constant on $B = TU$ left coset, so

$$R_T^G(\mathbb{C}) = \langle \delta_{Bx} | Bx \in B\backslash G \rangle_{\mathbb{C}}$$

where

$$\delta_{Bx}(y) = \begin{cases} 1 & y \in Bx \\ 0 & y \notin Bx \end{cases}.$$

The action of G is given by

$$R_T^G\mathbf{1}(g)\delta_{Bx} = \delta_{Bxg^{-1}}$$

since

$$R_T^G\mathbf{1}(g)\delta_{Bx}(By) = \delta_{Bx}(Byg) = \delta_{Bxg^{-1}}(By)$$

Take now $n = 2$, so $G = GL_2(q)$. Then $B\backslash G$ is isomorphic to the projective space of $\mathbb{F}_q^2$, with isomorphism given by

$$\begin{aligned} B\backslash G &\to \mathbb{P}^1(\mathbb{F}_q) \\ B\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) &\mapsto [c, d] \end{aligned}$$

with the isomorphism respecting the action of G given by right multiplication on both spaces. Therefore in this case we have $R_T^G(\mathbb{C}) = \langle \delta_x | x \in \mathbb{P}^(\mathbb{F}_q) \rangle_{\mathbb{C}}$ with action given by $R_T^G\mathbf{1}(g)\delta_x = \delta_{xg^{-1}}$.

In particular if $q = 2$, we have $\mathbb{P}^1(\mathbb{F}_2) = \{[0, 1], [1, 0], [1, 1]\}$ and so

$$R_T^G\mathbb{C} = \langle \delta_{[0,1]}, \delta_{[1,0]}, \delta_{[1,1]} \rangle_{\mathbb{C}}.$$

With notation as in Example 5, $GL_2(2) = \langle s = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right), t = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right) \rangle$. For any $[a, b] \in \mathbb{P}^1(\mathbb{F}_2)$ it holds

$$[a, b]s^{-1} = [b, a] \qquad\qquad [a, b]t^{-1} = [a, a + b].$$

Hence we can compute the action of $R_T^G(\mathbf{1}(s)$ and $R_T^G(\mathbf{1}(t)$ on the basis vector $delta_{[0,1]}, \delta_{[1,0]}, \delta_{[1,1]}$ and it yields

$$\begin{aligned} R_T^{GL_2(2)}\mathbf{1} : GL_2(2) &\to GL(R_T^G\mathbb{C}) \\ s &\mapsto \left(\begin{smallmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{smallmatrix}\right) \\ t &\mapsto \left(\begin{smallmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{smallmatrix}\right) \end{aligned}$$

That is $R_T^{GL_2(2)}\mathbf{1} = Stnd$ from Example 5.

3.2 Decomposition into Harish-Chandra series

From Example 10 we see that even starting with an irreducible representation (π, V) of a subgroup L of the form (1), its parabolic induction $(R_L^G\pi, R_L^G V)$ is not irreducible in general. But by Maschke's Theorem (Corollary 2.2) every representation decomposes as a direct sum of irreducible ones, so we can consider the irreducible constituent appearing in a representation parabolically induced.

The first natural question one could ask is the following: running over all the subgroups L of G of the form (1) and all their irreducible representations $\pi \in Irr(L)$, do we obtain all the irreducible representations of G as irreducible constituents of some parabolically induced representation $R_L^G\pi$? The answer is negative, and this gives rise to the following definition.

Definition 12 An irreducible representation σ of G is said to be cuspidal if it does not appear as an irreducible constituent of $R_L^G\pi$ for any $L = L_{n_1,\dots n_k} < G$ proper subgroup of G of the form (1) and any $\pi \in Irr(L)$.

Example 11 Let $G = GL_2(2)$. Let $(sgn, \mathbb{C})$ be the representation given by

$$\begin{aligned} GL_2(2) &\to \mathbb{C}^* \\ s &\mapsto -1 \\ t &\mapsto -1 \end{aligned}$$

This is a cuspidal representation.

Indeed, the only proper subgroup of G of the form (1) matrices is T, the subgroup of diagonal matrices. Since the only diagonal matrix in G is the identity, we have $T = \{\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)\}$, and therefore in this case $Irr(T) = \{\mathbf{1}\}$. We have seen in Example 10 and Example 9 that in this case $R_T^G\mathbf{1} = Stnd = \mathbf{1} \oplus St$. So sgn is not an irreducible constituent of any (properly) parabolically induced representation.

We extend the definition of cuspidal representations to any subgroups $L = L_{n_1,\dots n_K} \cong GL_{n_1}(q) \times \dots \times GL_{n_k}(q)$ of the form (1) by saying that a representation $\sigma = \sigma_1 \otimes \dots \otimes \sigma_k$ is cuspidal for L if each σ_i is cuspidal for $GL_{n_i}(q)$ for $i = 1 \dots k$.

Theorem 3.2 *Let $\pi \in Irr(G)$. Then there exists a pair (L, σ) consisting of a subgroup $L = L_{n_1,\dots n_k} \leq G$ of the form (1) and a cuspidal representation σ of L such that π is an irreducible constituent of $R_L^G\pi$. The pair (L, σ) is unique up to simultaneous permutation of the n_i (i.e. the diagonal blocks of L) and of the σ_i (i.e. the factors of σ) .*

Note that the theorem includes the case of cuspidal representation of G, since it admits the case $L = L_n = G$.

So we have

$$Irr(G) = \bigsqcup_{(L,\sigma)} \{\pi \in Irr(G) |\ \pi \leq R_L^G\sigma\}$$

where (L, σ) runs over the pairs consisting of a subgroup $L = L_{n_1,\dots n_k}$ of G the form (1) and a cuspidal representation σ of L, taken up to simultaneous permutation of the blocks

of L and the factors of σ. The sets of irreducible constituents in the decomposition above of $Irr(G)$ are called Harish-Chandra series.

Example 12 Let $G = GL_2(2)$. By Example 11, we have that the decomposition in Harish-Chandra series for G is given by

$$Irr(GL_2(2)) = \{\mathbf{1}, St\} \bigsqcup \{sgn\},$$

where the first set corresponds to the pair $(T, \mathbf{1})$, the second set to the pair (G, sgn).

So classify $Irr(G)$ reduces to the following two problems:

- Classify cuspidal representations of $GL_m(q)$ for any $m \leq n$
- Understand the decomposition in irreducible representations of the parabolic induction of a cuspidal representation

The first problem is difficult, and has been solved for any finite group of Lie type, thanks to Deligne-Lusztig theory [5]. In the case of the finite general linear group $GL_n(q)$, a classification of cuspidal representations in terms of irreducible polynomials of degree n with coefficients in $\mathbb{F}_q$ was already known thanks to the work of Green [3]. We will focus now on the second problem.

3.3 Decomposition of parabolically induced representations

In this section let $L = L_{n_1 \dots n_k}$ a subgroup of G of the form (1), and σ be a cuspidal representation of L.

In the following we have a small digression involving algebras. The reader unfamiliar with such a structure can skip to Theorem 3.5: this is the main result, and its statement does not require any additional knowledge to the ones assumed so far.

We denote $End_G(R_L^G\sigma) = Hom_G(R_L^G\sigma, R_L^G\sigma)$ the representation homomorphisms from the parabolic induction of σ to itself. Since representation homomorphisms are linear, the set $End_G(R_L^G\sigma$ has naturally a vector space structure over $\mathbb{C}$, and moreover it is an algebra, with product given by composition.

We can define the representation of an algebra A similarly as representations of groups, as algebra morphisms $\phi : A \to End(V)$ for some vector space V over $\mathbb{C}$. Analogously to the group case, a subrepresentation of A is given by a subspace $W \leq V$ such that $\phi(A)W \subseteq W$, and a representation is said to be irreducible if it has no subrepresentation. If $A = \mathbb{C}G$ is the group algebra of a finite group, then the representations of A are the same as the representations of G. The following theorem is a particular case of a general result, valid in any semisimple category (i.e. category where an analogous of Maschke's theorem holds).

Theorem 3.3 *There is a bijection*

$$\{\pi \in Irr(G) |\ \pi \leq R_L^G\sigma\} \leftrightarrow Irr(End_G(R_L^G\sigma))$$

Moreover, the algebra $End_G(R_L^G\sigma)$ has a really nice structure.

Theorem 3.4 ([4]) *Let $L = L_{n_1,n_2,\dots n_k}$ and $\sigma \in Irr(L)$ be a cuspidal representation. Write $\sigma = \sigma_1 \otimes \dots \otimes \sigma_k \in Irr(L)$. Then there exist $r_1, \dots r_l$ natural numbers, with $\sum_{i=}^{l} r_l = k$ such that*

$$\begin{aligned} n_1 &= \dots = n_{r_1}, & \sigma_1 &\cong \dots \cong \sigma_{r_1}, \\ n_{r_1+1} &= \dots = n_{r_1+r_2}, & \sigma_{r_1+1} &\cong \dots \cong \sigma_{r_1+r_2}, \\ n_{\sum_{i=1}^{l-1} r_i+1} &= \dots = n_k & \sigma_{\sum_{i=1}^{l-1} r_i+1} &\cong \dots \cong \sigma_k. \end{aligned}$$

It holds

$$End_G(R_L^G\sigma) \cong \mathbb{C}[S_{r_1} \times \dots \times S_{r_l}]$$

Concatenating Theorems 3.3 and 3.4 gives the following

Theorem 3.5 *There is a bijection*

$$\{\pi \in Irr(G) |\ \pi \leq R_L^G\sigma\} \leftrightarrow Irr(S_{r_1} \times \dots \times S_{r_l})$$

The set of the right hand side is well understood from a combinatorial point of view. We have $Irr(S_{r_1} \times \dots \times S_{r_l}) \cong Irr(S_{r_1}) \times \dots \times Irr(S_{r_l})$. It is well known that for any $r \in \mathbb{N}$ there exists bijection

$$Irr(S_r) \leftrightarrow Par(r)$$

where $Par(r)$ denotes the set of the partitions of r, that is

$$Par(r) = \{\lambda = (\lambda_1, \dots \lambda_j) | j \in \mathbb{N}, \lambda_i \in \mathbb{N}, \ \sum_{i=1}^{j} \lambda_i = n \ \text{ and } \ \lambda_i \leq \lambda_{i+1} \text{ for } 1 \leq i \leq j\}$$

Example 13 Let $G = GL_2(q)$. Let $L = L_{1,1} = T = \mathbb{F}_2^* \times \mathbb{F}_2^*$, and $\mathbf{1} = \mathbf{1} \otimes \mathbf{1} \in Irr(T)$. In this case, using the same notation as in Theorem 3.5 we have $l = 1$ and $r_1 = 2$. Therefore

$$\{\pi \in Irr(G) |\ \pi \leq R_T^G\mathbf{1}\} \leftrightarrow Irr(S_2) \leftrightarrow Par(2) = \{(2), (1,1)\}$$

In particular, for the case $q = 2$ this agrees with our result from Example 10: $R_T^{GL_2(q)}\mathbf{1} = Stnd = \mathbf{1} \oplus St$, so it has 2 irreducible components.

More in general, for $G = GL_n(q)$ we have

$$\{\pi \in Irr(G) |\ \pi \leq R_T^G\mathbf{1}\} \leftrightarrow Irr(S_n) \leftrightarrow Par(n)$$

Example 14 On the other side of the spectrum with respect to the previous example, we have the case in which all the factors in the cuspidal representation σ of L are distinct. In this case $r_1 = \cdots = r_l = 1$, so

$$\{\pi \in Irr(G) |\ \pi \leq R_L^G \sigma\} \leftrightarrow Irr(S_1 \times \ldots \times S_1) = \{\mathbf{1}\}$$

and hence $R_L^G \sigma$ is irreducible.

For instance, let $G = GL_n(q)$ with $q > n$. Let $L = L_{1^n} = T = (\mathbb{F}_q^*)^n$ and, using the same notation as in Remark 1, let $\sigma = \chi_1 \otimes \cdots \chi_n \in Irr(T)$. In this case the χ_i for $1 \leq i \leq n$ are pairwise distinct irreducible representations of $\mathbb{F}_q^*$, and so $R_T^G \sigma$ is irreducible.

References

[1] Roger W. Carter, "Finite groups of Lie type". Wiley, 1993.

[2] Françoise Digne and Jean Michel, "Representations of Finite Groups of Lie Type". Cambridge University Press, 2020.

[3] James A. Green, *The characters of the finite general linear groups.* Transactions of the American Mathematical Society, 80(2):402–447, 1995.

[4] Robert B. Howlett and Gustav I. Lehrer, *Induced cuspidal representations and generalised Hecke rings.* Inventiones Mathematicae, 58: 37–64, 1980.

[5] George Lusztig, "Characters of Reductive Groups Over a Finite Field". Annals of Mathematics Studies. Princeton University Press, 1984.

[6] Gunter Malle and Meinolf Geck, "The Character Theory of Finite Groups of Lie Type". Cambridge University Press, 1991.

Parameter Estimation of Integrated Fractional Brownian Motions with Application to Energy Markets

Marco Mastrogiovanni (*)

This seminar is based on joint work with Yuliya Mishura (Taras Shevchenko National University of Kyiv), Stefania Ottaviano and Tiziano Vargiolu (Department of Mathematics "Tullio Levi Civita", University of Padova)

1 Introduction

The modeling of electricity prices has gained growing attention due to the unique characteristics of energy markets. Fractional Brownian motion (fBm) has been used extensively for modeling purposes thanks to its capacity to capture both short-range and long-range dependencies. However, an important inconsistency arises in many applications: while fBm is assumed to model observed price paths, these paths are typically averages of high-frequency data. As such, daily prices—often used in statistical models—do not directly reflect fBm realizations.

To overcome this, we consider a more suitable process: the integrated fractional Brownian motion (IfBm), representing a time-averaged fBm. This adjustment provides a more realistic framework for modeling electricity prices and allows the derivation of consistent estimators for the underlying parameters.

This seminar begins with an introduction to fractional Brownian motion (fBm), its main characteristics and an overview of electricity markets. We then introduce the integral-mean process of fBm and analyze the impact of time-averaging on its properties. Using ergodic theory, we construct strongly consistent estimators for the Hurst parameter adapted to the averaged process and validate them through an extensive simulation study.

Next, we extend our approach to linear combinations of two distinct time-averaged fBm processes, again estimating the relevant parameters. Finally, we apply our methodology

(*)Ph.D. in Mathematics and Modeling, Università degli Studi dell'Aquila, Dip. Ingegneria, Scienze dell'Informazione e Matematica (DISIM), via Vetoio, I-67100 L'Aquila, Italy. E-mail: `marco.mastrogiovanni@graduate.univaq.it` . Seminar held on 22 May 2025.

to empirical electricity spot price data and discuss potential future developments in this research area.

2 Fractional Brownian Motion

Fractional Brownian motion B_t^H is a centered Gaussian process defined by:

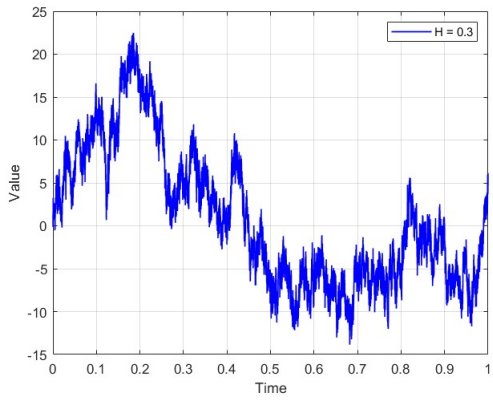
$$\mathbb{E}[B_t^H B_s^H] = \frac{1}{2}\left(t^{2H} + s^{2H} - |t-s|^{2H}\right), \tag{1}$$

with Hurst parameter $H \in (0,1)$. For $H = 1/2$, we recover standard Brownian motion. Unlike Brownian motion, fBm is not a Markov process nor a semimartingale for $H \neq 1/2$, which has important implications for its mathematical treatment.

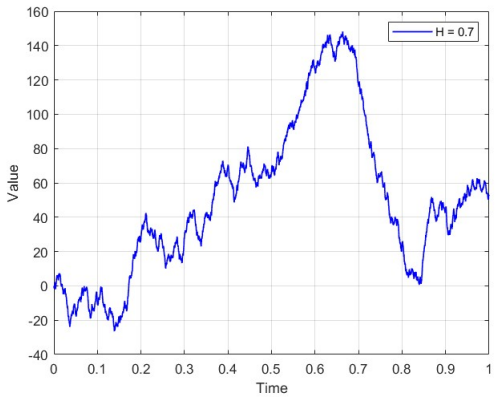
fBm possesses the following properties:

- **Self-similarity**: $B_{ct}^H \overset{d}{=} c^H B_t^H$
- **Stationary increments**: the distribution of the increments is invariant under time shifts.
- **Hölder continuity**: paths are almost surely Hölder continuous of any order $\alpha < H$
- **Long-range dependence**: for $H > 1/2$, increments are positively correlated whereas for $H < 1/2$, the increments are negatively correlated.

The Hurst parameter H determines both the smoothness and memory properties of the process. A process with $H > 1/2$ is said to be persistent, while $H < 1/2$ indicates anti-persistence. The trajectories become smoother as H approaches 1, and rougher as H approaches 0.



(a) Fractional Brownian motion with $H = 0.3$



(b) Fractional Brownian motion with $H = 0.7$

Figure 1: Comparison of fBm paths with different Hurst parameters.

3 Motivation from Electricity Markets

The market in which electricity prices are determined is known as the day-ahead market. In this market, prices are established through an auction process for blocks of electricity that will be delivered the following day. Buyers, market participants looking to purchase electricity, must submit their bids between 8:00 a.m. and 12:00 noon on the day before delivery. These bids cover hourly blocks, meaning each buyer submits 24 separate offers, each with a different price and quantity for each hour of the following day. These are all submitted simultaneously. On the other side, sellers also submit their offers, specifying how much electricity they are willing to supply and at what price. The market price for each hour is determined by the intersection of the aggregate demand and supply curves.

Once the market price is established for a given hour, all electricity traded during that hour is settled at this single market price, regardless of the individual bids and offers.

There are different factors that can influence the spot price, including daily and seasonal demand fluctuations, the availability of generation capacity, particularly from renewable energy sources like wind and solar, which are intermittent and weather-dependent, the cost and limitations of energy storage. Additionally, the prices of fuels used in electricity generation, such as natural gas, coal, and oil, can vary and significantly affect spot prices.

As a result, spot prices exhibit seasonal patterns, high volatility with their own seasonal behavior, frequent spikes, and some dependence on past prices.

In the literature, the price is commonly modeled as:

$$P(t) = f(t) + S(t), \tag{2}$$

where $f(t)$ is a deterministic seasonal component and $S(t)$ a stochastic component. There exists a wide range of models in the literature to represent the stochastic component $S(t)$, including one-factor and two-factor models, as well as approaches based on fractional Brownian motion (fBm). Our claim is that the fundamental dynamics of the stochastic component in electricity spot prices can be effectively captured by fBm.

Indeed, several empirical studies have adopted fBm-based models and found them to align reasonably well with the behavior of daily electricity prices, particularly when spikes and strong seasonal components are excluded (see, e.g., [7], [3]). However, such approaches overlook an important structural feature of electricity price formation: prices are typically not observed continuously but are computed as averages over high-frequency data. In most electricity markets, the basic traded interval is much finer—typically hourly or 15-minute intervals in Europe, and even as short as 5 minutes in some Australian markets.

Thus, if one models the underlying high-frequency prices as a realization of an fBm, the daily prices, being averages of these, do not follow an fBm themselves but rather correspond to integrated (time-averaged) fBm paths. This transformation significantly alters the statistical properties of the process, particularly the autocorrelation structure, and invalidates standard estimation techniques designed for fBm. A more consistent approach requires modeling the observed prices using integrated fBm and adapting inference methods accordingly.

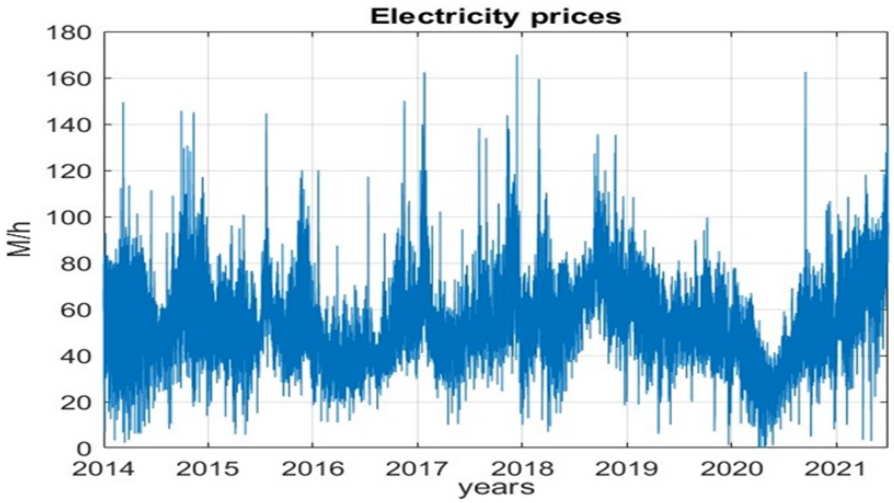


Figure 2: Hourly electricity prices, from January 2014 until March 2021.

4 Integrated Fractional Brownian Motion

We define the normalized integrated fractional Brownian motion (IfBm) as:

$$X_t^{h,H} = \frac{1}{h}\int_t^{t+h} B_s^H ds, \tag{3}$$

where B^H is a fBm with Hurst parameter H. The variance of $X_t^{h,H}$ is:

$$\mathrm{Var}[X_t^{h,H}] = \frac{1}{h^2}\int_t^{t+h}\int_t^{t+h} \mathbb{E}[B_s^H B_r^H] ds dr. \tag{4}$$

The increment process $\Delta X_k^{h,H} = X_{(k+1)h}^{h,H} - X_{kh}^{h,H}$ is stationary and Gaussian. Though it inherits some features from fBm, its autocovariance structure differs significantly. The decay rate of autocovariance depends on H and is slower for $H > 1/2$, similar to fBm. However, it exhibits a different covariance structure, and for this reason, it is necessary to construct alternative estimators.

5 Statistical Estimation via Ergodicity

To estimate the parameters of the integrated fractional Brownian motion (IfBm) model, we exploit the ergodic properties of the increment process. Assuming stationarity and Gaussianity, the ergodic theorem allows us to derive strongly consistent estimators for the model parameters.

We consider a process given by a linear combination of two independent IfBm components:

$$\widetilde{X}_k^h = aX_k^{h,H_1} + bX_k^{h,H_2}, \quad k \in \mathbb{N}, \tag{5}$$

where $a, b \in \mathbb{R}$ and X_k^{h,H_i} is the normalized IfBm associated with the Hurst parameter H_i, for $i = 1, 2$.

Our goal is to estimate the parameter vector

$$\theta = (H_1, H_2, a^2, b^2).$$

To this end, we apply the ergodic theorem to the increments of the process $\widetilde{X}_k^h$. We construct four estimators, one for each parameter, using the following family of statistics:

$$\xi_N^{(j)} := \frac{1}{N} \sum_{k=0}^{N-1} \left(\widetilde{X}_{k+1}^{jh} - \widetilde{X}_k^{jh} \right)^2, \quad j = 1, 2, 4, 8. \tag{6}$$

The choice of four different time scales provides sufficient information to identify the four parameters through a system of nonlinear equations. The consistency of these estimators is ensured by the ergodic theorem:

$$\hat{\theta}_N \xrightarrow{a.s.} \theta \quad \text{as } N \to \infty. \tag{7}$$

We construct these strongly consistent estimators of parameters $\theta = (H_1, H_2, a^2, b^2)$. Introduce the following notations

$$\log_{q+} x = \begin{cases} \log_q x, & \text{if } x > 0, \\ 0, & \text{if } x \le 0, \end{cases} \qquad \log x = \log_e x, \qquad \sqrt[+]{x} = \begin{cases} \sqrt{x}, & \text{if } x > 0, \\ 0, & \text{if } x \le 0. \end{cases}$$

Also, denote

$$D_N = (\xi_N^4 \xi_N^2 - \xi_N^8 \xi_N^1)^2 - 4(\xi_N^4 \xi_N^1 - (\xi_N^2)^2)(\xi_N^8 \xi_N^2 - (\xi_N^4)^2),$$

$$x_N = \frac{\xi_N^8 \xi_N^1 - \xi_N^4 \xi_N^2 + \sqrt[+]{D_N}}{2(\xi_N^4 \xi_N^1 - (\xi_N^2)^2)}, \qquad y_N = \frac{\xi_N^8 \xi_N^1 - \xi_N^4 \xi_N^2 - \sqrt[+]{D_N}}{2(\xi_N^4 \xi_N^1 - (\xi_N^2)^2)}$$

Theorem 1 *Let $0 < H_2 < H_1 < 1$. The random vector $\hat{\theta} = (\hat{H}_{1,N}, \hat{H}_{2,N}, \hat{a}_N^2, \hat{b}_N^2)$, where*

$$\hat{H}_{1,N} = \frac{1}{2 \log 2} \log_+ (x_N), \tag{8}$$

$$\hat{H}_{2,N} = \frac{1}{2 \log 2} \log_+ (y_N), \tag{9}$$

$$\hat{a}_N^2 = \frac{(2\hat{H}_{1,N} + 1)(\hat{H}_{1,N} + 1)(\xi_N^2 - y_N \xi_N^1)}{2h^{2\hat{H}_{1,N}}(x_N - y_N)(x_N - 1)},$$

and

$$\hat{b}_N^2 = \frac{(2\hat{H}_{2,N} + 1)(\hat{H}_{2,N} + 1)(\xi_N^2 - x_N \xi_N^1)}{2h^{2\hat{H}_{2,N}}(y_N - x_N)(y_N - 1)}$$

is a strongly consistent estimator of the parameters $\theta = (H_1, H_2, a^2, b^2)$.

In the special case where $b = 0$, the model reduces to a single IfBm component. In this simpler setting, we can construct up to five distinct strongly consistent estimators for the parameters (H, a^2) by considering ratios between the statistics $\xi_N^{(j)}$ at different scales.

Corollary 1 *Let* $0 < H < 1$ *the statistics* $\hat{\theta}_j = (\hat{H}_N^{(j)}, (\hat{a}_N^{(j)})^2)$ *with* $j = 1, 2, 4$ *where*

$$(10)\quad \hat{x}_N^{(j)} = \frac{\xi_N^{(2j)}}{\xi_N^{(j)}}, \qquad \hat{H}_N^{(j)} = \frac{1}{2}\log_{2+}\left(\hat{x}_N^{(j)}\right), \qquad (\hat{a}_N^{(j)})^2 = \frac{\xi_N^{(j)}\left(\hat{H}_N^{(j)}+1\right)\left(2\hat{H}_N^{(j)}+1\right)}{2h^{2\hat{H}_N^{j}}(\hat{x}_N^{(j)})^{\log_2 j}\left(\hat{x}_N^{(j)}-1\right)}$$

and the statistics $\tilde{\theta}_i = (\tilde{H}_N^{(i)}, (\tilde{a}_N^{(i)})^2)$ *with* $i = 1, 2$ *where*

$$(11)\quad \tilde{x}_N^{(i)} = \sqrt{\frac{\xi_N^{(4i)}}{\xi_N^{(i)}}}, \qquad \tilde{H}_N^{(i)} = \frac{1}{2}\log_{2+}\left(\hat{x}_N^{(i)}\right), \qquad (\tilde{a}_N^{(i)})^2 = \frac{\xi_N^{(i)}\left(\hat{H}_N^{(i)}+1\right)\left(2\hat{H}_N^{(i)}+1\right)}{2h^{2\hat{H}_N^{i}}(\hat{x}_N^{(i)})^{\log_2 i}\left(\hat{x}_N^{(i)}-1\right)}$$

are strongly consistent estimators for $\theta = (H, a^2)$.

6 Simulation Results

We simulate 1000 paths of fBm with various values of $H \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and discretizations up to 2^{20} points. Estimators are computed using increments of varying lags.

The estimators exhibit:

- Small bias for H not near boundaries (i.e., ≈ 0.1 or ≈ 0.9)
- Decreasing variance with larger sample sizes
- Convergence to normal distribution under central limit conditions.

7 Conclusion

We propose a modeling framework based on integrated fractional Brownian motion (IfBm) to account for the averaging nature of observed electricity price data. Our statistical methodology leverages ergodic properties of the process to construct consistent estimators for model parameters. Simulations confirm the validity and reliability of our estimators, which opens avenues for empirical applications to real electricity market data and more general integrated stochastic models.

References

[1] J. Beran, "Statistics for Long-Memory Processes". Chapman and Hall/CRC, 1994.

[2] F. Biagini, Y. Hu, B. Øksendal, T. Zhang, "Stochastic Calculus for Fractional Brownian Motion and Applications". Springer, 2008.

[3] L. Giordano, D. Morale, *A fractional Brownian-Hawkes model for the Italian electricity spot market: estimation and forecasting.* Journal of Energy Markets 14, n. 3 (2021), 65–109.

[4] K. Kubilius, Y. Mishura, and K. Ralchenko, "Parameter Estimation in Fractional Diffusion Models". Springer, 2017.

[5] B. Mandelbrot and J.W. Van Ness, *Fractional Brownian motions, fractional noises and applications.* SIAM Review, vol. 10, no. 4 (1968), 422–437.

[6] Y. Mishura, "Stochastic Calculus for Fractional Brownian Motion and Related Processes". Springer, 2008.

[7] R. Weron, *Energy price risk management.* Physica A: Statistical Mechanics and its Applications, vol. 285, no. 1–2 (2000), 127–134.

Parallel parking 101

Marco Di Marco (*)

Une géométrie ne peut pas être plus vraie qu'une autre; elle peut seulement être plus commode.

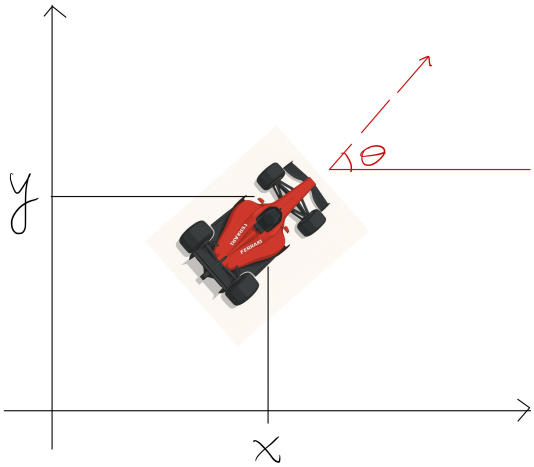
Henri Poincaré, *La Science et l'Hypothèse*

1 Introduction

Do you know that when you parallel park you are actually invoking the Chow-Rashevskii Theorem [1, 12], a cornestone of sub-Riemannian geometry? Let us be more "mathematical" and formalize the state and the possible movements of your car.

You can describe the state of your car (see the picture below) by a point $(x, y, \theta) \in \mathbb{R}^2 \times (-\Theta, \Theta)$ where $\Theta < \frac{\pi}{2}$ is the maximum angle you can rotate your wheels and:

- (x, y) is the position of your car on the plane (with respect to a fixed origin),
- θ is the angle of your wheels (with respect to a fixed axis).



(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: dimarco@math.unipd.it . Seminar held on 5 June 2025.

Now let us formalize the movement of your car.

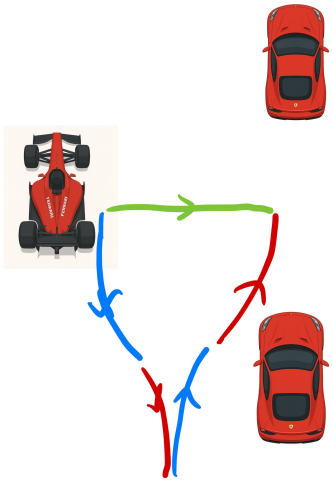
(i) You can go straight (or backwards) in the "direction" of your wheels; your car will "move" along the vector field

$$X = \cos\theta \frac{\partial}{\partial x} + \sin\theta \frac{\partial}{\partial y}$$

(ii) You can rotate your steering wheel (and therefore your wheels); your car will "move" along the vector field:

$$Y = \frac{\partial}{\partial \theta}$$

When you want to parallel park you will probably do some procedure like the one in the picture below. The movements drawn below have their velocity contained in $\mathrm{span}(X, Y)$ therefore they are "admissible"; on the other hand moving orthogonally to the road in a straight line would imply going along a curve whose velocity is *not* contained in $\mathrm{span}(X, Y)$ and therefore is not "admissible".



The secret to moving perpendicularly to the road, as said before, lies in Chow-Rashevskii Theorem: if for every point we have that

$$\dim(\mathrm{span}(X, Y, [X, Y]))) = 3$$

then between every two points there exists a curve whose velocity belongs to $\mathrm{span}(X, Y)$. In our case this is verified since

$$\begin{cases} X = \cos\theta \frac{\partial}{\partial x} + \sin\theta \frac{\partial}{\partial y} \\ Y = \frac{\partial}{\partial \theta} \\ [X, Y] = XY - YX = \sin\theta \frac{\partial}{\partial x} - \cos\theta \frac{\partial}{\partial y} \end{cases}$$

For a more detailed analyis of "parallel parking" see [11]. The example above can be considered a toy model of sub-Riemannian geometry: what you can get from it is the following.

It may be useful to think of a sub-Riemannian manifold as a Riemannian manifold, but with fewer directions in which you can move, so you have to take some detours. In other words, in the sub-Riemannian world, you can go wherever you want, but not however you want.

The aim of this note is to present, in the setting of Heisenberg groups, sub-Riemannian analogues of certain classical Euclidean results. In Section 2 we introduce Heisenberg groups and the sub-Riemannian counterparts of some classical notions. In Section 3 we summarize some of the results contained in [4] and in Section 4 we present a weaker version of some of the results contained in [3].

2 A few words on Heisenberg groups

Among sub-Riemannian manifolds, probably the most studied are the Heisenberg groups. Let us recall their definition.

Definition 2.1 Given $n \geq 1$, the *Heisenberg group* is the connected, simply connected, step 2 nilpotent Lie group associated with the algebra

$$X_1, \dots, X_n, Y_1, \dots, Y_n, T, \qquad [X_j, Y_j] = T.$$

In exponential coordinates $\mathbb{H}^n = (\mathbb{R}^{2n+1}, \cdot) = \mathbb{R}^n_x \times \mathbb{R}^n_y \times \mathbb{R}_t$ and

$$X_j = \frac{\partial}{\partial x_j} - \frac{y_j}{2}\frac{\partial}{\partial t}, \qquad Y_j = \frac{\partial}{\partial y_j} + \frac{x_j}{2}\frac{\partial}{\partial t}, \qquad T = \frac{\partial}{\partial t}.$$

The *dilations* $\delta_\lambda(x, y, t) = (\lambda x, \lambda y, \lambda^2 t)$ define a one-parameter family of group isomorphisms.
Our admissible directions will be the ones given by $\operatorname{span}(X_1, \dots, X_n, Y_1, \dots, Y_n)$, which are the *horizontal vector fields*; T is the *vertical direction*.

As said before, one question we aim to answer in this note is the following: what are the sub-Riemannian counterparts of these notions in Heisenberg groups?

- Distance,
- C^1-functions,
- C^1-submanifolds (with and without boundary),
- differential forms,
- BV/SBV functions.

The "canonical distance" in sub-Riemannian manifold is the Carnot-Carathéodory distance, that we recall below.

Definition 2.2 The *Carnot-Carathéodory* distance d is defined for $p, q \in \mathbb{H}^n$ as

$$d(p,q) := \inf \left\{ \|h\|_{L^1([0,1],\mathbb{R}^{2n})} : \begin{array}{l} \text{the curve } \gamma_h : [0,1] \to \mathbb{H}^n \text{ defined by} \\ \gamma_h(0) = p,\ \dot{\gamma}_h = \sum_{j=1}^n (h_j X_j + h_{j+n} Y_j)(\gamma_h) \\ \text{has final point } \gamma_h(1) = q \end{array} \right\}.$$

For $m \geq 1$ we canonically define the measures $\mathcal{H}^m, \mathcal{S}^m$.

Remark 2.3 In general sub-Riemannian manifolds open balls defined via the Carnot-Carathéodory distance may have strange shapes; however under mild regularity assumption it is possible to show, via a simple calibration argument, that $\operatorname{diam}(B(p,r)) = 2r$ for small radii. See [6].

Definition 2.4 Let $\Omega \subseteq \mathbb{H}^n$ be an open set and $f \colon \Omega \to \mathbb{R}$. We say that $f \in C^1_{\mathbb{H}}(\Omega)$ if f is continuous and its *horizontal gradient* $\nabla_{\mathbb{H}} f := (X_1 f, \dots, Y_n f)$, in the sense of distributions, is represented by a continuous function.

The following definition about intrinsic C^1 (or $C^1_{\mathbb{H}}$) submanifold was given in [9].

Definition 2.5 Let $1 \leq k \leq 2n+1$ and $S \subseteq \mathbb{H}^n$. We say that S is a $C^1_{\mathbb{H}}$ submanifold of dimension k if

- $k \leq n$: S is a C^1 submanifold and $TS \subseteq \operatorname{span}(X_i, Y_i)_{1 \leq i \leq n}$.
- $k \geq n+1$: locally $S = \{p : f(p) = 0\}$ for $f : \mathbb{H}^n \to \mathbb{R}^{2n+1-k}$ s.t. $f \in C^1_{\mathbb{H}}$ and $\nabla_{\mathbb{H}} f$ has maximal rank.

Remark 2.6 If $k \geq n+1$, S can be a fractal (see [10]), but it retains good intrinsic properties, e.g., its blow up is the vertical k-plane $\langle \nabla_{\mathbb{H}} f_1, \dots, \nabla_{\mathbb{H}} f_{2n+1-k} \rangle^{\perp}$. One can define the *tangent k-vector* $t^{\mathbb{H}}_S$ as the unit multivector associated with $\langle \nabla_{\mathbb{H}} f_1, \dots, \nabla_{\mathbb{H}} f_{2n+1-k} \rangle^{\perp}$.

High dimensional $C^1_{\mathbb{H}}$-submanifold can be also described by *intrinsic* graph. In other words, the following implicit function theorem (proved in [9]) holds.

Theorem 2.7 *Given $k \geq n+1$, a $C^1_{\mathbb{H}}$ k-dimensional submanifold $\{f = 0\}$ can be written as a continuous intrinsic graph, i.e., locally there exists a continuous function $\phi : \mathbb{W} \to \mathbb{V}$ acting between complementary subgroups of $\mathbb{H}^n$ such that $\dim(\mathbb{V}) = k$ (therefore $\mathbb{V}$ is abelian and horizontal) and*

$$\{f(x) = 0\} = \{w \cdot \phi(w)\}$$

Remark 2.8 For intrinsic graph one can define appropriate notions of intrinsic Lipschitz and intrinsic differentiability condition which are far more *geometric* than their Euclidean counterpart. Rademacher (see [15]) and Stepanov (see [5]) theorems holds (in their "sub-Riemannian geometric translation").

Definition 2.9 M. Rumin (see [13, 14]) introduced in the 90s a complex of differential forms in $\mathbb{H}^n$:

$$0 \xrightarrow{d} \mathcal{D}^0_{\mathbb{H}} \xrightarrow{d} \mathcal{D}^1_{\mathbb{H}} \dots \xrightarrow{d} \mathcal{D}^n_{\mathbb{H}} \xrightarrow{D} \mathcal{D}^{n+1}_{\mathbb{H}} \xrightarrow{d} \dots \xrightarrow{d} \mathcal{D}^{2n+1}_{\mathbb{H}} \xrightarrow{d} 0$$

where

- the forms $\mathcal{D}^0_{\mathbb{H}}, \dots, \mathcal{D}^n_{\mathbb{H}}$ are *different* from the forms $\mathcal{D}^{n+1}_{\mathbb{H}}, \dots, \mathcal{D}^{2n+1}_{\mathbb{H}}$,
- d is the exterior derivative but D is a *non trivial* second order operator.

Rumin complex is highly not trivial so in this note, for the sake of simplicity, we will just assume, *cum grano salis*, that

- $\mathcal{D}^k_{\mathbb{H}} \subseteq \{\text{smooth } k\text{-forms in } \mathbb{H}^n\}$,
- the differential operator D is something like

$$D\omega = d(\omega + \text{“vertical stuff}(d\omega)\text{”})$$

  where by *vertical stuff* we intend something that, when integrated (see Definition 2.10 below) on a submanifold with its tangent space contained in the horizontal distribution, it vanishes.

Definition 2.10 We can integrate a Rumin k-form ω on an oriented k-dimensional $C^1_{\mathbb{H}}$ submanifold S in the following way

- If $k \leq n$ we define $\displaystyle\int_S \omega$ as the classical one.
- If $k \geq n+1$ and S is also a C^1 submanifold, one has (see [15]) that

$$\int_S \omega = C \int_S \langle \omega | t^{\mathbb{H}}_S \rangle d\mathcal{S}^{k+1}$$

  for some explicit constant $C = C(n, k, d)$. Hence we define

$$\int_S \omega \coloneqq C \int_S \langle \omega | t^{\mathbb{H}}_S \rangle d\mathcal{S}^{k+1}.$$

3 Stokes' Theorem in Heisenberg groups

In this section we present a sketch of the proof of some of the results contained in [4]. First we present our definition of $C^1_{\mathbb{H}}$-regular submanifold with boundary.

Definition 3.1 Given a $C^1_{\mathbb{H}}$ submanifold S, we define its *boundary* as $\overline{S} \setminus S$.
We say that $S \subseteq \mathbb{H}^n$ is a k-dimensional $C^1_{\mathbb{H}}$ submanifold with boundary if

(i) S is a k-dimensional $C^1_{\mathbb{H}}$ submanifold,

(ii) ∂S is a $(k-1)$-dimensional $C^1_{\mathbb{H}}$ submanifold,

(iii) for every $p \in \partial S$ there exist a neighbourhood $U \ni p$ and a k-dimensional $C^1_{\mathbb{H}}$ submanifold S' such that

$$U \cap \overline{S} \subseteq S' \text{ and } S' \cap \overline{S} \cap B(p,r) \neq \emptyset \text{ for every } r > 0.$$

When S is oriented, an orientation is naturally induced on ∂S.

The main result of this section is the following.

Theorem 3.2 *Let $S \subset \mathbb{H}^n$ be a k-dimensional orientable $C^1_{\mathbb{H}}$ submanifold with boundary and $\omega \in \mathcal{D}^{k-1}_{\mathbb{H}}$. Then*

$$\int_S d_c\omega = \int_{\partial S} \omega,$$

where $d_c = d$ if $k \neq n+1$ and $d_c = D$ if $k = n+1$.

Let us make the following remarks.

Remark 3.3

(i) Theorem 3.2 is just the classical Stokes Theorem if $k \leq n$.

(ii) Theorem 3.2 is just the classical Stokes Theorem if $k \geq n+2$ and S is also a C^1 submanifold with boundary.

(iii) Theorem 3.2 is *almost* just the classical Stokes Theorem if $k = n+1$ and S is also a C^1 submanifold with boundary:

$$\int_S D\omega = \int_S d(\omega + \text{vertical stuff}) = \int_{\partial S} \omega + \text{vertical stuff} = \int_{\partial S} \omega.$$

Sketch of the proof of Theorem 3.2 when $k \geq n+2$. The proof works by approximating S by a sequence $(S_j)_{j\in\mathbb{N}}$ of C^1 submanifolds with boundary, so that

$$\int_S d_c\omega = \lim_{j\to+\infty} \int_{S_j} d_c\omega = \lim_{j\to+\infty} \int_{\partial S_j} \omega = \int_{\partial S} \omega.$$

This can be done (locally, which is enough) exploiting the following result:

Theorem 3.4 *Assume $k \geq n+2$ and consider a k-dimensional $C^1_{\mathbb{H}}$ submanifold with boundary $S \subseteq \mathbb{H}^n$, let $h := 2n+1-k$ be the codimension of S. Then, for every $p \in \partial S$ there exist a neighbourhood $U \ni p$ and functions $f_1, \dots, f_{h+1} \in C^1_{\mathbb{H}}(U)$ such that*

(i) $\nabla_{\mathbb{H}} f_1, \dots, \nabla_{\mathbb{H}} f_{h+1}$ *are linearly independent in U,*

(ii) $U \cap S = \{q \in U : f_1(q) = \cdots = f_h(q) = 0, f_{h+1}(q) > 0\}$,

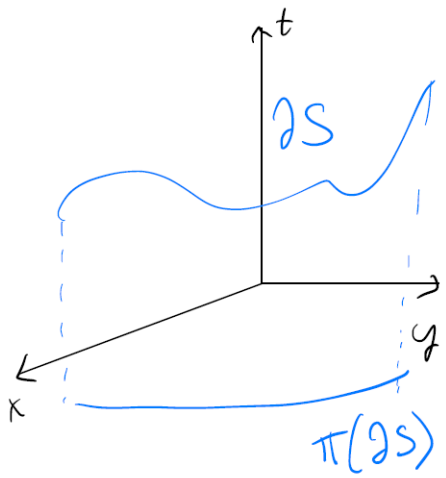
(iii) $U \cap \partial S = \{q \in U : f_1(q) = \cdots = f_h(q) = f_{h+1}(q) = 0\}$.

Moreover, the functions $f_h, \dots, f_{h+1}$ *can be chosen to be of class* C^∞ *on* $U \setminus \{q \in U : f_1(q) = \cdots = f_h(q) = 0\}$.

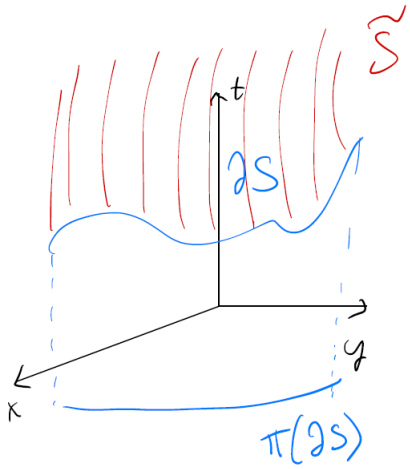
□

Sketch of the proof of Theorem 3.2 when $k = n + 1$. The idea is to produce a sequence of k-dimensional C^1 submanifolds with boundary such that $\partial S_j = \partial S$ (recall that ∂S is an n-dimensional horizontal C^1 submanifold).

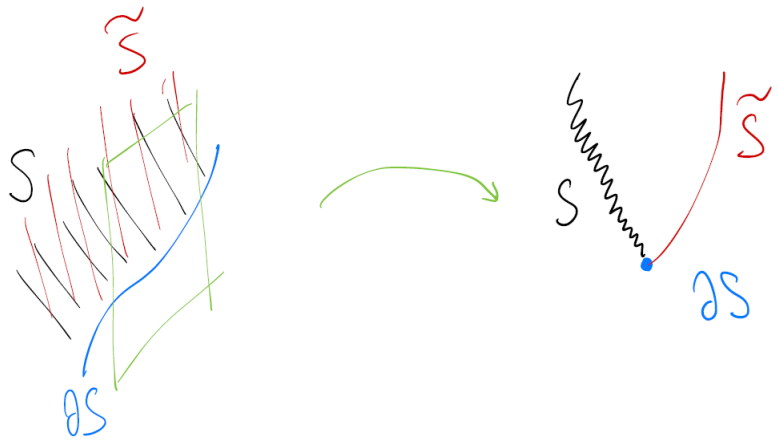
- We want to produce a C^1 and $C^1_{\mathbb{H}}$ submanifold with boundary $\tilde{S}$ such that $\partial \tilde{S} = \partial S$: to do this we project ∂S on the horizontal distribution.



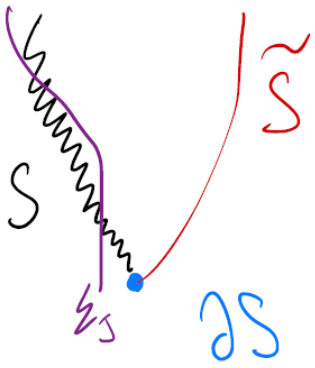
- Then we produce $\tilde{S}$ by producing one “half" of a vertical “cylinder".



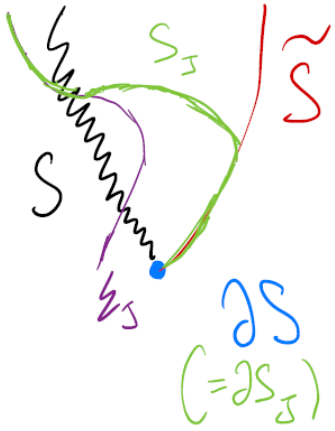
For the rest of this sketch we will look at a section through a vertical plan, in order to avoid “graphical" confusion.



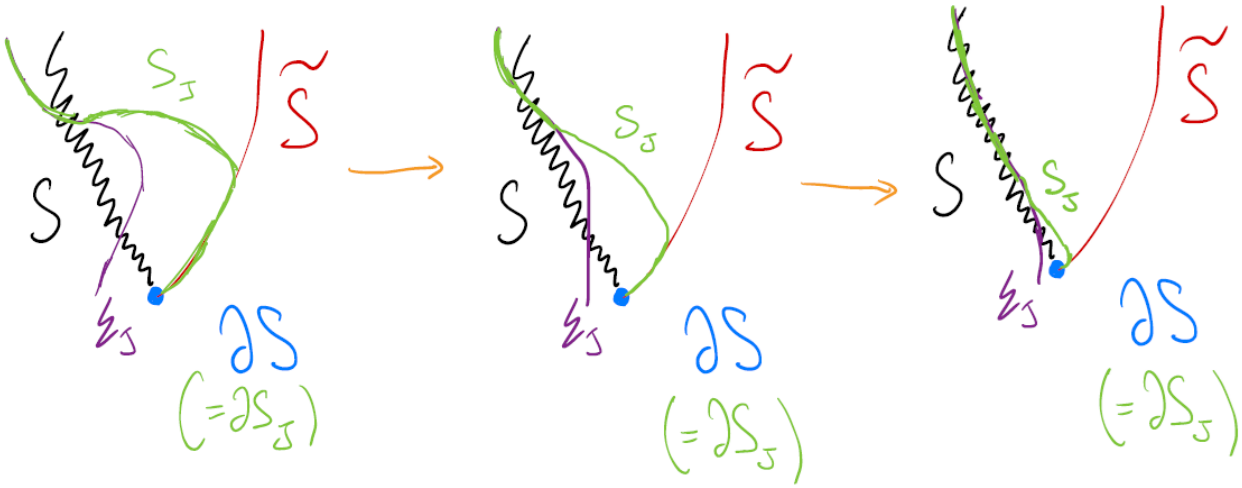
- Now we produce smooth approximations (as we did before when $k \geq n+2$) (in $C^1_{\mathbb{H}}$) Σ_j of S, without any concern for their boundaries.



- Finally we produce S_j by interpolating between $\tilde{S}$ (in a neighbourhood of ∂S) and Σ_j away from ∂S.



Below there is a graphic visualization of what happens when $j \to +\infty$.



□

4 $\mathrm{SBV}_{\mathbb{H}}$ functions in Heisenberg groups

In this section we present a simplified version of some of the results contained in [3]. First we recall some definitions about functions of bounded variation in Heisenberg groups ($\mathrm{BV}_{\mathbb{H}}$ functions).

Definition 4.1 Fix a bounded open set $\Omega \subseteq \mathbb{H}^n$. We say that $u \in L^1(\Omega)$ is a *function of bounded $\mathbb{H}$-variation*, and we write $u \in \mathrm{BV}_{\mathbb{H}}(\Omega)$, if there exists a $\mathbb{R}^{2n}$-valued Radon measure $D_{\mathbb{H}}u = (D_{X_1}u, \dots D_{Y_n}u)$ on Ω with finite total variation such that, for every open set $A \subset\subset \Omega$, for every $1 \leq i \leq n$ and for every $\varphi \in C_c^1(A)$ one has

$$\int_A \varphi d(D_{X_i}u) = -\int_A u X_i^* \varphi d\mathcal{L}^{2n+1}, \qquad \int_A \varphi d(D_{Y_i}u) = -\int_A u Y_i^* \varphi d\mathcal{L}^{2n+1}.$$

where $\cdot^*$ denotes the formal adjoint of $\cdot$.
For every $u \in \mathrm{BV}_{\mathbb{H}}(\Omega)$ we define the norm

$$\|u\|_{\mathrm{BV}_{\mathbb{H}}(\Omega)} \coloneqq \|u\|_{L^1(\Omega)} + |D_{\mathbb{H}}u|(\Omega).$$

The space $\mathrm{BV}_{\mathbb{H}}(\Omega)$ equipped with the above norm is a Banach space.

Definition 4.2 Let $S \subseteq \mathbb{H}^n$. We say that S is *countably $\mathbb{H}$-rectifiable* if there exists a family $\{S_h : h \in \mathbb{N}\}$ of $C^1_{\mathbb{H}}$-hypersurfaces (aka $2n$-dimensional/1-codimensional $C^1_{\mathbb{H}}$-submanifolds) such that

$$\mathcal{H}^{2n+1}\left(S \setminus \bigcup_{h\in\mathbb{N}} S_h\right) = 0.$$

Moreover, if $\mathcal{H}^{2n+1}(S) < +\infty$, we say that S is $\mathbb{H}$-*rectifiable*.
We can define $\mathcal{H}^{2n+1}$-a.e., up to a sign, the *horizontal normal* $\nu_S(p)$ of a countably $\mathbb{H}$-rectifiable set S at $p \in S$.

Definition 4.3 Let $u \in L^1_{\mathrm{loc}}(\Omega)$, $z \in \mathbb{R}$ and $p \in \Omega$. We say that z is the *approximate limit* of u at p if

$$\lim_{r\to 0} \fint_{B(p,r)} |u - z| d\mathcal{L}^{2n+1} = 0.$$

If the approximate limit of u at p exists, it is also unique and we denote it by $u^\star(p)$.
Let p be such that $u^\star(p)$ exists. We say that u is *approximately* $\mathbb{H}$-*differentiable* at p if there exist a neighbourhood $U \subset \Omega$ of p and $f \in C^1_{\mathbb{H}}(U)$ such that $f(p) = 0$ and

$$\lim_{r\to 0} \fint_{B(p,r)} \frac{|u - u^\star(p) - f|}{r} d\mathcal{L}^{2n+1} = 0.$$

The vector $\nabla_{\mathbb{H}} f(p) \in \mathbb{R}^m$ is uniquely determined and we call it *approximate* $\mathbb{H}$-*gradient* of u at p and we denote it by $D^{\mathrm{ap}}_{\mathbb{H}} u(p)$.

Definition 4.4 Fix $p \in \mathbb{H}^n$, $R > 0$ and $\nu \in \mathbb{S}^{2n-1}$. Let $f \in C^1_{\mathbb{H}}(B(p,R))$ be such that $f(p) = 0$ and $\frac{\nabla_{\mathbb{H}} f(p)}{|\nabla_{\mathbb{H}} f(p)|} = \nu$. For every $r \in (0, R)$ we set

$$B^+_\nu(p,r) \coloneqq B(p,r) \cap \{f > 0\},$$
$$B^-_\nu(p,r) \coloneqq B(p,r) \cap \{f < 0\}.$$

Let $u \in L^1_{loc}(\Omega)$ and $p \in \Omega$. We say that u has an *approximate* $\mathbb{H}$-*jump* at p if there exist $u^+, u^- \in \mathbb{R}$ with $u^+ \neq u^-$ and $\nu \in \mathbb{S}^{2n-1}$ such that

$$\lim_{r\to 0} \fint_{B^+_\nu(p,r)} |u - u^+| d\mathcal{L}^{2n+1} = \lim_{r\to 0} \fint_{B^-_\nu(p,r)} |u - u^-| d\mathcal{L}^{2n+1} = 0. \tag{1}$$

The *jump set* $\mathcal{J}_u$ is defined as the set of points where u has an approximate $\mathbb{H}$-jump.

Definition 4.5 For every $u \in \mathrm{BV}_{\mathbb{H}}(\Omega)$ we decompose

$$D_{\mathbb{H}} u = D^a_{\mathbb{H}} u + D^s_{\mathbb{H}} u$$

where $D^a_{\mathbb{H}} u$ denotes the *absolutely continuous part* of $D_{\mathbb{H}} u$ (with respect to the usual Lebesgue measure $\mathcal{L}^{2n+1}$) and $D^s_{\mathbb{H}} u$ denotes the *singular part* of $D_{\mathbb{H}} u$. We define the *jump part* of $D_{\mathbb{H}} u$ as

$$D^j_{\mathbb{H}} u \coloneqq D^s_{\mathbb{H}} u \llcorner \mathcal{J}_u$$

and the *Cantor part* of $D_{\mathbb{H}} u$ as

$$D^c_{\mathbb{H}} u \coloneqq D^s_{\mathbb{H}} u \llcorner (\Omega \setminus \mathcal{J}_u).$$

We are now ready to give the definition of special functions with bounded variation in Heisenberg groups ($\mathrm{SBV}_{\mathbb{H}}$ functions).

Definition 4.6 Let $u \in \mathrm{BV}_{\mathbb{H}}(\Omega)$. We say that u is a *special function of bounded $\mathbb{H}$-variation*, and we write $u \in \mathrm{SBV}_{\mathbb{H}}(\Omega)$, if $D^c_{\mathbb{H}} u = 0$.

The following result about the representation of the distributional derivative of an $\mathrm{SBV}_{\mathbb{H}}$ function is an immediate consequence of some results from [7].

Theorem 4.7 *For every $u \in \mathrm{SBV}_{\mathbb{H}}(\Omega)$ the jump set $\mathcal{J}_u$ is $\mathbb{H}$-rectifiable and we can write*

$$D_{\mathbb{H}} u = D^{\mathrm{ap}}_{\mathbb{H}} u \mathcal{L}^{2n+1} + \sigma(\cdot, \nu_{\mathcal{J}_u})(u^+ - u^-)\nu_{\mathcal{J}_u} \mathcal{S}^{2n+1} \llcorner \mathcal{J}_u$$

for some function $\sigma : \mathbb{H}^n \times \mathbb{S}^{2n-1} \to (0, +\infty)$.

We are now ready to state the main result of this section which draws inspiration from its Euclidean counterpart, proved in [2].

Theorem 4.8 *Let Ω be a bounded open subset of $\mathbb{H}^n$ and let $u \in \mathrm{SBV}_{\mathbb{H}}(\Omega)$. Then, there exists a sequence of functions $(u_k)_{k\in\mathbb{N}} \subset \mathrm{SBV}_{\mathbb{H}}(\Omega)$ and of $C^1_{\mathbb{H}}$-hypersurfaces $(M_k)_{k\in\mathbb{N}} \subset \Omega$ such that, for every $k \in \mathbb{N}$, $\mathcal{J}_{u_k} \subseteq M_k \cap \mathcal{J}_u$, $\mathcal{J}_{u_k}$ is compact, and*

$$\|u - u_k\|_{\mathrm{BV}_{\mathbb{H}}(\Omega)} \xrightarrow{k\to+\infty} 0, \qquad u_k \in C^\infty(\Omega \setminus \mathcal{J}_{u_k}).$$

Sketch of the proof. First we need to "approximate" the jump set $\mathcal{J}_u$.
We start with a countably $\mathbb{H}$-rectifiable set $\mathcal{J}_u$.
Then for every $k \in \mathbb{N}$ we consider the set

$$\mathcal{J}^k_u \coloneqq \left\{ x \in \mathcal{J}_u : |u^+(x) - u^-(x)| \geq \frac{1}{k} \right\} \cap B(0, k).$$

These sets are $\mathbb{H}$-rectifiable and $|D_{\mathbb{H}} u|(\mathcal{J}_u \setminus \mathcal{J}^k_u) \xrightarrow{k\to+\infty} 0$.
Then we can find a $C^1_{\mathbb{H}}$-hypersurface M_k such that $|D_{\mathbb{H}} u|(M_k \setminus \mathcal{J}^k_u) < \frac{1}{k}$.
Finally we can find (by Lusin Theorem) a compact set $C_k \subseteq M_k \cap \mathcal{J}_u$ such that

$$|D_{\mathbb{H}} u|((\mathcal{J}^k_u \cap M_k) \setminus C_k) < \frac{1}{k}$$

C_k will be the jump set $\mathcal{J}_{u_k}$ of the approximating function u_k.
Now we construct the approximating functions $(u_k)_{k\in\mathbb{N}}$. For $\ell \in \mathbb{N}$ we define the bounded open sets

$$A^1_k \coloneqq \left\{ x \in \Omega : d_E(x, C_k) > \frac{1}{2} \right\},$$
$$A^\ell_k \coloneqq \left\{ x \in \Omega : \frac{1}{\ell + 1} < d_E(x, C_k) < \frac{1}{\ell - 1} \right\} \text{ if } \ell > 1.$$

where d_E denotes the classical Euclidean distance. We have

$$\bigcup_{\ell\in\mathbb{N}} A_k^\ell = \Omega \setminus C_k.$$

Consider a partition of unity on $\Omega \setminus C_k$ associated with $(A_k^\ell)_{\ell\in\mathbb{N}}$,

$$\xi_k^\ell \in C_c^\infty(A_k^\ell) \text{ such that } 0 \le \xi_k^\ell \le 1 \text{ and } \sum_{\ell\in\mathbb{N}} \xi_k^\ell \equiv 1 \text{ on } \Omega \setminus C_k.$$

Fix a mollification kernel, i.e., a spherically symmetric non-negative function $K \in C_c^\infty(B_E(0,1))$ such that $\int_{\mathbb{R}^n} K d\mathcal{L}^{2n+1} = 1$. For $\varepsilon > 0$ we define $K_\varepsilon(x) \coloneqq \varepsilon^{-2n-1} K(x/\varepsilon)$.
Finally for some $\varepsilon_k^\ell > 0$ choosen appropriately we define

$$u_k \coloneqq \sum_{\ell\in\mathbb{N}} (\xi_k^\ell u) * K_{\varepsilon_k^\ell} \quad \text{on} \quad \Omega \setminus C_k,$$

where $*$ denotes the classical Euclidean convolution. By the choice we made on $(A_k^\ell)_{\ell\in\mathbb{N}}$ the sum is locally finite, hence $u_k \in C_c^\infty(\Omega \setminus C_k)$. By choosing the ε_k^ℓ small enough we clearly have (since $\mathcal{L}^{2n+1}(C_k) = 0$) that

$$\|u - u_k\|_{L^1(\Omega)} \xrightarrow{k\to+\infty} 0.$$

Moreover, one can prove that $u_k \in \mathrm{SBV}_{\mathbb{H}}(\Omega)$ and

$$u_k^- \equiv u^- \text{ on } C_k = \mathcal{J}_{u_k} \quad \text{and} \quad u_k^+ \equiv u^+ \text{ on } C_k = \mathcal{J}_{u_k}. \tag{2}$$

We are left to prove that $\|u - u_k\|_{\mathrm{BV}_{\mathbb{H}}(\Omega)} \xrightarrow{k\to+\infty} 0$. By definition,

$$\|u - u_k\|_{\mathrm{BV}_{\mathbb{H}}(\Omega)} = \underbrace{\|u - u_k\|_{L^1(\Omega)}}_{(A)} + |D_{\mathbb{H}}(u - u_k)|(\Omega).$$

The term (A), as a result of what has been said above, can be made smaller than $1/k$ as long as k is sufficiently big. We are left to estimate $|D_{\mathbb{H}}(u - u_k)|(\Omega)$. By the construction we made for C_k we have

$$\begin{aligned}|D_{\mathbb{H}}(u - u_k)|(\Omega) \le & |D_{\mathbb{H}}(u - u_k)|(\Omega \setminus \mathcal{J}_u) + \underbrace{|D_{\mathbb{H}}(u - u_k)|(\mathcal{J}_u \setminus \mathcal{J}_u^k)}_{(B)} + \underbrace{|D_{\mathbb{H}}(u - u_k)|(\mathcal{J}_u^k \setminus M_k)}_{(C)} \\ & + \underbrace{|D_{\mathbb{H}}(u - u_k)|((\mathcal{J}_u \cap M_k) \setminus C_k)}_{(D)} + |D_{\mathbb{H}}(u - u_k)|(C_k).\end{aligned}$$

The terms $(B), (C)$ and (D) also can be made smaller than $1/k$ as long as k is sufficiently big. Moreover, thanks to Theorem 4.7 and (2), we can infer that

$$|D_{\mathbb{H}}(u - u_k)|(C_k) = 0.$$

We are left to estimate $|D_{\mathbb{H}}(u-u_k)|(\Omega\setminus\mathcal{J}_u)$. On $\Omega\setminus\mathcal{J}_u$ we have that $D_{\mathbb{H}}(u-u_k)$ is absolutely continuous with respect to $\mathcal{L}^{2n+1}$ so that

$$|D_{\mathbb{H}}(u-u_k)|(\Omega\setminus\mathcal{J}_u)=\left\|D^{\mathrm{ap}}_{\mathbb{H}}u-\nabla_{\mathbb{H}}u_k\right\|_{L^1(\Omega)}.$$

On $\Omega\setminus\mathcal{J}_u$ we have

$$\nabla_{\mathbb{H}}u_k=\sum_{\ell\in\mathbb{N}}\left((\xi^\ell_k D_{\mathbb{H}}u)*K_{\varepsilon^\ell_k}+R^\ell_k\right)$$

where R^ℓ_k is a *reminder* coming from the *non-commutativity* of Heisenberg groups. Then we have

$$\left\|D^{\mathrm{ap}}_{\mathbb{H}}u-\nabla_{\mathbb{H}}u_k\right\|_{L^1}\le\left\|D^{\mathrm{ap}}_{\mathbb{H}}u-\sum_{\ell\in\mathbb{N}}[(\xi^\ell_k D_{\mathbb{H}}u)*K_{\varepsilon^\ell_k}]\right\|_{L^1}+\underbrace{\left\|\sum_{\ell\in\mathbb{N}}R^\ell_k\right\|_{L^1}}_{(E)}$$

Mimicking some results from [8], and possibly reducing ε^ℓ_k, the term (E) can be made smaller than $1/k$ as long as k is sufficiently big. We are left with

$$\begin{aligned}\left\|\sum_{\ell\in\mathbb{N}}[(\xi^\ell_k D_{\mathbb{H}}u)*K_{\varepsilon^\ell_k}]-D^{\mathrm{ap}}_{\mathbb{H}}u\right\|_{L^1(\Omega)}\le&\underbrace{\left\|\sum_{\ell\in\mathbb{N}}[(\xi^\ell_k D^{\mathrm{ap}}_{\mathbb{H}}u)*K_{\varepsilon^\ell_k}]-\sum_{\ell\in\mathbb{N}}\xi^\ell_k D^{\mathrm{ap}}_{\mathbb{H}}u\right\|_{L^1(\Omega)}}_{(F)}\\&+\left\|\sum_{\ell\in\mathbb{N}}(\xi^\ell_k D^j_{\mathbb{H}}u)*K_{\varepsilon^\ell_k}\right\|_{L^1(\Omega)}.\end{aligned}$$

Up to reducing ε^ℓ_k, we can assume that $(F)\le 1/k$ as long as k is big enough. Finally we have

$$\begin{aligned}\left\|\sum_{\ell\in\mathbb{N}}(\xi^\ell_k D^j_{\mathbb{H}}u)*K_{\varepsilon^\ell_k}\right\|_{L^1(\Omega)}&\le\sum_{\ell\in\mathbb{N}}\left\|(\xi^\ell_k D^j_{\mathbb{H}}u)*K_{\varepsilon^\ell_k}\right\|_{L^1(A^\ell_k)}\le\sum_{\ell\in\mathbb{N}}|D^j_{\mathbb{H}}u|(A^\ell_k+\varepsilon^\ell_k)\\&\le\sum_{\ell\in\mathbb{N}}|D^j_{\mathbb{H}}u|(A^{\ell-1}_k\cup A^\ell_k\cup A^{\ell+1}_k)\le\sum_{\ell\in\mathbb{N}}3|D^j_{\mathbb{H}}u|(A^\ell_k)\le 9|D^j_{\mathbb{H}}u|(\Omega\setminus C_k)\end{aligned}$$

but, thanks to the construction of C_k, the last term can be made smaller than $1/k$ as long as k is sufficiently big, concluding the proof. □

References

[1] Chow, W.-L., *Über Systeme von linearen partiellen Differentialgleichungen erster Ordnung.* Math. Ann. 117 (1939), 98–105.

[2] De Philippis, G., Fusco, N., and Pratelli, A., *On the approximation of SBV functions.* Atti Accad. Naz. Lincei Rend. Lincei Mat. Appl. 28, 2 (2017), 369–413.

[3] Di Marco, M., Don, S., and Vittone, D., *SBV functions in Carnot-Carathéodory spaces.* 2025. Preprint available online at `https://arxiv.org/abs/2503.04285`.

[4] Di Marco, M., Julia, A., Nicolussi Golo, S., and Vittone, D., *Submanifolds with boundary and Stokes' Theorem in Heisenberg groups.* Trans. Amer. Math. Soc. 378 (2025), 4955—4990.

[5] Di Marco, M., Pinamonti, A., Vittone, D., and Zambanini, K., *Stepanov differentiability theorem for intrinsic graphs in Heisenberg groups.* Adv. Calc. Var. (2025). In press, `https://doi.org/10.1515/acv-2024-0118`.

[6] Di Marco, M., Somma, G., and Vittone, D., *A note on the diameter of small sub-Riemannian balls.* 2025. Preprint available online at `https://arxiv.org/abs/2505.02790`.

[7] Don, S., and Vittone, D., *Fine properties of functions with bounded variation in Carnot-Carathéodory spaces.* J. Math. Anal. Appl. 479, 1 (2019), 482–530.

[8] Franchi, B., Serapioni, R., and Serra Cassano, F., *Meyers-Serrin type theorems and relaxation of variational integrals depending on vector fields.* Houston J. Math. 22, 4 (1996), 859–890.

[9] Franchi, B., Serapioni, R., and Serra Cassano, F., *Regular submanifolds, graphs and area formula in Heisenberg groups.* Adv. Math. 211, 1 (2007), 152–203.

[10] Kirchheim, B., and Serra Cassano, F., *Rectifiability and parameterization of intrinsic regular surfaces in the Heisenberg group.* Ann. Sc. Norm. Super. Pisa Cl. Sci. (5) 3, 4 (2004), 871–896.

[11] Nelson, E., *Tensor analysis. Preliminary informal notes of university courses and seminars in mathematics.* Mathematical Notes. Princeton, N.J.: Princeton University Press. iv, 127 p. (1967).

[12] Rashevsky, P., *Any two points of a totally nonholonomic space may be connected by an admissible line.* Uch. Zap. Ped. Inst. im. Liebknechta Sr. Phys. Math. 2 (1938), 83–94. In Russian.

[13] Rumin, M., *Un complexe de formes différentielles sur les variétés de contact.* C. R. Acad. Sci. Paris Sér. I Math. 310, 6 (1990), 401–404.

[14] Rumin, M., *Formes différentielles sur les variétés de contact.* J. Differential Geom. 39, 2 (1994), 281–330.

[15] Vittone, D., *Lipschitz graphs and currents in Heisenberg groups.* Forum Math. Sigma 10 (2022), Paper No. e6, 104.

Hamilton Jacobi Equations in the Space of Probability Measures

Giacomo Ceccherini Silberstein (*)

Abstract. In this short note, I present the main ideas discussed during the PhD seminar. Optimal control theory is a branch of Calculus of Variations aiming to guide efficiently a system to achieve a specific goal, minimizing a given "cost" along the way. In this manuscript, we'll embark on a controlled excursion into the Hamilton-Jacobi equation, a powerful partial differential equation that encodes optimality conditions for this variational problem. We will start by presenting the classical Euclidean setting, where the system's state space is finite-dimensional and familiar. Then, we'll extend these fundamental ideas to the more complex mean field setting, where the dynamics play out on the space of probability measures, allowing us to understand collective behaviors in large systems.

1 Control Problems

For the sake of clarity of the exposition all the discussion will be delivered considering the ambient space $\mathbb{T}^d$. Many details and assumptions are omitted> However all precise references for the assumptions are given.

Let's start with

Definition 1.1 (Controlled Dynamical system) *A controlled dynamical system in* $\mathbb{T}^d$ is the datum of the triple $(\mathbb{T}^d, \mathcal{A}, \mathrm{Dyn})$, where

- A is the set of admissible control values, namely

  $$\mathcal{A} = \Big\{\alpha : [0,\infty) \to A : \alpha(\cdot) \text{ is Borel measurable}\Big\}, \tag{1.1}$$

  with $A \subset \mathbb{R}^k$ the set of parameter values.

- Dyn is an ODE dynamical system parametrized by $\mathcal{A}$

  $$\begin{cases} \dot{X}_t = f(X_t, \alpha_t) \\ X_0 = x_0 \in \mathbb{T}^d \end{cases} \tag{Dyn}$$

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: ceccheri@math.unipd.it . Seminar held on 19 June 2025.

In application, X_t represents the state of an agent, that at time zero is in x_0, acting on its own position via the control α_t

1.1 Payoffs

In modeling, we want the control parameter to be chosen according to an optimization criterion. Given $L : \mathbb{T}^d \times A \to \mathbb{R}$, $g : \mathbb{T}^d \to \mathbb{R}$ and $T > 0$, the *cost or objective functional* is

$$J(x_0, \alpha) := \int_0^T L(X_t, \alpha(t))dt + g(X_t) \tag{Obj}$$

where X_t solves (Dyn) for the control α with initial condition $X_0 = x_0$.

The goal is to minimize this payoff, i.e. find $\alpha^* \in \mathcal{A}$ s.t.

$$J(x_0, \alpha^*) \leq J(x_0, \alpha). \tag{1.2}$$

Such a α^* is said to be an optimal control for (Obj). Therefore the goal is solve the following minimization problem

$$u(x_0) = \inf_{\alpha \in \mathcal{A}} J(x_0, \alpha) \tag{Value}$$

In the first part of this note we will show

- How to produce optimal control.
- How to extend this construction to a finite number of agents.

1.2 Dynamic Programming Principle

Some mathematical problems can be solved by deformation. Namely, we inject the original problem into a family of problems:

$$P \hookrightarrow \Big\{(P_t)_{t>0},\, P_0 = P\Big\}.$$

The idea is that maybe the infinitesimal problem

$$\frac{d}{dt}P_t$$

can be more easily solved, and by "integration," you can recover the starting problem. To illustrate the idea of a deformation, let us first consider a classical example from Calculus I

Example 1.1 (Feynman Integration Technique) Consider the following integral

$$\int_0^\infty \frac{\sin(x)}{x}dx \tag{I}$$

whose value is not immediately computable with standard Calculus I techniques.

Then, we deform this problem, by considering a family of integrals

$$\int_0^\infty e^{-tx}\frac{\sin(x)}{x}dx \tag{I_t}$$

Note that (I_t) is a deformation of (I) as t varies, i.e. $\mathrm{I}_0 = \mathrm{I}$. The flux $t \to \mathrm{I}_t$ satisfies

$$\begin{cases} \frac{d}{dt}\mathrm{I}_t = -\int_0^x e^{-tx}\sin x dx = -\frac{1}{1+t^2} \\ \mathrm{I}_0 = \mathrm{I}. \end{cases} \tag{1.3}$$

This ODE can be solved by a simple integration and we have $\mathrm{I}_t = -\arctan(t) + \frac{\pi}{2}$. In particular,

$$\mathrm{I} = \mathrm{I}_0 = \frac{\pi}{2}.$$

Now, we show how to interpret the dynamic programming principle and the Hamilton Jacobi equation as consequences of a deformation of our original problem (MF-VF).

In the sixties, R. Bellman [3] introduced the following deformation

$$J(t, x_0, \alpha) := \int_t^T L(X_s, \alpha_s)\, ds + g(x(T)) \tag{Obj_t}$$

where

$$\begin{cases} \dot{X}_s = f(X_s, \alpha_s), \\ X_t = x_0 \end{cases} \tag{Dyn_t}$$

Note that Obj_t is s.t. $\mathrm{Obj}_0 = \mathrm{Obj}$.

We will call

$$\mathcal{A}_{[t,T]}(x_0) := \left\{ [t,T] \ni s \mapsto X_s \text{ solution of } (\mathrm{Dyn}_t) : \alpha \in \mathcal{A} \right\} \tag{1.4}$$

the set of admissible trajectories.

The function

$$u(t,x) := \inf_{x(\cdot)\in\mathcal{A}_{[t,T]}(x_0)} J(t, x_0) \tag{Value_t}$$

is called the **value function** associated to the objective function (3.15) subjected to the controlled dynamic (Dyn_t).

Note that $u(0,x) = u(x)$ and $u(T,x) = g(x)$. We proceed by focusing on the equation satisfied by $\partial_t u(t,x)$ to emulate the procedure in Example 1.1. This can be done thanks to the Dynamic Programming Principle, introduced by R. Bellman.

$$u(t,x) = \inf_{\alpha\in\mathcal{A}_{[t,t+h]}(x)} \int_t^{t+h} L(X_s, \alpha_s)\, ds + u(t+h, X_{t+h}), \quad \forall h \in (0, T-t). \tag{DPP}$$

Essentially, it expresses the fact that the original variational problem can be broken down into subproblems in a recursive manner.

The HJ equation is nothing else than an infinitesimal version of this principle. Indeed, by Taylor expansion:

$$u(t,x) = \inf_{\alpha \in \mathcal{A}_{[t,t+h]}(x)} \Big\{ \int_t^{t+h} L(X_s, \alpha_s)\, ds \\ + u(t, X_t) + h\partial_t u(t, X_t) + \langle \nabla u(t,x), \dot{X}_t \rangle \Big\}$$

Therefore, since $X_t = x$, dividing by $h > 0$, we get

$$0 = \inf_{\alpha \in \mathcal{A}_{[t,t+h]}(x)} \frac{1}{h} \int_t^{t+h} L(X_s, \alpha_s)\, ds + \partial_t u(t,x) + \langle \nabla u(t,x), f(x,u) \rangle.$$

We now pass to the limit $h \to 0$:

$$0 = \inf_{\alpha \in A} \Big\{ L(x,\alpha) + \partial_t u(t,x) + \langle \nabla u(t,x), f(x,\alpha) \rangle \Big\}.$$

Finally, calling

$$H(t,x,\nabla u) := -\sup_{\alpha \in A} \Big\{ -L(x,\alpha) + -\langle \nabla u(t,x), f(x,\alpha) \rangle \Big\}.$$

we get:

$$-\partial_t u(t,x) + H(t,x,\nabla u) = 0 \tag{HJ}$$

coupled with the terminal condition $u(T,x) = g(x)$. It is not difficult to show that we also derived a procedure to exhibit an optimal control α, that at each time must satisfy the inclusion

$$\alpha(t,x) \in \operatorname{argmax}_{\alpha \in A} \Big\{ L(x,\alpha) + \langle \nabla u(t,x), f(x,\alpha) \rangle \Big\}. \tag{Opt}$$

The PDE (HJ) should be seen as the analogue of the ODE $\frac{d}{dt} \mathrm{I}_t$. Then the idea is to solve (HJ) by evaluating $u(0,x) = u(x)$, obtaining the value of optimal cost. On the other hand, the optimal control can be obtained via the workflow described in Figure 1.

In particular, the control is in *feedback form*: the state of the system determines the way to act on it.

Unfortunately, this procedure assumes the differentiability of the function $u(t,x)$. It is not always the case, as can be easily seen by the method of characteristics.

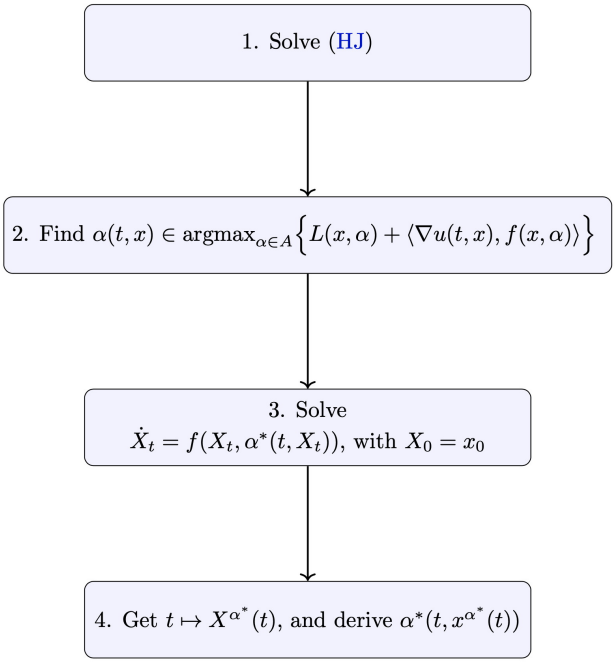


Figure 1: Workflow to solve optimal control problem via the DPP.

Viscosity Solutions

To address the lack of classical solutions for (HJ), the notion of viscosity solutions was introduced by M.G. Crandall and P.-L. Lions in 1981 for first-order Hamilton–Jacobi equations. In what follows, we provide an overview of this approach.

First of all we have to relax the notion of differential.

Definition 1.2 Let $u \in C((0,T)\times \mathbb{T}^d)$. We call the superdifferential of u at the point $x \in \mathcal{T}^d$ the closed, convex subset of $\mathbb{R}\times\mathbb{R}^d$
(D^+)
$$D^+u(x) := \Big\{(r,p) \in (0,T)\times\mathbb{R}^d : u(t+h, x+v) - u(t,x) - \langle p, v\rangle + r(s-t) \le o(|v|+|h|)\Big\}$$

We call the subdifferential of u at the point $x \in \mathbb{R}^n$ the closed, convex subset of $\mathbb{R}^n$
(D^-)
$$D^-u(x) := \Big\{(r,p) \in (0,T)\times\mathbb{R}^d : u(t+h, x+v) - u(t,x) - \langle p, v\rangle + r(s-t) \ge o(|v|+|h|)\Big\},$$

where $o(|v|+|h|)$ is s.t. $\lim_{v,h\to 0} \frac{o(|v|+|h|)}{|v|+|h|} = 0$.

In the case u is differentiable at x, we have $D^+u(x) = D^-u(x) = \{Du(x)\}$. On the other hand, the intersection $D^+u(x) \cap D^-u(x)$ between the two is not empty iff u is differentiable at x.

Remark 1.1 In the finite dimensional case, an equivalent, and more intrinsic way of defining super/sub differential can be considered. We refer to Proposition 3.1.7 [5] for details. The point $p \in \mathbb{R}^d$ is in $D^+u(x)$(resp. $D^-u(x)$) iff there exists $\phi \in C^1(\mathbb{T}^d)$, with $D\phi(x) = p$ s.t. x is local maximum (resp. minimum) of $u - \phi$. Since this condition is not sentitive to adding and substracting constants we can equivalently say that ϕ touches u from above(resp. below) at the point x. Namely, there exists $r > 0$ s.t.

$$u(x) = \phi(x) \quad \& \quad u \leq (\geq)\phi \text{ on } B_r(x).$$

Definition 1.3 Let $u \in C([0,T] \times \mathbb{T}^d)$. We say that u is a **subsolution** of (HJ) iff, for all $(t,x) \in (0,T) \times \mathbb{T}^d$:

$$(\mathrm{S}^-) \qquad -r + H(x,p) \leq 0 \quad \forall p \in D^+u(x)$$

Let $u \in C([0,T] \times \mathbb{T}^d)$. We say that u is a **supersolution** of (HJ) iff, for all $(t,x) \in (0,T) \times \mathbb{T}^d$:

$$(\mathrm{S}^+) \qquad -r + H(x,p) \geq 0 \quad \forall p \in D^-u(x)$$

Let $u \in C([0,T] \times \mathbb{T}^d)$. We say that u is a **solution** if it is both a supersolution and a subsolution.

This notion is *the good one* in the sense of Hadamard: it enjoys the following property (**(U) Uniqueness**), (**(S) Stability**), (**(E) Existence**).

The first two properties can be obtained by this stronger principle

Theorem 1 (Comparison Principle) *Let u, v be a subsolution and a supersolution of* (HJ)*, respectively. Under suitable continuity assumptions on H and g* (see [2] Theorem 3.1) *we have*

$$(\mathrm{CP}) \qquad \max_{(t,x)\in[0,T]\times\mathbb{T}^d} u(t,x) - v(t,x) \leq \max_{x\in\mathbb{T}^d} u(T,x) - v(T,x).$$

In particular, the equation (HJ) *has at most a unique solution for a fixed terminal condition.*

Note that the previous result also gives a stability estimate in terms of the terminal condition at time T.

Sketch of the proof. Denote by M the RHS of (CP). Proceed by contradiction: There exists $\lambda > 0$ s.t.

$$(1.5) \qquad M_\lambda = \max_{(t,x)\in[0,T]\times\mathbb{R}^n} u(t,x) - v(t,x) + \lambda(T-t) \geq M$$

We introduce a penalization that doubles the variable

$$(1.6) \qquad M_{\varepsilon,\eta} = \max_{(t,x)\in[0,T]\times\mathbb{R}^n} u(t,y) - v(s,x) + \lambda(T-t) - \frac{1}{2\varepsilon}|y-x|^2 - \frac{1}{2\eta}|t-s|^2.$$

As $\varepsilon, \eta \to 0$, the penalizations vanish, and $M_{\varepsilon,\eta} \to_{\varepsilon,\eta\to 0} M_\lambda$. Now, by using the Remark 1.1, we produce elements $(\frac{\bar{t}-\bar{s}}{\eta} - \lambda, \frac{\bar{x}-\bar{y}}{\varepsilon}) \in D^+u(\bar{t},\bar{x})$, and $(\frac{\bar{t}-\bar{s}}{\eta} - \lambda, \frac{\bar{x}-\bar{y}}{\varepsilon}) \in D^-v(\bar{s},\bar{y})$, where $(\bar{t},\bar{s},\bar{x},\bar{y})$ is a maximum point in (1.6). We can then use the definitions (S^-) and (S^+):

$$\lambda \leq H(\bar{t},\bar{x},\frac{\bar{x}-\bar{y}}{\varepsilon}) - H(\bar{s},\bar{x},\frac{\bar{x}-\bar{y}}{\varepsilon}) \underbrace{\leq}_{\text{Continuity assumptions on } H} o(1). \tag{1.7}$$

That yields a contradiction. □

Remark 1.2 (On the choice of the penalization) The use of a squared penalization is classical in the theory of viscosity solutions, and the doubling of variables technique originates from the work of Kružkov. In a forthcoming paper with C. Bertucci,[(1)] we revisit this argument in a more abstract framework. Specifically, we consider a Riemannian manifold as the underlying space and focus on a convex Hamiltonian H. Within this setting, we introduce a more intrinsic penalization strategy, which enables us to reinterpret the difference $u-v$ as a kind of Lyapunov function. This perspective provides a natural explanation for its monotonicity in time.

Concerning the **existence**, assuming the Hamiltonian comes from a control problem, the goal is to prove that the value function is *the* viscosity solution of (HJ). We will not give here a proof of this classical result that can be found in [2], Theorem 3.17.

1.3 Vanishing Viscosity

There is also another way to recover the viscosity solution of (HJ), that motivates the attribute *viscosity*. It relies on a regularization procedure.
Consider the second-order PDE

$$\begin{cases} -\partial_t u^\varepsilon + H(x,\nabla u^\varepsilon) - \varepsilon\Delta u^\varepsilon = 0 \\ u^\varepsilon(T,x) = g(x) \end{cases} \tag{HJ_ε}$$

The second-order term changes the scale of the equation and has a regularizing effect. Under mild assumption on the Hamiltonian, it can be proved

(a) u_ε is uniformly bounded

(b) u_ε uniformly α-Holder.

(c) By Ascoli-Arzelà, the sequence u_ε is precompact in the uniform topology, and every accumulation point is a viscosity solution.

(d) By the uniqueness result expressed in Theorem 1, any accumulation points must coincide, i.e. the sequence *converges* to a function u, solution of the equation (HJ).

[(1)] CNRS researcher in Mathematics at the Department of Applied Mathematics, École Polytechnique, Palaiseau, France

(e) Theorem 1.5 [2]. The quantitative estimates holds: there exists $C > 0$ s.t. for all $\varepsilon > 0$

$$\|u^\varepsilon - u\|_{L^\infty(\mathbb{T}^d)} \leq C\sqrt{\varepsilon}.$$

2 N-agents Problems

In application, many interesting problems arise when several agents interact in the optimization problem. Unfortunately, as in classical mechanics, this is not computationally feasible. We briefly describe the main idea behind the so-called mean field approximation, introduced to overcome this complexity.

2.1 N-agent value problem

The $(\mathbb{T}^d)^N$-valued state dynamic evolves according to the controlled dynamics

$$\begin{cases} \dot{X}_t^i = f(X_t^i, \alpha_t^i) \\ \mathbf{X}_0 = \mathbf{x}_0 \in (\mathbb{T}^d)^N, \end{cases} \tag{N-Dyn}$$

where $\boldsymbol{\alpha} = (\alpha^1, \dots, \alpha^N) \in \mathcal{A}^N$ is the population control and $\mathbf{x}_0 = (x_0^1, \dots, x_0^N)$ is the initial condition. We endow the space $(\mathbb{T}^d)^N$ with the distance

$$\|\mathbf{x}_0 - \mathbf{x}_1\|_{(\mathbf{T}^d)^N} = \sqrt{\frac{1}{N}\sum_{i=1}^N |x_0^i - x_1^i|^2} \tag{0.4}$$

We denote by

$$\mathcal{A}^N_{[t,T]}(\mathbf{x_0}) := \left\{ [t,T] \ni s \mapsto \mathbf{X}_s \text{ solution of (N-dyn)} : \mathbf{ff} \in \mathcal{A}^N \right\} \tag{2.1}$$

the set of controlled trajectories of the N-agents.

The value function $V^N : [0,T] \times (\mathbb{T}^d)^N \to \mathbb{R}$ is defined by

$$V^N(t_0, \mathbf{x}_0) := \inf_{\boldsymbol{\alpha} = (\alpha^1, \dots, \alpha^N) \in \mathcal{A}^N_{[t_0,T]}(\mathbf{x}_0)} J^N(t_0, \mathbf{x}_0, \boldsymbol{\alpha}),$$

where

$$J^N(t_0, \mathbf{x}_0, \boldsymbol{\alpha}) := \int_{t_0}^T \left(\frac{1}{N}\sum_{i=1}^N L(X_t^i, \alpha_t^i) + F(m^N_{\mathbf{x}_t}) \right) dt + G(m^N_{\mathbf{x}_T}). \tag{N-Score}$$

Here

$$m_{\mathbf{x}} = \frac{1}{N}\sum_{i=0}^N \delta_{x^i} \tag{EmpMea}$$

denotes the empirical measure.

The associated Hamilton Jacobi equation is

$$\text{(HJB}_N\text{)} \qquad \begin{cases} -\partial_t u^N(t,\mathbf{x}) + \frac{1}{N}\sum_{i=1}^N H(x_i, ND_{x_i}u^N) = 0, \quad (0,T)\times(\mathbb{T}^d)^N \\ u(0,\mathbf{x}) = G(m^N_{\mathbf{x}}) \end{cases}$$

The mean field part relies in the particular structure of the interactions that are all averaged and the common law of evolution given by f. The limit setting is formally obtained by $N \to \infty$, but a difficulty arises in interpreting the meaning of

$$\lim_{N\to\infty} (\mathbb{T}^d)^N. \tag{2.2}$$

The idea, that will be clarified in the next section, is to consider a space that (1) contains all the spaces $(\mathbb{T}^d)^N$, as $N \in \mathbb{N}_{>0}$ and (2) reflects still well the geometry if the underlying space. This space will be the space of probability measures endowed with the Wasserstein distance.

3 Short Introduction to the Wasserstein space

Let X be compact Hausdorff space, and let $\mathcal{M}(X)$ be the set of finite signed Borel measures over it. By Riesz Markov Theorem, see Theorem IV.14 [10], $\mathcal{M}(X)$ can be realized as the dual space of $C(X)$:

$$\forall l \in C(X)^* \quad \exists!\, \mu \in \mathcal{M}(X) \text{ s.t. } l(\phi) = \int_X \phi d\mu \quad \forall \phi \in C(X), \text{ and } \|l\|_* = |\mu|(X), \tag{3.1}$$

where $\|\cdot\|_*$ is the operator norm and $|\mu|(X) = \sup\left\{ \sum_{i=1}^N |\mu(A_i)| \colon A_i \text{ Borel measurable } X = \bigsqcup A_i \right\}$ is the total variation of μ. In this setting, the weak* topology can be described as follows

$$\{\mu_n\} \subset \mathcal{M}(X),\ \mu_n \rightharpoonup \mu \in \mathcal{M}(X) \iff \int_X f\mu_n \to \int_X f\mu \quad \forall f \in C(X). \tag{3.2}$$

Usually, the convergence in this setting is referred to as *narrow convergence* We denote by

$$\mathcal{P}(X) := \left\{ \mu \in \mathcal{M}(X) \,|\, \mu \geq 0, \int_X \mu = 1 \right\}, \tag{3.3}$$

the narrowly closed convex set of the probability measures over X. Moreover, by Prokhorov Theorem [11] pag. 4, $\mathcal{P}(X)$ is a compact space in the narrow topology.

Given a Borel measurable map $T : X \to Y$ and a measure $\mu \in \mathcal{P}(X)$, we define the *pushforward measure* $T_\# : \mathcal{P}(X) \to \mathcal{P}(Y)$ as follows

$$T_\#\mu(A) = \mu(T^{-1}(A)) \quad \forall A \subset X \text{ Borel measurable.}$$

It is easy to check that $T_\#\mu$ does the dine a probability measure over Y. Moreover, we have the following useful formula

$$\int_X \phi(x) dT_\#\mu = \int_X \phi(T(x)) d\mu, \quad \forall \phi \in C(X).$$

We say that T *transports* μ to ν if $T_{\#}\mu = \nu$. Note that, given two measures μ, ν might not exist a T that transport μ to ν, as shown by the following simple observation

$$\mu = \delta_0 \Longrightarrow T_{\#}\mu = \delta_{T(0)}.$$

We will denote by $\Gamma_M(\mu, \nu)$ the set of Borel measureable maps transporting μ to ν.

We consider the set of *couplings or transport plans* between μ_0 and μ_1

$$\Gamma_K(\mu_0, \mu_1) := \Big\{\gamma \in \mathcal{P}(X \times X) \,|\, \pi_{i\#}\gamma = \mu_i\Big\}, \tag{3.4}$$

where π_i is the projection of $X \times X$ on the $i = 0, 1$ fractor. This set turns out to be a non-empty ($\mu \otimes \nu \in \Gamma_K(\mu, \nu)$) compact set for the narrow convergence (see Theorem 1.4 [11].) Moreover,

$$\begin{aligned} &\Gamma_M(\mu, \nu) \hookrightarrow \Gamma_K(\mu, \nu) \\ (M \hookrightarrow K) \qquad &T \longmapsto (\mathrm{Id} \times T)_{\#}\mu.. \end{aligned} \tag{3.5}$$

A plan $\gamma = (\mathrm{Id} \times T)_{\#}\mu$ is said to be induced by the map T.

3.1 Monge Problem

In 1781 in *Mémoire sur la théorie de déblais et des remblais* G. Monge introduced for the first time the following[(2)]

Problem 3.1 Fix $\mu, \nu \in \mathcal{P}(X)$. Describe the variational problem

$$C_M(\mu, \nu) := \inf_{T \in \Gamma_M(\mu,\nu)} \int_X |T(x) - x|^2 d\mu(x). \tag{MP}$$

The value $C_M(\mu, \nu)$ is the so-called *Monge transportation cost*. As previously said, this value can be $+\infty$, since the constraint might not be satisfied.[(3)] In '38 Kantorovich proposed a relaxed version

Problem 3.2 Fix $\mu, \nu \in \mathcal{P}(X)$. Describe the variational problem

$$C_K(\mu_0, \mu_1) = \inf_{\gamma \in \Gamma(\mu_0,\mu_1)} \Big\{ \int_{\mathbb{T}^d \times \mathbb{T}^d} |x - y|^2 d\gamma(x, y) \Big\} \tag{KP}$$

Essentially, the fundamental idea of Kantorovic is that the element of mass located at a point x can be transported among several locations. This is forbidden in the Monge formulation. By standard tools we have

[(2)] Actually, the original problem was formulated for $|x - y|$ as transport cost.

[(3)] There are also an other issue: Even if $\Gamma_M(\mu, \nu) \neq \emptyset$, the constraint is not closed under weak convergence.

Proposition 3.1 *The value*

$$C_K(\mu_0,\mu_1)=\min_{\gamma\in\Gamma(\mu_0,\mu_1)}\Big\{\int_{\mathbb{T}^d\times\mathbb{T}^d}|x-y|^2d\gamma(x,y)\Big\} \tag{3.6}$$

is well defined by the direct method of Calculus of Variations and its square root $W_2(\mu_0,\mu_1)=\sqrt{C(\mu_0,\mu_1)}$ *is a distance (Proposition 5.1 [11].) The non empty set of optimal transport plans will be denoted by* $\Gamma_0(\mu_0,\mu_1)$*. Moreover, the space* $\mathcal{P}(\mathbb{T}^d)$ *endowed with this distance has the following properties*

(1) *has topology inducing the narrow convergence* (Theorem 5.11 [11]),

(2) *is complete* (Theorem 6.18 [12]),

(3) *is geodesic* (Theorem 5.27 [11]).

Note that, in general, $C_K(\mu,\nu)\le C_M(\mu,\nu)$.

Example 3.1 (Solution of the previous problems for empirical measures) The above formulations are stated for general measures in $\mathcal{P}(\mathbb{T}^d)$. Here, we specialize in the particular case of empirical measures. It turns out that we are dealing with a Linear Programming problem. We refer to [9] for more details on this discrete setting. Fix $\mu=\sum_{i=1}^n\alpha_i\delta_{x_i}$ and $\nu=\sum_{j=1}^m\beta_j\delta_{x_i}$, with $\sum_{i=1}^n\alpha_i=\sum_{j=1}^m b_j=1$. Then

$$\Gamma^K(\mu,\nu)\longleftrightarrow \mathrm{U}(\alpha,\beta)=\Big\{\mathrm{P}\in\mathbb{R}_+^{nm}:\mathrm{P}\mathbb{1}_n=\alpha\ ,\mathrm{P}^{\mathrm{T}}\mathbb{1}_m=\beta\Big\}, \tag{3.7}$$

while

$$\Gamma^M(\mu,\nu)\longleftrightarrow \mathrm{U}_M(\alpha,\beta)=\Big\{\mathrm{P}\in\mathbb{R}_+^{nm}:\mathrm{P}\mathbb{1}_n=\alpha\ ,\mathrm{P}^{\mathrm{T}}\mathbb{1}_m=\beta,\mathrm{P}\text{ permutation matrix}\Big\}, \tag{3.8}$$

Morever,

$$C_K(\mu,\nu)=\min_{\mathrm{P}\in\mathrm{U}(\alpha,\beta)}\mathrm{Tr}(C^{\mathrm{T}}P), \tag{LP}$$

where $C=(C_{i,j})\in\mathbb{R}^{n\times m}$ and $C_{i,j}=|x_i-x_j|^2$. In the particular case $n=m$ and $\alpha_i=\beta_i=\frac{1}{n}$, as in the case of (EmpMea), a nice characterization is available. It is a consequence of the celebrated Birkhoff- Von Neumann Theorem and every minimizer of (LP) is a permutation matrix. In this particular case the Kantorovic Problem is equivalent to the Monge one. Moreover, we have the following

(a) **lifting property:**

$$W_2(\delta_{x_0},\delta_{x_1})=|x_0-x_1|; \tag{Lift}$$

(b) **contraction property**:

$$\begin{aligned} W_2^2(m^N_{\mathbf{x_0}},m^N_{\mathbf{x_1}}) &\underbrace{=}_{\sigma\text{ optimal permutation}} \frac{1}{N}\sum_{i=1}^N|x_0^i-x_1^{\sigma(i)}|^2\\ &\le \frac{1}{N}\sum_{i=1}^N|x_0^i-x_1^i|^2 \underbrace{=}_{(\mathrm{d}_N)} \|\mathbf{x}_0-\mathbf{x}_1\|^2_{(\mathbf{T}^d)^N}. \end{aligned} \tag{3.9}$$

In other words, we have the 1-Lipschitz embedding

$$\begin{aligned} i_N : (\mathbb{T}^d)^N &\hookrightarrow \mathcal{P}(\mathbb{T}^d) \\ \mathbf{x} &\mapsto m_{\mathbf{x}}^N. \end{aligned}$$

We denote by λ_d the Lebesgue measure over $\mathbb{T}^d$.

3.2 When Monge=Kantorovic?

Suppose $C_M(\mu,\nu) < +\infty$, what can be said about the relation with $C_K(\mu,\nu)$? Obviously, $C_K(\mu,\nu) \leq C_M(\mu,\nu)$, by (3.5). In the Example (EmpMea), we have seen a sufficient condition. Here, we state another important sufficient condition.

Theorem 2 (Brenier) *Assume $\mu << \lambda_d$. Then*

- *The problem (3.6) has a unique solution γ. Moreover, γ is induced by a map T and $T = \nabla\psi$, where $\psi : \mathbb{R}^n \to (-\infty, +\infty]$ l.s.c., convex function, differentiable μ-a.e.;*
- *Conversely, if ψ is convex, l.s.c., differentiable μ a.e, with $\nabla\psi \in L^2(\mu)$, then $T := \nabla\psi$ is optimal from μ to $\nu = T_{\#}\mu$.*
- $C_K(\mu,\nu) = C_M(\mu,\nu)$

3.3 Control Problems

To introduce control problems in the Wasserstein space, we have to understand what is the analogue in this framework of an ODE. We start with the following

Example 3.2 Let $v : (0,T) \times \mathbb{T}^d \to \mathbb{R}^d$ be a time dependent vector field. We suppose $v \in L^1_t \mathrm{Lip}_x$, namely v is uniformly Lipschitz in the spatial variable and uniformly L^1 integrable in time. (This is not the sharp condition, see Lemma 8.1.4 [1]. By the Cauchy-Lipthitz Theorem, the ODE

$$\begin{cases} \dot{X}_t = v(t, X_t) \\ X_s = x \end{cases} \tag{3.10}$$

is uniquely solvable. Now, take $\mu \in \mathcal{P}(\mathbb{T}^d)$ and consider $\mu_t := (X_t)_{\#}\mu$. In this case, the measure μ_t is the unique weak solution[(4)] of the following continuity equation

$$\begin{cases} \partial_t \mu_t + \mathrm{div}(v(t,x)\mu_t) = 0 \\ \mu_0 = \mu \end{cases} \tag{CE}$$

[(4)] $\frac{d}{dt}\int \phi(x) d\mu_t = \int \langle \nabla\phi, v_t \rangle d\mu_t$ for a.e. $t \in [0,T]$ and $\forall \phi \in C^1(\mathbb{T}^d)$

The previous example shows that under some regularity on vector field we have a correspondence from micro to macro: It is sufficient to understand the behavior in time of each particle X_t to reconstruct the dynamics of the entire population μ_t.

Another way to look at the (CE) relies on its connection with the metric structure of $(\mathcal{P}(\mathbb{T}^d), W_2)$.

Definition 3.1 Let (X, d) be a complete metric space and $a, b \in \mathbb{R}$. A curve $\gamma \in C([a, b]; X)$ is 2-absolutely continuous, and we say $\gamma \in \mathrm{AC}_2([a, b]; X)$, if there exists $g \in L^2([a, b])$ s.t.

$$d(\gamma(t), \gamma(s)) \leq \int_s^t g(\tau) d\tau \quad \forall [s, t] \subset [a, b]. \tag{AC$_2$}$$

Theorem 3 ([1], Theorem 1.1.2) *For any $\gamma \in AC_2([a, b])$*

$$|\gamma'|(t) : \lim_{s \to t} \frac{d(\gamma(s), \gamma(t))}{|t - s|} \tag{3.11}$$

exists $\mathcal{L}_1$ -a.e. $t \in (a, b)$. Moreover, $|\gamma'|(\cdot) \in L^2(a, b)$ and it is the minimal function satisfying (AC$_2$).

Remarkably, we have a finer result in the case $(X, d) = (\mathcal{P}(\mathbb{T}^d), W_2)$.

Theorem 4 (Absolutely continuous curves and the continuity equation, [1]) *Let I be an open interval in $\mathbb{R}$, let $\mu_t : I \to \mathcal{P}(\mathbb{T}^d)$ be an absolutely continuous curve and let $|\mu'| \in L^1(I)$ be its metric derivative, given by (3.11). Then there exists a Borel vector field $v : (x, t) \mapsto v_t(x)$ such that*

$$v_t \in L^2_{\mu_t}(\mathbb{T}^d), \qquad \|v_t\|_{L^2_{\mu_t}(\mathbb{T}^d)} \leq |\mu'|(t) \qquad \text{for } \mathcal{L}^1\text{-a.e. } t \in I \tag{3.12}$$

and the continuity equation

$$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0 \qquad \text{in } \mathbb{T}^d \times I \tag{3.13}$$

holds in the sense of distributions.

Moreover, for $\mathcal{L}^1$-a.e. $t \in I$, v_t belongs to the closure in $L^2_{\mu_t}(\mathbb{T}^d)$ of the subspace

$$\left\{ \nabla \phi : \phi \in C^\infty(\mathbb{T}^d) \right\} \tag{3.14}$$

.

Conversely, if a narrowly continuous curve $\mu_t : I \to \mathcal{P}(\mathbb{T}^d)$ satisfies the continuity equation for some Borel velocity field v_t with $L^2_{\mu_t}(\mathbb{T}^d) \in L^1(I)$, then $\mu_t : I \to \mathcal{P}_2(\mathbb{T}^d)$ is absolutely continuous and $|\mu'|(t) \leq \|v_t\|_{L^2_{\mu_t}(\mathbb{T}^d)}$ for $\mathcal{L}^1$-a.e. $t \in I$.

Problem 3.3 Consider the following path cost

$$\begin{cases} \mathcal{J}(\mu, (\alpha_t)_t) = \int_0^T L(x, \mu_t, \alpha_t) d\mu_t + G(\mu_T) \\ \partial_t \mu_t + \operatorname{div}(\alpha_t \mu_t) = 0 \\ \mu_0 = \mu \end{cases} \tag{3.15}$$

and denote by

$$\mathcal{A}_{[a,b]}(\mu) = \Big\{ t \mapsto (\mu_t, \alpha_t) \in \mathcal{P}(\mathbb{T}^d) \times L^2_{\mu_t} : \partial_t \mu_t + \mathrm{div}(\mu_t \alpha_t) = 0 \quad \text{for } (t,\mu) \in (a,b) \times \mathcal{P}(\mathbb{T}^d) \text{ weakly } \& \, \mu_a = \mu \Big\} \tag{3.16}$$

the extended set of admissible controlled curves in the interval [a,b].

Compute

$$\mathcal{J}(\mu) := \inf_{(\mu_t,\alpha_t) \in \mathcal{A}_{[0,T]}(\mu)} \mathcal{J}(\mu, (\alpha_t)_t). \tag{MCP}$$

It is also possible in this case to derive

Proposition 3.2 (Mean Field Dynamic Programming Principle) *For $t \in [0,T)$, define*

$$\mathcal{J}(t, \mu, (\alpha_t)_t) := \inf_{(\mu_t,\alpha_t) \in \mathcal{A}_{t,T}(\mu)} \int_t^T L(x, \mu_s, \alpha_s) d\mu_s + G(\mu_T). \tag{3.17}$$

Then, the value function

$$\mathcal{U}(t,\mu) := \inf_{(\mu_\cdot,\alpha_\cdot) \in \mathcal{A}_{[t,T](\mu)}} \mathcal{J}(t, \mu, (\alpha_t)_t) \tag{MF-VF}$$

satisfied the DPP property, namely

$$\mathcal{U}(t,\mu) = \inf_{(\mu_t,\alpha_t) \in \mathcal{A}_{t,T}(\mu)} \int_t^s L(x, \mu_\tau, \alpha_\tau) d\mu_\tau + \mathcal{U}(s, \mu_s), \quad \forall\, t, s \ s.t. \, 0 \le t < s \le T. \tag{MFDPP}$$

Example 3.3 (Benamou-Brenier Formulation of W_2) A particular case of this class of control problems is the Benamou-Brenier control representation of W_2. Fix $T = 1$, choose $L(x, \mu, \alpha) = \frac{|\alpha|^2}{2}$ and

$$G_\nu(\mu) = \begin{cases} 0 & \text{if } \mu = \nu \\ +\infty & \text{otherwise .} \end{cases} \tag{3.18}$$

In this case, we have

$$W_2^2(\mu,\nu) = \inf \Big\{ \int_0^1 \int_{\mathbb{T}^d} |\alpha_t|^2 d\mu_t \quad | \quad \partial_t \mu_t + \mathrm{div}(\mu_t \alpha_t) = 0; \mu_0 = \mu, \mu_1 = \nu \Big\} \tag{3.19}$$

This formula motivates Otto's interpretation of the Wasserstein space as a Riemannian manifold [8]. Indeed, setting

$$\|\partial_t \mu_t\|^2_{\mu_t} := \inf_{\alpha_t} \Big\{ \int_0^1 \int_{\mathbb{T}^d} |\alpha_t|^2 d\mu_t \quad | \quad \mathrm{div}(\mu_t \alpha_t) = -\partial_t \mu_t \, ; \, \mu_0 = \mu, \mu_1 = \nu \Big\} \tag{3.20}$$

we derive

$$W^2(\mu,\nu) = \inf_{\mu_t} \int_0^1 \|\partial_t \mu_t\|^2_{\mu_t} dt, \tag{3.21}$$

that it is reminiscent of the Riemannian formula for the geodesic distance squared.

See also [7] for futher details about this Riemannian interpretation.

3.4 Subdifferential Calculus

In [1] a powerful subdifferential calculus has been developed in order to study gradient flows in the space of probability measures. Following [4] we adapt this notion in order to better deal with HJ equations.

Given $\gamma, \xi \in \mathcal{P}(\mathbb{T}^d \times \mathbb{R}^d)$, s.t $(\pi_1)_\# \xi = (\pi_1)_\# \gamma = \mu$, and in this case we say $\gamma, \xi \in \mathcal{P}_\mu(\mathbb{T}^d \times \mathbb{R}^d)$, we denote by

$$\Gamma_\mu(\gamma, \xi) = \Big\{ \theta \in \mathcal{P}(\mathbb{T}^d \times \mathbb{R}^d \times \mathbb{R}^d) : (\pi_1, \pi_2)_\# \theta = \gamma, (\pi_1, \pi_3)_\# \theta = \xi \Big\}. \tag{3.22}$$

In analogy with the Riemannian framework, we introduce the exponential map

$$\exp_\mu : \mathcal{P}_\mu(\mathbb{T}^d \times \mathbb{R}^d) \to \mathcal{P}(\mathbb{T}^d) \tag{3.23}$$

$$\gamma \longmapsto (\pi_2 \circ (\pi_1, \pi_2 - \pi_1))_\# \gamma \tag{3.24}$$

Definition 3.2 Let $U : [0,T] \times \mathcal{P}(\mathbb{T}^d) \to \mathbb{R}$ be a continuous function. We call the superdifferential of U at the point $(t, \mu) \in [0,T] \times \mathcal{P}(\mathbb{T}^d)$ the closed convex subset of $\mathbb{R} \times \mathcal{P}(\mathbb{T}^d \times \mathbb{R}^d)$ of elements (r, γ) satisfying

$$(\partial^+) \qquad \begin{aligned} U(s, \exp_\mu \xi) - U(t,\mu) - \textstyle\int_{\mathbb{T}^d \times \mathbb{R}^d} \langle z, v \rangle d\theta(x,z,v) - r(s-t) \\ \leq o\Big(\sqrt{\textstyle\int |v|^2 d\xi(x,v)}\Big) + o(|t-s|), \end{aligned}$$

for all $\xi \in \mathcal{P}(\mathbb{T}^d \times \mathbb{R}^d)$ and $\theta \in \Gamma_\mu(\gamma, \xi)$. We will say $(r, \gamma) \in \partial^+_{t,\mu} U$

Let $U : [0,T] \times \mathcal{P}(\mathbb{T}^d) \to \mathbb{R}$ be a continuous function. We call the subdifferential of U at the point $(t, \mu) \in [0,T] \times \mathcal{P}(\mathbb{T}^d)$ the closed convex subset of $\mathbb{R} \times \mathcal{P}(\mathbb{T}^d \times \mathbb{R}^d)$ of elements (r, γ) satisfying

$$(\partial^-) \qquad \begin{aligned} U(s, \exp_\mu \xi) - U(t,\mu) - \textstyle\int_{\mathbb{T}^d \times \mathbb{R}^d} \langle z, v \rangle d\theta(x,z,v) - r(s-t) \\ \geq o\Big(\sqrt{\textstyle\int |v|^2 d\xi(x,v)}\Big) + o(|t-s|), \end{aligned}$$

for all $\xi \in \mathcal{P}(\mathbb{T}^d \times \mathbb{R}^d)$ and $\theta \in \Gamma_\mu(\gamma, \xi)$. We will say $(r, \gamma) \in \partial^-_{t,\mu} U$.

As in the finite dimensional case, $\partial^+ U(t,)\mu \cap \partial^- U(t,\mu)$ contains at most one element. If so, we say that U is differentiable and we call $\partial_{t,\mu} U = (\partial_t U(t,\mu), \partial_\mu U(t,\mu))$ the *Wasserstein gradient*, the common element in the intersection.

In general, the Wassertein distance squared is not differentiable[(5)]. This result is an immediate consequence of the following

Proposition 3.3 *Fix $\nu \in \mathcal{P}(\mathbb{T}^d)$. Then $\mu \mapsto \frac{W^2(\mu,\nu)}{2}$ is superdifferentiable. Moreover,*

$$\partial^+_\mu W_2^2(\cdot, \nu) \supset \Big\{ \gamma \in \mathcal{P}(\mathbb{T}^d \times \mathbb{R}^d) : \exp_\nu \gamma = \mu \Big\}. \tag{3.25}$$

[(5)] It is the same for the geodesic distance squared in a positively curved Riemannian manifold, e.g. the sphere. However, in this smooth case it is locally differentiable. This last property is not true in the Wasserstein setting

Note the analogy with the usual Riemannian framework:

$$(M,g) \text{ Riemannian manifold} \underbrace{\Longrightarrow}_{d_g \text{ geodesic distance}} \partial_p d_g^2(\cdot, q) \supset \{v \in T_q M : \exp_q(v) = p\}. \tag{3.26}$$

Remark 3.1 It is not difficult to see that if $U : \mathcal{P}(\mathbb{T}^d) \to \mathbb{R}$ is differentiable, it satisfies the following chain rule property

$$\frac{d}{dt} U(\mu_t) = \int \langle \partial_{\mu_t} U, v(t,x) \rangle d\mu_t, \tag{0.5}$$

where μ_t is a weak solution of (CE).

3.5 Hamilton Jacobi Equations in $\mathcal{P}(\mathbb{T}^d)$ & Viscosity Solutions

Performing a DPP as in the first section of this manuscript, it is possible to show that, at least formally (see also Remark 3.1), the value function should satisfy

$$-\partial_t U(t,\mu) + \int H(x, \mu, \partial_\mu U) d\mu = 0, \tag{MF-HJ}$$

where $H(x,\mu,p) = \sup_{v \in \mathbb{R}^d} \Big\{ -p(v) - L(x,\mu,v) \Big\}$.

We want to give a notion of solution of this equation inspired by the finite dimensional viscosity theory.
Since $\partial U(t,\mu) \subset \mathbb{R} \times \mathcal{P}(\mathbb{T}^d \times \mathbb{R}^d)$, the author of [4] introduced the following notion of relaxation for the Hamiltonian[6]

$$\mathcal{H}(\gamma) := \int H(x,p) d\gamma(x,p), \tag{3.28}$$

that of course coincides with the usual one if γ is induced by a map.

Definition 3.3 Let $U : [0,T] \times \mathcal{P}(\mathbb{T}^d)$. We say that U is a **subsolution** of (MF-HJ) iff, for all $(t,\mu) \in (0,T) \times \mathbb{R}^d$:

$$-r + \mathcal{H}(\gamma) \leq 0 \quad \forall (r,\gamma) \in \partial^+_{t,\mu} U \tag{MF S^+}$$

Let $U : [0,T] \times \mathcal{P}(\mathbb{T}^d)$. We say that u is a **supersolution** of (MF-HJ) iff, for all $(t,\mu) \in (0,T) \times \mathbb{R}^d$:

$$-r + \mathcal{H}(\gamma) \geq 0 \quad \forall (r,\gamma) \in \partial^-_{t,\mu} U \tag{MF S^-}$$

Let $U : [0,T] \times \mathcal{P}(\mathbb{T}^d)$. We say that U is a **solution** of (MF-HJ) if it is both a supersolution and a subsolution.

[6] We understood it is quite natural in optimal transport.

Under this definition, it can be proved the following

Theorem 5 (Comparison Principle) *Let $U, V : (0,T) \times \mathcal{P}(\mathbb{T}^d)$ be a subsolution and a supersolution of (MF-HJ), respectively. Then*

$$\max_{(t,x)\in[0,T]\times\mathcal{P}(\mathbb{T}^d)} U(t,\mu) - V(t,\mu) \leq \max_{\mu\in\mathcal{P}(\mathbb{T}^d)} U(T,\mu) - V(T,\mu). \tag{MF-CP}$$

The argument for the aforementioned result closely parallels the one used in the finite-dimensional setting. However, in the present context, the only essential ingredient needed to complete the proof is the *superdifferentiability* of the distance function, along with the explicit characterization of the superdifferential of the squared distance, both derived in Proposition 3.3.

Moreover, we also have the usual characterization of the value function.

Theorem 6 (Existence) *Under the continuous assumption in section 4.2 [4] the value function (MF-VF) is the unique viscosity solution of (MF-HJ).*

4 Perspective

We list some results that will appear soon.

- In the main project of my PhD, under the supervision of Prof. D. Tonon, I focused on the vanishing viscosity limit depicted in Section 1.3 in the infinite dimensional setting $\mathcal{P}(\mathbb{T}^d)$. The analogue of $(\mathrm{HJ}_\varepsilon)$ in the Wasserstein framework is

$$-\partial_t U + \int_{\mathbb{T}^d} H(x, \partial_\mu U) d\mu + \varepsilon \int_{\mathbb{T}^d} \mathrm{div}_x \partial_\mu U d\mu = 0 \tag{4.1}$$

  From the control point of view, this equation expresses optimality of the value in controlling the Fokker Planck

$$\partial_t \mu_t + \mathrm{div}(\alpha_t \mu_t) - \varepsilon \Delta \mu_t = 0. \tag{4.2}$$

  Unfortunately, adding this second operator does not produce a regularization effect. However, we were able to establish the vanishing viscosity limit and exibit a rate of convergence analogous to the finite dimensional case.

- Extend the theory into the space of positive measures $\mathcal{M}_+(\mathbb{T}^d)$. In other words we allow the measures to have different masses. The metric structure has to be renovated and a natural choice is the one introduced in [6]. This is a joint work with C. Bertucci.

References

[1] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré, "Flows in Metric Spaces and in the Space of Probability Measures". Lectures in Mathematics ETH Zürich. Birkhäuser, 2nd ed., 2008.

[2] Martino Bardi and Italo Capuzzo-Dolcetta, "Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations". Systems & Control: Foundations & Applications. Birkhäuser, Boston, 1997.

[3] Richard Bellman, *The theory of dynamic programming.* Bulletin of the American Mathematical Society, 60(6), 503–516, 1954. Includes an extensive bibliography of the literature in the area, up to the year 1954.

[4] Charles Bertucci, *Stochastic optimal transport and Hamilton-Jacobi-Bellman equations on the set of probability measures.* 2023. Preprint, HAL: `hal-04118729v1`.

[5] Piermarco Cannarsa and Carlo Sinestrari, "Semiconcave Functions, Hamilton-Jacobi Equations, and Optimal Control". Volume 58 of Progress in Nonlinear Differential Equations and Their Applications. Birkhäuser, Boston, 2004.

[6] Matthias Liero, Alexander Mielke, and Giuseppe Savaré, *Optimal entropy-transport problems and a new Hellinger-Kantorovich distance between positive measures.* Invent. Math., 211(3): 969–1117, 2018.

[7] John Lott, *Some geometric calculations on Wasserstein space.* 2007.

[8] Felix Otto, *The geometry of dissipative evolution equations: the porous medium equation.* Communications in Partial Differential Equations, 26(1-2):101–174, 2001. Preprint available as MPI MIS preprint 8/1999.

[9] Gabriel Peyré and Marco Cuturi, *Computational optimal transport.* Foundations and Trends in Machine Learning, 11(5-6): 355–607, 2019. Revised version correcting typos in Eqs. (4.43),(4.44).

[10] Michael Reed and Barry Simon, "Methods of Modern Mathematical Physics, Volume I: Functional Analysis". Academic Press, New York, 1972. Library of Congress Catalog Card Number: 75-182650.

[11] Filippo Santambrogio, "Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling". Volume 87 of Progress in Nonlinear Differential Equations and Their Applications. Birkhäuser, Cham, 2015.

[12] Cédric Villani, "Optimal Transport: Old and New". Springer Berlin, Heidelberg, 1st ed., 2008.